1-2017

Penalized Spline Estimation in the Partially Linear Model

Ashley D. Holland
Cedarville University, aholland@cedarville.edu

Journal of Multivariate Analysis 153 (2017) 211–235

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Penalized spline estimation in the partially linear model

Ashley D. Holland

Science and Mathematics Department, Cedarville University, 251 North Main Street, Cedarville, OH 45314, United States

ARTICLE INFO

Article history:
Received 18 December 2015
Available online 11 October 2016

AMS 2000 subject classifications:
62E20
62F12
62G05
62G08
62G20
62H12

Keywords:
Semilinear model
Regression splines
Smoothing splines
Convergence rates
Asymptotic normality

ABSTRACT

Penalized spline estimators have received considerable attention in recent years because of their good finite-sample performance, especially when many regressors are employed. In this paper, we propose a penalized B-spline estimator in the context of the partially linear model and study its asymptotic properties under a two-sequence asymptotics: both the number of knots and the penalty factor vary with the sample size. We establish asymptotic distributions of the estimators of both the parametric and nonparametric components in the model. In addition, as a previous step, we obtain the rate of convergence of the estimator of the regression function in a nonparametric model. The results in this paper contribute to the recent theoretical literature on penalized B-spline estimators by allowing for (i) multivariate covariates, (ii) heteroskedasticity of unknown form, (iii) derivative estimation, and (iv) statistical inference in the semi-linear model, under the two-sequence asymptotics. Our main findings rely on some apparently new technical results for splines that may be of independent interest. We also report results from a small-scale simulation study.

1. Introduction

The partially linear model has a long tradition in statistics and econometrics. In this model, for a dependent variable y and covariates $\mathbf{x} \in \mathbb{R}^{d_x}$ and $\mathbf{z} \in \mathbb{R}^{d_z}$, the conditional mean function is assumed to satisfy

$$\mathrm{E}(y|\mathbf{x}, \mathbf{z}) = \mathbf{x}^\top \boldsymbol{\theta} + g(\mathbf{z}),$$

where both the finite-dimensional parameter $\boldsymbol{\theta}$ and the infinite-dimensional parameter g are of potential interest. This is a very popular model in empirical work because it provides a parsimonious, yet flexible, approach to inference in many different contexts. See, e.g., [23,38] for recent textbook discussions. Typically, the dimension of $\mathbf{x}$ is small while the dimension of $\mathbf{z}$ is large. In the program evaluation literature, for example, $\mathbf{x}$ is usually just a treatment indicator and $\boldsymbol{\theta}$ is the scalar treatment effect of interest, while g is a nonparametric nuisance function which is present to "control" for many ($d_z > 1$) potential confounding factors in a flexible way. See, e.g., [11,26] for a discussion of treatment effects with many covariates. The multivariate function g and its derivatives are also parameters of interest in other cases, for instance in policy analysis, e.g., [43].

Inference in the partially linear model is a well-studied semiparametric problem. For instance, large-sample results are available for inference on $\boldsymbol{\theta}$ and g when the nonparametric component is estimated using kernel regression [37] or regression splines [19]. These results rely on classical smoothing techniques that are sometimes quite sensitive to the specifics of their implementation in applications, a problem that is only exacerbated when the dimension of $\mathbf{z}$ is large, e.g., [10]. Partially motivated by the poor finite-sample performance of conventional smoothing techniques, a recent literature on penalized

E-mail address: aholland@cedarville.edu.

http://dx.doi.org/10.1016/j.jmva.2016.10.001
0047-259X/

spline estimation has emerged and is receiving considerable attention. Proposed by O'Sullivan [35], and later popularized by Eilers and Marx [20], this alternative smoothing technique has generated great interest because it is perceived as a very competitive alternative to classical nonparametric estimators. Section 1.1 offers a brief literature review on penalized spline estimation.

Motivated by their recent popularity, and with the goal of increasing the finite-sample performance of the resulting statistical procedures, we propose a multivariate penalized B-spline estimator for the infinite-dimensional parameter in the partially linear model and study the large-sample properties of the resulting estimators of $\boldsymbol{\theta}$, g, and its derivatives. Our asymptotic results allow for a two-sequence asymptotics where both the number of knots and the penalty term vary with the sample size, following the recent work of Claeskens et al. [15]. This more general asymptotic approximation gives a rich array of possible limiting behaviors for the estimators. Our results extend and complement recent theoretical work in the literature on penalized spline estimators by allowing for multivariate covariates, heteroskedasticity of unknown form, derivative estimation, and statistical inference in the semi-linear model under the two-sequence asymptotics. The main findings rely on some apparently new technical results for B-splines that may be on independent interest. In a Monte Carlo experiment, we find that the estimators perform well in finite samples. The specific results are consistent with the theoretical results given in Sections 3 and 4.

The rest of the paper is organized as follows. The remainder of this section provides a brief literature review on penalized spline estimation and introduces the main notation used throughout this paper. Section 2 introduces the partially linear model and discusses the penalized spline estimator. Section 3 derives the mean-square convergence rates for multivariate penalized spline estimators of the regression function and its derivatives in the nonparametric regression model. These results are then employed in Section 4 to establish asymptotic distributional approximations, with consistent standard-errors, for the resulting estimators of $\boldsymbol{\theta}$, g, and its derivatives. Section 5 summarizes the results of a small-scale Monte Carlo study aimed to assess the finite-sample performance of these estimators. Appendix A contains brief proofs of our main results.

1.1. Related literature

Despite the recent popularity of penalized spline smoothing techniques, there is only a handful of papers analyzing their theoretical properties. Early work has obtained asymptotic results under fixed-knot asymptotics, that is, when the number of knots is assumed to be fixed and only the penalty factor is assumed to vary with the sample size at some appropriate rate. Examples of this approach in the literature include [2,47,48,50]. An alternative large-sample approximation for penalized spline estimators of the regression function was more recently proposed by Hall and Opsomer [22], who considered a sequential asymptotic experiment where first the number of knots is assumed to diverge to infinity and then the (scaled) penalty term is assumed to vanish.

These asymptotic approximations are arguably restrictive and may not always characterize appropriately the finite-sample behavior of the penalized spline estimators. For this reason, more recent work has focused on the asymptotic properties of penalized splines when both the knots and penalty vary with the sample size simultaneously. Claeskens et al. [15] study univariate penalized splines under quite general sequences of tuning parameters (knots and penalty), and show that these estimators are asymptotically equivalent in a mean-square error (MSE) sense to either regression splines or smoothing splines depending on the sequence of tuning parameters considered. Li et al. [30] study univariate penalized splines when the number of knots is "large" and derive an asymptotic equivalence between kernel smoothing and penalized (smoothing) splines. Kauermann et al. [27] extend some of the previous results to the context of univariate generalized spline smoothing, while Krivobokova et al. [28] propose asymptotically conservative uniform confidence bands for univariate penalized spline estimators of the regression function.

Other semi-parametric methods used in the partially linear model include empirical likelihood [31,40], kernels [32,37], and wavelets [12]. Specific to the additive partially linear model, [29] considered general series estimation, including regression splines.

In this paper, we contribute to this emerging literature on two-sequence asymptotics for penalized spline estimators by allowing for multivariate covariates, heteroskedasticity of unknown form, derivative estimation, and statistical inference on both the parametric and nonparametric components in the partially linear model. These extensions are nontrivial, as the mathematical techniques employed for univariate (and level of the) regression function do not immediately extend to multivariate (and derivatives of the) regression functions. We discuss this point more explicitly in Remarks 2 and 3, after the necessary notation and main results have been introduced.

1.2. Notation

We employ the following notation throughout the paper. Let $\mathcal{C}^p(\mathscr{A})$ denote the set of p-times continuously differentiable functions on $\mathscr{A}$. For clarity we bold vectors and matrices but not scalars. For any $k \geq 1$, $\mathbf{0}_k$ denotes the zero vector of dimension k, and $\mathbf{I}_k$ denotes the identity matrix of dimension k. We employ conventional multi-index notation: for a multi-index $\mathbf{k}_n^d \in \mathbb{Z}_*^d$ with $d \in \mathbb{Z}_+$, let $|\mathbf{k}_n^d| = k_1 + \cdots + k_d$, along with the usual Euclidean norm $\|\mathbf{A}\| = \sqrt{\mathrm{tr}(\mathbf{A}^\top \mathbf{A})}$ for scalar, vector, or matrix $\mathbf{A}$. We also let $r_n \asymp \bar{r}_n$ indicate $r_n \geq c_1 \bar{r}_n$ and $r_n \leq c_2 \bar{r}_n$ for some $c_1 > 0$ and $c_2 < \infty$.

2. Model, estimator, and other preliminaries

We assume a random sample $\{(y_i, \mathbf{x}_i^\top, \mathbf{z}_i^\top)^\top : i = 1, \dots, n\}$ of the random vector $(y, \mathbf{x}^\top, \mathbf{z}^\top)^\top$ is observed, where $y \in \mathbb{R}$ is a dependent variable and $\mathbf{x} \in \mathbb{R}^{d_x}$ and $\mathbf{z} \in \mathbb{R}^{d_z}$ are explanatory vectors. We will also assume $\mathbf{z}$ is continuously distributed, but we do not restrict y and $\mathbf{x}$ beyond the usual support and smoothness restrictions, formally discussed below. The semilinear model is

$$y_i = \mathbf{x}_i^\top \boldsymbol{\theta} + g(\mathbf{z}_i) + \boldsymbol{\varepsilon}_i, \qquad \mathrm{E}(\boldsymbol{\varepsilon}_i|\mathbf{x}_i, \mathbf{z}_i) = 0, \qquad \sigma_\varepsilon^2(\mathbf{x}_i, \mathbf{z}_i) = \mathrm{E}(\boldsymbol{\varepsilon}_i^2|\mathbf{x}_i, \mathbf{z}_i),$$

where the vector-valued parameter $\boldsymbol{\theta}$ and the real-valued functions g and σ_ε^2 are unknown. We impose the following conventional assumption on this model.

Assumption 1. (a) $\{(y_i, \mathbf{x}_i^\top, \mathbf{z}_i^\top)^\top : i = 1, \dots, n\}$ is i.i.d., with $\mathbf{z}_i \in \mathscr{Z} = [0, 1]^{d_z}$.
(b) $\mathbf{z}_i$ is continuously distributed on $\mathscr{Z}$ with Lebesgue density $f(\mathbf{z})$ bounded above and bounded away from zero.
(c) $g \in \mathcal{C}^{\alpha_g}(\mathscr{Z})$ for some $\alpha_g \geq 1$.

Our main goal is to develop asymptotically valid inference procedures for $\boldsymbol{\theta}$, g, and its derivatives, employing penalized B-splines to approximate nonparametrically the unknown function g. Specifically, we consider the estimation problem

$$\{\hat{\boldsymbol{\theta}}, \hat{g}\} = \underset{\boldsymbol{\theta}\in\mathbb{R}^{d_x}, s\in\mathscr{S}_{n,r}}{\arg\min} \sum_{i=1}^{n} \{y_i - \mathbf{x}_i^\top \boldsymbol{\theta} - s(\mathbf{z}_i)\}^2 + \lambda_n \int_{\mathscr{Z}} \sum_{|\boldsymbol{\ell}|=m} \{\partial^{\boldsymbol{\ell}} s(\mathbf{z})\}^2 \mathrm{d}\mathbf{z}, \tag{1}$$

where $m \in \mathbb{Z}_+$, λ_n is the penalty sequence, and $\mathscr{S}_{n,r}$ is the set of tensor product B-spline functions of order $r \in \mathbb{Z}_+$ (and degree $r-1$) with knot sequence controlled by the tuning parameter $k_n \to \infty$. For simplicity, we assume the same knot sequence is used for each covariate in $\mathbf{z} = (z_1, \dots, z_{d_z})^\top$, denoted by $\kappa(j, k_n) = \{\kappa_{j,0} < \cdots < \kappa_{j,r-1} = 0 < \kappa_{j,r} < \cdots < \kappa_{j,k_n} = 1 < \kappa_{j,k_n+1} < \cdots < \kappa_{j,k_n+r-1}\}$, and that each knot is simple.

More specifically, the set $\mathscr{S}_{n,r}$ is constructed as follows (see, e.g., [18,39,46] for more details). For each covariate in $\mathbf{z} = (z_1, \dots, z_{d_z})^\top$, a B-spline basis of order $r \in \mathbb{Z}_+$ for the jth covariate, denoted by $\{p_{j,k,r}(z_j)\}_{k=1}^{k_n}$, is constructed by first partitioning the support of z_j, $[0, 1]$ under Assumption 1, into the $k_n - r + 1$ intervals $[\kappa_{j,r}, \kappa_{j,r+1}], \dots, [\kappa_{j,k_n}, \kappa_{j,k_n+1}]$, with knots $0 = \kappa_{j,r} < \cdots < \kappa_{j,k_n+1} = 1$. To manage boundary effects, an extra $2(r-1)$ knots are added with $\kappa_{j,1} < \cdots < \kappa_{j,r} = 0$ and $1 = \kappa_{j,k_n+1} < \cdots < \kappa_{j,k_n+r}$, creating an extended partition. Then the B-splines for the jth covariate are constructed using the well-known Cox–de Boor recursion relation:

$$p_{j,k,1}(z_j) = \begin{cases} 1 & \kappa_{j,k} \leq z_j < \kappa_{j,k+1} \\ 0 & \text{otherwise} \end{cases}$$

and, for $\ell \geq 2$,

$$p_{j,k,\ell}(z_j) = \frac{z_j - \kappa_{j,k}}{\kappa_{j,k+\ell-1} - \kappa_{j,k}} p_{j,k,\ell-1}(z_j) + \frac{\kappa_{j,k+\ell} - z_j}{\kappa_{j,k+\ell} - \kappa_{j,k+1}} p_{j,k+1,\ell-1}(z_j),$$

where $p_{j,k,\ell}$ is the kth spline of order ℓ, and the convention $0/0 = 0$ is used. The set $\{p_{j,k,r}\}_{k=1}^{k_n}$ spans $\mathscr{S}_{n,j,r}$, where

$$\mathscr{S}_{n,j,r} = \{s \in \mathcal{C}^{r-2}([0, 1]) : s \text{ is a polynomial of order } r \text{ on each subinterval } [\kappa_{j,k}, \kappa_{j,k+1}]\}.$$

Multivariate tensor-product splines are formed using

$$\mathbf{p}_n(\mathbf{z}) = (p_1(\mathbf{z}), p_2(\mathbf{z}), \dots, p_{k_n^{d_z}}(\mathbf{z}))^\top = (p_{1,1,r}(z_1), \dots, p_{1,k_n,r}(z_1))^\top \otimes \cdots \otimes (p_{d_z,1,r}(z_{d_z}), \dots, p_{d_z,k_n,r}(z_{d_z}))^\top.$$

Other references on B-splines and related nonparametric estimators include [9,13,20,44].

We also impose the following rate restriction throughout.

Assumption 2. (a) $|\kappa_{j,k_n-1} - \kappa_{j,k_n}| \asymp 1/k_n$, for all k_n and $j = 1, \dots, d_z$.
(b) $\ln(k_n) k_n^{d_z}/n \to 0$.

This assumption describes the conditions imposed in the construction of the B-splines. Part (a) of these conditions is weaker than those imposed in [15,51,52], and analogous to those imposed in [8], which allows for several bases of approximation, and [25]. Part (b) imposes a simple, well-known side condition on the (growth rate of the) number of knots relative to the sample size. See for example [24].

Now observe that for $s \in \mathscr{S}_{n,r}$, the Cartesian product of $\mathscr{S}_{n,1,r}, \dots, \mathscr{S}_{n,d_z,r}$,

$$\int_{\mathscr{Z}} \sum_{|\boldsymbol{\ell}|=m} \{\partial^{\boldsymbol{\ell}} s(\mathbf{z})\}^2 \mathrm{d}\mathbf{z} = \int_{\mathscr{Z}} \sum_{|\boldsymbol{\ell}|=m} \left\{\partial^{\boldsymbol{\ell}} \mathbf{p}_n(\mathbf{z})^\top \boldsymbol{\beta}\right\}^2 \mathrm{d}\mathbf{z} = \boldsymbol{\beta}^\top \mathbf{D} \boldsymbol{\beta}$$

with $\mathbf{D}$ a $k_n^{d_z} \times k_n^{d_z}$ matrix with typical element

$$(D)_{k,j} = \int_{\mathscr{Z}} \sum_{|\boldsymbol{\ell}|=m} \{\partial^{\boldsymbol{\ell}} p_k(\mathbf{z})\}\{\partial^{\boldsymbol{\ell}} p_j(\mathbf{z})\} \mathrm{d}\mathbf{z},$$

and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{k_n^{d_z}})^\top$. Define $\mathbf{y} = (y_1, \ldots, y_n)^\top, \mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top, \mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)^\top$. Also, let $\mathbf{P} = (\mathbf{p}_n(\mathbf{z}_1), \ldots, \mathbf{p}_n(\mathbf{z}_n))^\top$. Under the regularity conditions imposed, it follows that

$$\hat{\boldsymbol{\theta}} = (\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})^- \tilde{\mathbf{X}}^\top \tilde{\mathbf{y}}, \qquad \tilde{\mathbf{X}} = \mathbf{X} - \mathbf{R}\mathbf{X}, \qquad \tilde{\mathbf{y}} = \mathbf{y} - \mathbf{R}\mathbf{y}, \qquad \mathbf{R} = \mathbf{P}(\mathbf{P}^\top \mathbf{P} + \lambda_n \mathbf{D})^- \mathbf{P}^\top,$$

and

$$\hat{g}(\mathbf{z}) = \hat{g}(\mathbf{z}; \hat{\boldsymbol{\theta}}), \quad \hat{g}(\mathbf{z}; \boldsymbol{\theta}) = \mathbf{p}_n(\mathbf{z})^\top (\mathbf{P}^\top \mathbf{P} + \lambda_n \mathbf{D})^- \mathbf{P}^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}),$$

where $\mathbf{A}^-$ denotes a generalized inverse of $\mathbf{A}$. We drop the second evaluation point of $\hat{g}(\mathbf{z}; \boldsymbol{\theta})$ whenever possible for notational simplicity. We consider simultaneously k_n and λ_n varying with the sample size, which heuristically encompasses both regression and smoothing splines procedures for the approximation of the unknown control function $g(\mathbf{z})$ in large samples.

3. Convergence rates in the nonparametric regression model

When $\boldsymbol{\theta} = \mathbf{0}$, the estimation problem (1) reduces to a conventional nonparametric multivariate penalized B-spline regression problem, and the resulting estimator of the regression function $g(\mathbf{z})$, which we will denote by $g_{\boldsymbol{\theta}=\mathbf{0}}(\mathbf{z})$, is given by $\hat{g}_{\boldsymbol{\theta}=\mathbf{0}}(\mathbf{z}) = \hat{g}(\mathbf{z}; \mathbf{0})$. In this section, we derive mean-square convergence rates for this estimator and its derivatives, under the two-sequence asymptotics $k_n \to \infty$ and $\limsup_{n\to\infty} \lambda_n \leq \infty$. Our results not only contribute to the recent literature on asymptotic properties of nonparametric penalized spline estimators, but also will be useful in the following section when studying the large-sample properties of the estimators discussed above for the partially linear model.

To describe the convergence rate result we define the following weighted L_2 norm:

$$\|g_{\boldsymbol{\theta}=\mathbf{0}}\|^2_{w,2,\ell} = \sup_{|\boldsymbol{\ell}|\leq \ell} \|\partial^{\boldsymbol{\ell}} g_{\boldsymbol{\theta}=\mathbf{0}}\|^2_{w,2} = \sup_{|\boldsymbol{\ell}|\leq \ell} \int \{\partial^{\boldsymbol{\ell}} g_{\boldsymbol{\theta}=\mathbf{0}}(\mathbf{z})\}^2 \mathrm{d}w(\mathbf{z}), \quad \boldsymbol{\ell} \in \mathbb{Z}_*^{d_z}.$$

Our results are based on $w(\mathbf{z}) = \hat{F}(\mathbf{z})$ with $\hat{F}(\mathbf{z}) = \sum_{i=1}^n \mathbb{1}(\mathbf{z}_i \leq \mathbf{z})/n$ the empirical distribution function of $\mathbf{z}$, but Remark 1 discusses other fixed norms including the more standard L_2 norm where $w(\mathbf{z}) = F(\mathbf{z})$ with $F(\mathbf{z})$ the distribution function of $\mathbf{z}$.

Theorem 1. *Suppose Assumptions 1 and 2 hold and* $\mathrm{E}(y_i^2|\mathbf{z}_i)$ *is bounded. Let* $r_g = \min(\alpha_g, r-2)$ *and* $m \leq r_g$.

(a) *If* $\lambda_n k_n^{2m}/n < 1$ *for all sufficiently large n, then*

$$\left\|\hat{g}_{\boldsymbol{\theta}=\mathbf{0}} - g_{\boldsymbol{\theta}=\mathbf{0}}\right\|^2_{\hat{F},2,\ell} = \mathrm{O_p}\left\{\frac{k_n^{d_z+2\ell}}{n} + \frac{\lambda_n^2}{n^2} k_n^{2(m+\ell)} + k_n^{-2(r_g-\ell)}\right\}.$$

(b) *If* $\lambda_n k_n^{2m}/n \geq 1$ *for all sufficiently large n, and* $d_z/d < m$ *if* $\limsup_{n\to\infty} \lambda_n k_n^{2m}/n = \infty$, *then*

$$\left\|\hat{g}_{\boldsymbol{\theta}=\mathbf{0}} - g_{\boldsymbol{\theta}=\mathbf{0}}\right\|^2_{\hat{F},2,\ell} = \mathrm{O_p}\left\{\frac{n^{(d_z-2m)/2m} k_n^{2\ell}}{\lambda_n^{d_z/2m}} + \frac{\lambda_n k_n^{2\ell}}{n} + k_n^{-2(r_g-\ell)}\right\},$$

for any $\boldsymbol{\ell} = (\ell_1, \ldots, \ell_{d_z}) \in \mathbb{Z}_*^{d_z}$ *such that* $\max_{\ell\leq j\leq d_z} \ell_j \leq r_g$.

This result establishes the (empirical) mean-square convergence rate for the penalized B-spline estimator of a multivariate regression function and its derivatives, under weak side rate-conditions (i.e., Assumption 2). It follows that the estimator behaves asymptotically as a regression splines estimator in case (a), while it behaves asymptotically as a smoothing splines estimator in case (b); see, e.g., [16,34,36]. The penalized B-spline estimator attains the optimal rate of convergence so that $\|\hat{g}_{\boldsymbol{\theta}=\mathbf{0}} - g_{\boldsymbol{\theta}=\mathbf{0}}\|^2_{\hat{F},2,\ell} = \mathrm{O_p}\{n^{-2(r_g-\ell)/(2r_g+d_z)}\}$ in case (a) if $k_n \asymp n^{1/(d_z+2r_g)}$ and $\lambda_n \lesssim n^{(r_g-m+d_z)/(2r_g+d_z)}$, and in case (b) if $k_n \asymp n^{m/\{r_g(2m+d_z)\}}$ and $\lambda_n \asymp n^{d_z/(2m+d_z)}$ with $m = r_g$.

We obtain the same rates of convergence as in the recent work of Claeskens et al. [15] in the univariate case, but allow for random regressors of any dimension, derivative estimation, weaker rate conditions, and heteroskedasticity.

Remark 1. Under the same assumptions and conditions of Theorem 1, we also show in the supplemental appendix (see Appendix A) that $\left\|\hat{g}_{\boldsymbol{\theta}=\mathbf{0}} - g_{\boldsymbol{\theta}=\mathbf{0}}\right\|^2_{\hat{F},2,q} \asymp_p \left\|\hat{g}_{\boldsymbol{\theta}=\mathbf{0}} - g_{\boldsymbol{\theta}=\mathbf{0}}\right\|^2_{F,2,q}$, with $F(\mathbf{z})$ the distribution function of $\mathbf{z}$. As the proof shows, other weighting functions could also be used under appropriate conditions. We present our results in terms of the empirical norm because the proof will be used heavily in the following section.

Remark 2. Extending [15] to handle multivariate covariates is nontrivial. The main challenge is the eigenvalues of the penalization matrix, which are the solutions $\hat{\mu}_1 \leq \cdots \leq \hat{\mu}_{k_n^{d_z}}$ to

$$\int_{\mathscr{Z}} \sum_{|\boldsymbol{\ell}|=m} \{\partial^{\boldsymbol{\ell}} u(\mathbf{z})\}\{\partial^{\boldsymbol{\ell}} w(\mathbf{z})\} d\mathbf{z} = \hat{\mu} \int_{\mathscr{Z}} u(\mathbf{z}) w(\mathbf{z}) f(\mathbf{z}) d\mathbf{z}, \tag{2}$$

for some $u \in \mathcal{S}_{n,r}$, for all $w \in \mathcal{S}_{n,r}$. These eigenvalues are usually approximated by the first $k_n^{d_z}$ eigenvalues $\mu_1 \leq \cdots \leq \mu_{k_n^{d_z}}$ of the continuous problem with the same equation but with $u \in H^m(\mathscr{Z})$ and $w \in H^m(\mathscr{Z})$, where $H^m(\mathscr{Z})$ is the set of those $L^2(\mathscr{Z})$ functions that have distributional derivatives up to order m in L^2, and $L^2(\mathscr{Z})$ is the set of square-integrable real-valued functions; see, e.g., [1].

In the univariate case, $\mu_1, \mu_2, \ldots$ are found by solving the differential equation $(-d^m w/dw^m)^2 = \mu f w$ with the Neumann boundary conditions, using a transformation (see, e.g., [3,7], [33, p.78], and [42]), which is the approach taken implicitly in [15]. In the multivariate case, the differential equation is $(-\Delta)^m w = \mu f w$. If $f(\mathbf{z}) \equiv 1$, the equation is easily solved for small values of m, as in [21,45] for example, but a solution for variable $f(\mathbf{z})$ was not available in the literature, due to the difficulty associated with mixed partial derivatives. We appeal to a geometric argument using concepts in functional analysis to present an expression for the eigenvalues of the discrete problem directly, without comparing to the continuous problem. See Lemmas 5 and 6 in Appendix A.

Remark 3. Our results permit derivative estimation of the multivariate regression function. We also obtain an apparently novel multivariate derivative approximation result (Lemma 4), extending in particular the results reported in [52] for univariate regression splines.

Remark 4. Recently, it was shown that splines achieve the optimal uniform rate of convergence [14]. We conjecture that Theorem 1 and Remark 1 could be used to establish optimal uniform rates of convergence for splines with a penalization, based on the uniform norm $\|g\|_{\infty,\ell}^2 = \sup_{|\boldsymbol{\ell}|\leq\ell} \|\partial^{\boldsymbol{\ell}} g\|_\infty^2 = \sup_{|\boldsymbol{\ell}|\leq\ell} \sup_{\mathbf{z}\in\mathscr{Z}} |\partial^{\boldsymbol{\ell}} g(\mathbf{z})|^2$. We do not spell out the details here to conserve space.

Remark 5. An extension of the partially linear model can be used in functional data analysis, which is an active field in the statistics literature and is of interest in applied statistics. Specifically, the semi-functional partially linear regression model is $y_i = \mathbf{x}_i^\top \boldsymbol{\theta} + g(t_i) + \varepsilon_i$, where as usual, $\mathrm{E}(\varepsilon_i|\mathbf{x}_i, t_i) = 0$ and g is an unknown smooth function, but now t_i is a random *function* instead of a single random variable. Some recent papers include [4,5]. Aneiros and Vieu [4] discuss using only some discretized values of the functional used to predict y_i, on some grid, which relates to variable selection with many variables, and Aneiros–Pérez and Vieu [5] use a functional version of the familiar Nadaraya–Watson-type kernel estimation. Aneiros and Vieu [4] suggest using a least penalized least squares estimator. A natural question, which could be a topic of future research, is how well penalized splines or even regression splines would perform in this context, in place of the functional version of kernel estimation or penalized least squares.

4. Inference in the partially linear model

In this section we establish asymptotic normality of the estimators $\hat{\boldsymbol{\theta}}$, $\hat{g}(\mathbf{z}) = \hat{g}(\mathbf{z}; \hat{\boldsymbol{\theta}})$, and its derivatives introduced in Section 2 in the context of the partially linear model. We also prove consistency of the natural plug-in heteroskedasticity-robust standard-error estimators. To this end, we introduce the following additional notation:

$$\mathbf{x}_i = \mathbf{h}(\mathbf{z}_i) + \boldsymbol{\nu}_i, \quad \mathbf{h}(\mathbf{z}_i) = \mathrm{E}(\mathbf{x}_i|\mathbf{z}_i), \qquad \mathrm{E}(\boldsymbol{\nu}_i|\mathbf{z}_i) = \mathbf{0}, \quad \boldsymbol{\Sigma}_\nu(\mathbf{z}_i) = \mathrm{E}(\boldsymbol{\nu}_i\boldsymbol{\nu}_i^\top|\mathbf{z}_i),$$

where $\mathbf{h} = (h_1, \ldots, h_{d_x})^\top \in \mathbb{R}^{d_x}$ and $\boldsymbol{\Sigma}_\nu \in \mathbb{R}^{d_x \times d_x}$ are unknown.

We impose the following additional assumption.

Assumption 3. (a) $\mathrm{E}(\boldsymbol{\varepsilon}_i^4|\mathbf{x}_i, \mathbf{z}_i)$ and $\mathrm{E}(\|\boldsymbol{\nu}_i\|^4|\mathbf{z}_i)$ are bounded.
(b) $\sigma_\varepsilon^2(\mathbf{x}, \mathbf{z})$ and $\boldsymbol{\Sigma}_\nu(\mathbf{z})$ are bounded away from zero.
(c) $h_j \in \mathcal{C}^{\alpha_h}(\mathscr{Z})$ for all $j = 1, \ldots, d_x$ and some $\alpha_h \geq 0$.

4.1. Parametric component

To obtain the asymptotic linear representation for $\hat{\boldsymbol{\theta}}$, and establish consistency of an associated plug-in standard-error estimator, we define

$$\mathbf{V}_n = \mathrm{V}(\hat{\boldsymbol{\theta}}|\mathbf{X}, \mathbf{Z}) = \boldsymbol{\Sigma}_\nu(\mathbf{z})_n^- \boldsymbol{\Omega}_n \boldsymbol{\Sigma}_\nu(\mathbf{z})_n^-, \quad \boldsymbol{\Sigma}_\nu(\mathbf{z})_n = \tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}/n, \ \boldsymbol{\Omega}_n = \tilde{\mathbf{X}}^\top(\mathbf{I}_n - \mathbf{R})\boldsymbol{\Sigma}_\varepsilon(\mathbf{I}_n - \mathbf{R})^\top\tilde{\mathbf{X}}/n,$$
$$\boldsymbol{\Sigma}_\varepsilon = \mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top|\mathbf{X}, \mathbf{Z}) = \mathrm{diag}\{\sigma_\varepsilon^2(\mathbf{x}_1, \mathbf{z}_1), \ldots, \sigma_\varepsilon^2(\mathbf{x}_n, \mathbf{z}_n)\},$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$.

Since the only unknown matrix is $\boldsymbol{\Sigma}_\varepsilon$, a simple plug-in standard-error estimator is

$$\hat{\mathbf{V}}_n = \boldsymbol{\Sigma}_\nu(\mathbf{z})_n^- \hat{\boldsymbol{\Omega}}_n \boldsymbol{\Sigma}_\nu(\mathbf{z})_n^-, \quad \hat{\boldsymbol{\Omega}}_n = \tilde{\mathbf{X}}^\top(\mathbf{I}_n - \mathbf{R})\hat{\boldsymbol{\Sigma}}_\varepsilon(\mathbf{I}_n - \mathbf{R})^\top\tilde{\mathbf{X}}/n$$

with

$$\hat{\boldsymbol{\Sigma}}_\varepsilon = \mathrm{diag}(\hat{\varepsilon}_1^2, \ldots, \hat{\varepsilon}_n^2), \quad \hat{\boldsymbol{\varepsilon}} = (\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n)^\top = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}} - \hat{\mathbf{G}},$$

and $\hat{\mathbf{G}} = (\hat{g}(\mathbf{z}_1), \ldots, \hat{g}(\mathbf{z}_n))^\top$.

The following theorem describes our result for $\hat{\boldsymbol{\theta}}$.

Theorem 2. *Suppose Assumptions 1–3 hold, and define* $r_h = \min(\alpha_h, r-2)$. *In addition, assume that* $nk_n^{-2r_g-2r_h} \to 0$ *and one of the following holds:*

(i) *If* $\lambda_n k_n^{2m}/n < 1$ *for all sufficiently large n, then* $\lambda_n^2 k_n^{2m}/n \to 0$.

(ii) *If* $\lambda_n k_n^{2m}/n \geq 1$ *for all sufficiently large n, then* $\lambda_n \to 0$.

Then

$$\mathbf{V}_n^{-1/2}\sqrt{n}\,(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = \mathbf{V}_n^{-1/2}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \boldsymbol{v}_i\boldsymbol{\varepsilon}_i + \mathrm{o_p}(1) \to_d \mathcal{N}(\mathbf{0}_{d_x}, \mathbf{I}_{d_x}),$$
$$\mathbf{V}_n = \boldsymbol{\Sigma}_v(\mathbf{z})^{-1}\boldsymbol{\Omega}\boldsymbol{\Sigma}_v(\mathbf{z})^{-1} + \mathrm{o_p}(1), \quad \boldsymbol{\Sigma}_v(\mathbf{z}) = \mathrm{E}(\boldsymbol{v}_i\boldsymbol{v}_i^\top),\ \boldsymbol{\Omega} = \mathrm{E}(\boldsymbol{v}_i\boldsymbol{v}_i^\top\boldsymbol{\varepsilon}_i^2).$$
$$\hat{\boldsymbol{\Omega}}_n = \boldsymbol{\Omega}_n + \mathrm{o_p}(1) = \boldsymbol{\Omega} + \mathrm{o_p}(1).$$

This theorem outlines a set of simple sufficient conditions to obtain the asymptotic linear representation of $\hat{\boldsymbol{\theta}}$ with heteroskedasticity-robust consistent standard-error estimators. For example, 95% confidence intervals for $\mathbf{c}^\top\boldsymbol{\theta}$, with $\mathbf{c} \in \mathbb{R}^{d_x}$, may be easily constructed using these results, taking the familiar form

$$\mathbf{c}^\top\hat{\boldsymbol{\theta}} \pm 1.96\sqrt{\mathbf{c}^\top\hat{\mathbf{V}}_n\mathbf{c}/n}.$$

In the next section, we explore the performance of these confidence intervals, as a function of the choice of tuning parameters k_n and λ_n.

The condition $nk_n^{-2\alpha_g-2\alpha_h} \to 0$ imposes the usual "undersmoothing" required to remove asymptotically the presence of smoothing-bias in semiparametric estimation. The conditions $\lambda_n^2 k_n^{2m}/n \to 0$ and $\lambda_n \to 0$, which may be binding depending on the asymptotic behavior of the sequence $\lambda_n k_n^{2m}/n$, ensure that the penalization-bias is also negligible in large samples.

4.2. Nonparametric component

Next, we establish the asymptotic distribution of the nonparametric component and its derivatives in the partially linear model. Define

$$W_{n,\boldsymbol{\ell}}(\mathbf{z};\boldsymbol{\theta}) = \mathrm{V}\{\partial^{\boldsymbol{\ell}}\hat{g}(\mathbf{z};\boldsymbol{\theta})|\mathbf{X},\mathbf{Z}\} = \partial^{\boldsymbol{\ell}}\mathbf{p}_n(\mathbf{z})^\top(\mathbf{P}^\top\mathbf{P} + \lambda_n\mathbf{D})^-\mathbf{P}^\top\boldsymbol{\Sigma}_\varepsilon\mathbf{P}(\mathbf{P}^\top\mathbf{P} + \lambda_n\mathbf{D})^-\partial^{\boldsymbol{\ell}}\mathbf{p}_n(\mathbf{z}),$$

for multi-index $\boldsymbol{\ell} \in \mathbb{Z}_*^{d_z}$. A simple plug-in estimator is

$$\hat{W}_{n,\boldsymbol{\ell}}(\mathbf{z}) = \partial^{\boldsymbol{\ell}}\mathbf{p}_n(\mathbf{z})^\top(\mathbf{P}^\top\mathbf{P} + \lambda_n\mathbf{D})^-\mathbf{P}^\top\hat{\boldsymbol{\Sigma}}_\varepsilon\mathbf{P}(\mathbf{P}^\top\mathbf{P} + \lambda_n\mathbf{D})^-\partial^{\boldsymbol{\ell}}\mathbf{p}_n(\mathbf{z}),$$

with $\hat{\boldsymbol{\Sigma}}_\varepsilon$ as given in the previous subsection.

Using this notation we have the following result.

Theorem 3. *Suppose Assumptions 1–3 hold,* $m \leq r_g$, *and* $k_n^{2|\boldsymbol{\ell}|+d_z}/n \to 0$. *In addition, the following hold:*

(i) *If* $\lambda_n k_n^{2m}/n < 1$ *for all sufficiently large n, then*

$$\frac{\lambda_n^2 k_n^{2m-d_z}}{n} \to 0 \quad \textit{and} \quad nk_n^{-d_z-2r_g} \to 0.$$

(ii) *If* $\lambda_n k_n^{2m}/n \geq 1$ *is bounded above for all sufficiently large n, then* $\lambda_n/k_n^{d_z} \to 0$.

(iii) *If* $\lambda_n k_n^{2m}/n \to \infty$, *then*

$$\frac{\max(\lambda_n, 1)\lambda_n^2 k_n^{4m-d_z}}{n^2} \to 0 \quad \textit{and} \quad d_z/4 < m.$$

Then, for $\hat{g}(\mathbf{z}) = \hat{g}(\mathbf{z}; \hat{\boldsymbol{\theta}})$,

$$\frac{\partial^{\boldsymbol{\ell}}\hat{g}(\mathbf{z}) - \partial^{\boldsymbol{\ell}} g(\mathbf{z})}{\sqrt{W_{n,\boldsymbol{\ell}}(\mathbf{z})}} \to_d \mathcal{N}(0,1), \quad \hat{W}_{n,\boldsymbol{\ell}}(\mathbf{z}) = W_{n,\boldsymbol{\ell}}(\mathbf{z}) + \mathrm{o_p}(1),$$

for any $\boldsymbol{\ell} = (\ell_1, \ldots, \ell_{d_z}) \in \mathbb{Z}_*^{d_z}$ *such that* $\max_{1\le j\le d_z} \ell_j \le r_g$.

This theorem gives asymptotic normality of the nonparametric component and its derivatives. The theorem could also be extended to functionals of the nonparametric component, as in [9], for example. The rate restrictions are necessary in order to ensure that the bias vanishes asymptotically.

5. Simulations

This section reports results from a Monte Carlo experiment designed to explore the interaction between the tuning parameters, along with their effect on the coverage rates of the 95% confidence intervals, the mean-square error of $\hat{\boldsymbol{\theta}}$, and the average mean squared error of $\hat{g}(\mathbf{z})$. Specifically, we considered the following model, with $d_z = 1$ and $d_z = 2$:

$$\begin{aligned} y_i &= x_i^\top \beta + g(z_i) + \varepsilon_i, \quad & \varepsilon_i &= \sigma_\varepsilon u_{1i} \\ x_i &= h(z_i) + \nu_i, & \nu_i &= \sigma_\nu u_{2i} \end{aligned}$$

In the univariate model, $d_z = 1, g(z_i) = e^{-32(z_i-0.5)^2} + 2z_i - 1$, and $h(z_i) = z_i/\sqrt{2+z_i}$. In the bivariate model, $d_z = 2, g(z_i) = 3x_i + e^{-32\{(z_{1i}-0.5)^2+(z_{2i}-0.5)^2\}} + 2(z_{1i}+z_{2i}) - 1$, and $h(z_i) = z_{1i}/\sqrt{2+z_{1i}} + z_{2i}/\sqrt{2+z_{2i}}$, chosen to be additive for simplicity. In both models, $d_x = 1$, $\beta = 3$, $z_i \sim \mathcal{U}(0,1)$, $\sigma_\varepsilon = 0.1$, $\sigma_\nu = 1$, and $u = (u_{1i}, u_{2i})^\top \sim \mathcal{N}(0, I_2)$.

In the first part of the study, we considered the empirical coverage rates of $\beta = 3$ and $g(0.67)$ in the first model, along with the mean-square error of $\hat{\beta}$ and the average mean squared error of $\hat{g}(z)$. We used a grid of $\lambda_n = 0, 0.00025, \ldots, 0.0015$ and $k_n - r + 1 = 10, 50, 100, 200, 400, 800$, where as above, $k_n - r + 1$ is the number of subintervals of $[0, 1]$. In the second model, we considered $\beta = 3$ and $g(z_1, z_2)$ with $z_1 = 0.67$ and $z_2 = 0.33$, and with a grid of $\lambda_n = 0, 0.0005, \ldots, 0.004$ and $k_n - r + 1 = 10, 18, 26$. The study is based on 1000 replications, with $n = 1000$, $n = 500$, and $n = 100$.

In each case, the results illustrate that the mean-square error of $\hat{\theta}$ and the average mean-square error of $\hat{g}$ decrease and the coverage rates of θ, $g(0.67)$, and $g(0.67, 0.33)$ get closer to 95% as the sample size increases.

The first set of results correspond to $n = 1000$ (see Tables 1–8), the second set correspond to $n = 500$ (see Tables 9–16), and the third set correspond to $n = 100$ (see Tables 17–24).

In this study, we also considered the bias and variance of $\hat{g}$ in the bivariate model, in order to better understand the empirical coverage rates. The column "95% CI, without bias" reports the coverage of the 95% confidence intervals using

$$z_b \equiv \frac{\hat{g}(0.67, 0.33) - \mathrm{E}_n\hat{g}(0.67, 0.33)}{\sqrt{W_{n,(0,0)}(0.67, 0.33)}},$$

Table 1
Empirical coverage rates of 95% confidence intervals for $g(0.67)$, univariate model, $n = 1000$.

λ_n ($\times 10^{-4}$)	$k_n - r + 1$					
	10	50	100	200	400	800
0	0.941	0.937	0.906	0.829	0.711	0.344
2.5	0.936	0.947	0.947	0.948	0.940	0.938
5	0.921	0.935	0.942	0.947	0.937	0.934
7.5	0.887	0.920	0.934	0.944	0.932	0.934
10	0.847	0.902	0.923	0.932	0.929	0.930
12.5	0.804	0.877	0.918	0.924	0.930	0.927
15	0.746	0.861	0.900	0.924	0.920	0.929

Table 2
Average mean squared error ($\times 10^{-2}$) of $\hat{g}(z_i)$, univariate model, $n = 1000$.

λ_n ($\times 10^{-4}$)	$k_n - r + 1$					
	10	50	100	200	400	800
0	0.013	0.053	0.103	0.204	0.404	0.768
3	0.015	0.325	1.148	2.368	3.354	3.935
5	0.018	0.562	1.719	3.148	4.182	4.765
8	0.021	0.763	2.141	3.680	4.734	5.314
10	0.025	0.941	2.484	4.094	5.158	5.736
13	0.030	1.103	2.777	4.437	5.507	6.083
15	0.035	1.252	3.034	4.733	5.806	6.379

Table 3
Empirical coverage rates of 95% confidence intervals for θ, univariate model, $n = 1000$.

λ_n ($\times 10^{-4}$)	$k_n - r + 1$					
	10	50	100	200	400	800
0	0.946	0.942	0.933	0.923	0.893	0.847
2.5	0.948	0.955	0.923	0.830	0.745	0.693
5	0.948	0.956	0.892	0.761	0.668	0.621
7.5	0.948	0.946	0.854	0.705	0.624	0.572
10	0.949	0.936	0.825	0.675	0.586	0.534
12.5	0.949	0.930	0.794	0.651	0.551	0.498
15	0.950	0.922	0.766	0.625	0.520	0.474

Table 4
Mean squared error ($\times 10^{-4}$) of $\hat{\theta}$, univariate model, $n = 1000$.

λ_n ($\times 10^{-4}$)	$k_n - r + 1$					
	10	50	100	200	400	800
0	0.101	0.106	0.111	0.124	0.164	0.424
2.5	0.101	0.118	0.256	0.697	1.252	1.659
5	0.101	0.143	0.422	1.104	1.823	2.307
7.5	0.101	0.173	0.580	1.438	2.263	2.796
10	0.101	0.205	0.731	1.730	2.634	3.203
12.5	0.101	0.239	0.876	1.993	2.959	3.558
15	0.101	0.275	1.014	2.233	3.253	3.875

Table 5
Empirical coverage rates of 95% confidence intervals for $g(0.67, 0.33)$, bivariate model, $n = 1000$.

λ_n ($\times 10^{-4}$)	$k_n - r + 1$		
	10	18	26
0	0.912	0.816	0.634
5	0.912	0.888	0.882
10	0.901	0.891	0.89
15	0.893	0.894	0.899
20	0.884	0.894	0.903
25	0.864	0.883	0.903
30	0.844	0.878	0.902
35	0.829	0.872	0.900
40	0.804	0.860	0.899

Table 6
Average mean squared error ($\times 10^{-2}$) of $\hat{g}(z_i)$, bivariate model, $n = 1000$.

λ_n ($\times 10^{-4}$)	$k_n - r + 1$		
	10	18	26
0	0.170	0.443	0.819
5	0.107	0.262	0.562
10	0.129	0.430	1.011
15	0.162	0.609	1.451
20	0.200	0.789	1.874
25	0.240	0.969	2.281
30	0.283	1.148	2.675
35	0.326	1.325	3.057
40	0.371	1.501	3.427

where $W_{n,(0,0)}$ is as in Theorem 3 and $\mathrm{E}_n\hat{g}(0.67, 0.33)$ equals the average estimate of $g(0.67, 0.33)$ in the simulations with 1000 repetitions.

The column "95% CI, sim SE" reports the coverage of the 95% confidence intervals using

$$z_{se} \equiv \frac{\hat{g}(0.67, 0.33) - g(0.67, 0.33)}{\sqrt{\mathrm{V}_n\{\hat{g}(0.67, 0.33)\}}},$$

with $\mathrm{V}_n\{\hat{g}(0.67, 0.33)\}$ is equal to the simulation variance with 1000 repetitions (see Tables 25–27).

We note that for samples sizes of 1000 and 500, the coverage rates were best using z_{se} for small values of λ_n and z_b for large values of λ_n. For a sample size of 100, the coverage was best with z_{se} and small λ_n. This illustrates that the standard error

Table 7
Empirical coverage rates of 95% confidence intervals for θ, bivariate model, $n = 1000$.

λ_n ($\times 10^{-4}$)	$k_n - r + 1$		
	10	18	26
0	0.924	0.895	0.798
5	0.935	0.936	0.921
10	0.949	0.94	0.864
15	0.952	0.936	0.786
20	0.959	0.915	0.698
25	0.961	0.894	0.612
30	0.965	0.873	0.536
35	0.962	0.846	0.459
40	0.961	0.818	0.393

Table 8
Average mean squared error ($\times 10^{-4}$) of $\hat{\theta}$, bivariate model, $n = 1000$.

λ_n ($\times 10^{-4}$)	$k_n - r + 1$		
	10	18	26
0	0.120	0.175	0.536
5	0.111	0.120	0.177
10	0.110	0.140	0.327
15	0.110	0.175	0.546
20	0.111	0.222	0.818
25	0.113	0.279	1.136
30	0.115	0.347	1.493
35	0.119	0.424	1.883
40	0.123	0.509	2.305

Table 9
Empirical coverage rates of 95% confidence intervals for $g(0.67)$, univariate model, $n = 500$.

λ_n ($\times 10^{-4}$)	$k_n - r + 1$				
	10	50	100	200	300
0	0.957	0.929	0.856	0.681	0.517
2.5	0.938	0.951	0.954	0.959	0.960
5	0.905	0.93	0.95	0.955	0.955
7.5	0.846	0.905	0.931	0.946	0.948
10	0.794	0.874	0.909	0.93	0.937
12.5	0.735	0.828	0.883	0.907	0.909
15	0.679	0.776	0.841	0.875	0.882

Table 10
Average mean squared error ($\times 10^{-2}$) of $\hat{g}(z_i)$, univariate model, $n = 500$.

λ_n ($\times 10^{-4}$)	$k_n - r + 1$				
	10	50	100	200	300
0	0.026	0.107	0.207	0.407	0.604
2.5	0.028	0.569	1.721	3.149	3.816
5	0.035	0.946	2.485	4.094	4.787
7.5	0.044	1.257	3.035	4.732	5.434
10	0.054	1.525	3.475	5.226	5.932
12.5	0.066	1.765	3.846	5.635	6.342
15	0.078	1.984	4.171	5.985	6.693

Table 11
Empirical coverage rates of 95% confidence intervals for θ, univariate model, $n = 500$.

λ_n ($\times 10^{-4}$)	$k_n - r + 1$				
	10	50	100	200	300
0	0.933	0.915	0.900	0.878	0.846
2.5	0.931	0.955	0.933	0.874	0.838
5	0.934	0.954	0.910	0.821	0.782
7.5	0.933	0.947	0.881	0.784	0.745
10	0.934	0.938	0.861	0.759	0.715
12.5	0.936	0.931	0.842	0.733	0.693
15	0.935	0.925	0.818	0.712	0.674

Table 12
Mean squared error ($\times 10^{-4}$) of $\hat{\theta}$, Univariate Model, $n = 500$.

λ_n ($\times 10^{-4}$)	$k_n - r + 1$				
	10	50	100	200	300
0	0.225	0.244	0.270	0.363	0.540
2.5	0.225	0.299	0.664	1.497	2.025
5	0.225	0.382	1.035	2.222	2.885
7.5	0.226	0.470	1.367	2.796	3.544
10	0.226	0.560	1.670	3.286	4.096
12.5	0.227	0.652	1.952	3.721	4.580
15	0.227	0.744	2.217	4.116	5.017

Table 13
Empirical coverage rates of 95% confidence intervals for $g(0.67, 0.33)$, bivariate model, $n = 500$.

λ_n ($\times 10^{-4}$)	$k_n - r + 1$		
	10	15	20
0	0.884	0.747	0.494
5	0.902	0.874	0.867
10	0.895	0.873	0.877
15	0.876	0.87	0.882
20	0.854	0.863	0.879
25	0.821	0.834	0.863
30	0.77	0.806	0.844
35	0.716	0.769	0.808
40	0.674	0.719	0.774

Table 14
Average mean squared error ($\times 10^{-2}$) of $\hat{g}(z_i)$, bivariate model, $n = 500$.

λ_n ($\times 10^{-4}$)	$k_n - r + 1$		
	10	15	20
0	0.339	0.646	0.998
5	0.178	0.323	0.553
10	0.233	0.518	0.979
15	0.306	0.733	1.413
20	0.387	0.953	1.840
25	0.472	1.175	2.258
30	0.560	1.397	2.667
35	0.650	1.618	3.067
40	0.741	1.838	3.458

Table 15
Empirical coverage rates of 95% confidence intervals for θ, bivariate model, $n = 500$.

λ_n ($\times 10^{-4}$)	$k_n - r + 1$		
	10	15	20
0	0.886	0.834	0.720
50	0.919	0.910	0.899
100	0.924	0.919	0.894
150	0.930	0.922	0.861
200	0.931	0.914	0.819
250	0.939	0.900	0.782
300	0.941	0.890	0.751
350	0.939	0.870	0.713
40	0.940	0.855	0.665

estimates converge more slowly than the estimate, and other methods for estimating the variance, such as bootstrapping, may perform better. This is a topic of future work.

Choosing the parameters k_n and λ_n in practice is a subject of much discussion in the smoothing spline, penalized spline, and ridge regression literature. The most common data-driven method is generalized cross-validation, which is used to determine the optimal λ_n for a fixed k_n. See, for example, [6,17,20,41,49]. In this method, the generalized cross validation

Table 16
Average mean squared error ($\times 10^{-4}$) of $\hat{\theta}$, bivariate model, $n = 500$.

λ_n ($\times 10^{-4}$)	$k_n - r + 1$		
	10	15	20
0	0.3236	0.5978	4.0237
5	0.262	0.2858	0.3558
10	0.2602	0.3241	0.5279
15	0.2662	0.3874	0.7741
20	0.2767	0.4694	1.079
25	0.2907	0.5681	1.434
30	0.3076	0.6821	1.833
35	0.3273	0.8103	2.2714
40	0.3495	0.952	2.7454

Table 17
Empirical coverage rates of 95% confidence intervals for $g(0.67)$, univariate model, $n = 100$.

λ_n ($\times 10^{-4}$)	$k_n - r + 1$				
	10	20	30	40	50
0	0.89	0.857	0.771	0.688	0.586
0.5	0.898	0.87	0.87	0.868	0.869
1	0.895	0.875	0.876	0.878	0.878
1.5	0.892	0.874	0.876	0.874	0.873
2	0.885	0.871	0.866	0.869	0.871
2.5	0.869	0.864	0.862	0.859	0.862
5	0.797	0.78	0.789	0.787	0.783
7.5	0.706	0.659	0.662	0.668	0.668
10	0.62	0.555	0.559	0.563	0.559
12.5	0.528	0.436	0.424	0.412	0.413
15	0.462	0.347	0.332	0.327	0.326

Table 18
Average mean squared error ($\times 10^{-2}$) of $\hat{g}(z_i)$, univariate model, $n = 100$.

λ_n ($\times 10^{-4}$)	$k_n - r + 1$				
	10	50	100	200	300
0	0.133	0.234	0.336	0.440	0.542
0.5	0.104	0.155	0.255	0.408	0.601
1	0.103	0.195	0.379	0.643	0.951
1.5	0.107	0.241	0.501	0.852	1.246
2	0.114	0.288	0.616	1.043	1.505
2.5	0.123	0.337	0.727	1.219	1.738
5	0.181	0.576	1.223	1.958	2.670
7.5	0.252	0.805	1.651	2.553	3.385
10	0.327	1.023	2.033	3.061	3.979
12.5	0.404	1.233	2.382	3.512	4.494
15	0.481	1.433	2.705	3.918	4.952

Table 19
Empirical coverage rates of 95% confidence intervals for θ, univariate model, $n = 100$.

λ_n ($\times 10^{-4}$)	$k_n - r + 1$				
	10	50	100	200	300
0	0.925	0.912	0.894	0.873	0.856
0.5	0.93	0.929	0.938	0.943	0.952
1	0.929	0.934	0.943	0.955	0.954
1.5	0.932	0.94	0.952	0.958	0.962
2	0.93	0.939	0.957	0.962	0.964
2.5	0.931	0.943	0.961	0.964	0.964
5	0.933	0.958	0.966	0.961	0.96
7.5	0.94	0.962	0.966	0.961	0.952
10	0.947	0.959	0.961	0.956	0.946
12.5	0.95	0.962	0.962	0.951	0.942
15	0.955	0.965	0.961	0.944	0.937

Table 20
Mean squared error ($\times 10^{-4}$) of $\hat{\theta}$, univariate model, $n = 100$.

λ_n ($\times 10^{-4}$)	$k_n - r + 1$				
	10	50	100	200	300
0	1.1447	1.2855	1.4706	1.7557	2.0835
0.5	1.1182	1.1511	1.2034	1.2982	1.4451
1	1.109	1.1478	1.2495	1.4338	1.7023
1.5	1.1021	1.1566	1.3091	1.5777	1.9528
2	1.0976	1.1707	1.3739	1.7223	2.1934
2.5	1.0946	1.1878	1.4414	1.8659	2.4252
5	1.0933	1.2956	1.795	2.5617	3.4877
7.5	1.1048	1.4218	2.1584	3.2231	4.4396
10	1.1247	1.5582	2.5234	3.854	5.3126
12.5	1.1513	1.7014	2.8862	4.4562	6.1216
15	1.1833	1.8491	3.244	5.0311	6.8758

Table 21
Empirical coverage rates of 95% confidence intervals for $g(0.67, 0.33)$, bivariate model, $n = 100$.

λ_n ($\times 10^{-4}$)	$k_n - r + 1$		
	4	6	8
0	0.808	0.639	0.035
5	0.758	0.813	0.744
10	0.687	0.771	0.686
15	0.614	0.703	0.606
20	0.549	0.643	0.550
25	0.498	0.584	0.505
30	0.451	0.552	0.458
35	0.417	0.511	0.423
40	0.394	0.464	0.386

Table 22
Average mean squared error ($\times 10^{-2}$) of $\hat{g}(z_i)$, bivariate model, $n = 100$.

λ_n ($\times 10^{-4}$)	$k_n - r + 1$		
	4	6	8
0	0.517	0.870	140.276
5	0.279	0.350	0.445
10	0.325	0.446	0.631
15	0.392	0.583	0.869
20	0.469	0.734	1.124
25	0.551	0.891	1.385
30	0.636	1.050	1.646
35	0.722	1.208	1.905
40	0.808	1.365	2.160

Table 23
Empirical coverage rates of 95% confidence intervals for θ, bivariate model, $n = 100$.

λ_n ($\times 10^{-4}$)	$k_n - r + 1$		
	4	6	8
0	0.889	0.760	0.056
5	0.926	0.912	0.899
10	0.939	0.938	0.924
15	0.948	0.948	0.936
20	0.958	0.953	0.946
25	0.961	0.963	0.954
30	0.963	0.963	0.955
35	0.964	0.967	0.963
40	0.965	0.971	0.961

score for the data generating process $y = \mathbf{x}^\top \boldsymbol{\theta} + g(\mathbf{z}) + \varepsilon$ is given by

$$\mathrm{GCV}(\lambda_n) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^\top \hat{\boldsymbol{\theta}}_{\lambda_n}^{[i]} + \hat{g}_{\lambda_n}^{[i]} - y_i)^2 w_i(\lambda_n), \quad w_i(\lambda_n) = \left[\frac{1 - r_{ii}(\lambda_n)}{\frac{1}{n}\mathrm{tr}\{\mathbf{I} - \mathbf{R}(\lambda_n)\}}\right]^2$$

Table 24
Mean squared error ($\times 10^{-4}$) of $\hat{\theta}$, bivariate model, $n = 100$.

λ_n ($\times 10^{-4}$)	$k_n - r + 1$		
	4	6	8
0	1.805	5.664	9993.012
5	1.257	1.372	1.513
10	1.250	1.328	1.521
15	1.255	1.358	1.648
20	1.271	1.424	1.829
25	1.296	1.511	2.044
30	1.328	1.613	2.286
35	1.365	1.725	2.547
40	1.406	1.844	2.824

Table 25
Empirical coverage rates of 95% confidence intervals for $\hat{g}(0.67, 0.33)$, bivariate model, $n = 1000$.

$k_n - r + 1$	$\lambda_n (\times 10^{-4})$	95% CI	95% CI, without bias	95% CI, sim SE
10	0	0.912	0.914	0.947
10	5	0.912	0.920	0.953
10	10	0.901	0.920	0.936
10	15	0.893	0.919	0.922
10	20	0.884	0.927	0.907
10	25	0.864	0.931	0.888
10	30	0.844	0.932	0.871
10	35	0.829	0.931	0.850
10	40	0.804	0.932	0.829
18	0	0.816	0.816	0.951
18	5	0.888	0.893	0.947
18	10	0.891	0.904	0.944
18	15	0.894	0.907	0.937
18	20	0.894	0.913	0.930
18	25	0.883	0.918	0.919
18	30	0.878	0.924	0.906
18	35	0.872	0.925	0.902
18	40	0.860	0.926	0.890
26	0	0.634	0.617	0.976
26	5	0.882	0.887	0.946
26	10	0.890	0.902	0.947
26	15	0.899	0.902	0.947
26	20	0.903	0.907	0.946
26	25	0.903	0.914	0.943
26	30	0.902	0.921	0.936
26	35	0.900	0.923	0.932
26	40	0.899	0.927	0.925

where $\hat{\boldsymbol{\theta}}^{[i]}_{\lambda_n}$ and $\hat{g}^{[i]}_{\lambda_n}$, the so-called leave-one-out estimators, are the minimizers of (1), with $\mathbf{R}(\lambda_n) \equiv \mathbf{R}$ and $r_{ii}(\lambda_n)$ equal to the ith diagonal element of $\mathbf{R}(\lambda_n)$. A computational formula of the generalized cross validation score is

$$\mathrm{GCV}(\lambda_n) = \frac{\frac{1}{n}\sum_{i=1}^{n}\{\mathbf{x}_i^\top \hat{\boldsymbol{\theta}} + \hat{g}(\mathbf{z}_i) - y_i\}^2}{\left[1 - \frac{1}{n}\mathrm{tr}\{\mathbf{R}(\lambda_n)\}\right]^2}$$

The optimal λ_n is the minimizer of $\mathrm{GCV}(\lambda_n)$ and is approximated using some type of grid search.

The second part of the simulation study used generalized cross-validation, with an appropriate grid in each case, to select an optimal λ_n for various values of the number of knots. The grid was selected in each case by adjusting the maximum value and step size of λ_n based on the optimal value chosen by GCV in the first several repetitions. The same data generating process was used as in the first part of the study, for both the univariate and bivariate cases. The columns "%, $g(0.67)$", "%, $g(0.67, 0.33)$", and "%, θ" report the coverage rates for the 95% confidence intervals. (See Tables 28–33).

The same method can be used to choose k_n and λ_n simultaneously. The equation for the generalized cross validation score becomes

$$\mathrm{GCV}(k_n, \lambda_n) = \frac{\frac{1}{n}\sum_{i=1}^{n}\{\mathbf{x}^\top \hat{\boldsymbol{\theta}} + \hat{g}(\mathbf{z}_i) - y_i\}^2}{\left[1 - \frac{1}{n}\mathrm{Tr}\{\mathbf{R}(k_n, \lambda_n)\}\right]^2},$$

Table 26
Empirical coverage rates of 95% confidence intervals for $\hat{g}(0.67, 0.33)$, bivariate model, $n = 500$.

$k_n - r + 1$	$\lambda_n (\times 10^{-4})$	95% CI	95% CI, without bias	95% CI, sim SE
10	10	0.884	0.889	0.944
10	10	0.902	0.902	0.942
10	10	0.895	0.909	0.929
10	10	0.876	0.916	0.920
10	10	0.854	0.922	0.893
10	10	0.821	0.926	0.858
10	10	0.770	0.931	0.822
10	10	0.716	0.929	0.783
10	10	0.674	0.934	0.727
15	15	0.747	0.752	0.941
15	15	0.874	0.884	0.935
15	15	0.873	0.907	0.936
15	15	0.870	0.913	0.929
15	15	0.863	0.917	0.918
15	15	0.834	0.920	0.891
15	15	0.806	0.922	0.861
15	15	0.769	0.925	0.818
15	15	0.719	0.929	0.775
20	20	0.494	0.097	0.992
20	20	0.867	0.878	0.937
20	20	0.877	0.905	0.937
20	20	0.882	0.911	0.936
20	20	0.879	0.911	0.931
20	20	0.863	0.916	0.915
20	20	0.844	0.918	0.894
20	20	0.808	0.919	0.856
20	20	0.774	0.923	0.814

Table 27
Empirical coverage rates of 95% confidence intervals for $\hat{g}(0.67, 0.33)$, bivariate model, $n = 100$.

$k_n - r + 1$	$\lambda_n (\times 10^{-4})$	95% CI	95% CI, without bias	95% CI, sim SE
4	0	0.808	0.816	0.947
4	5	0.758	0.885	0.871
4	10	0.687	0.897	0.791
4	15	0.614	0.898	0.724
4	20	0.549	0.906	0.656
4	25	0.498	0.909	0.600
4	30	0.451	0.911	0.562
4	35	0.417	0.910	0.532
4	40	0.394	0.910	0.508
6	0	0.639	0.641	0.985
6	5	0.813	0.831	0.937
6	10	0.771	0.853	0.890
6	15	0.703	0.863	0.842
6	20	0.643	0.873	0.787
6	25	0.584	0.882	0.734
6	30	0.552	0.883	0.685
6	35	0.511	0.887	0.636
6	40	0.464	0.889	0.598
8	0	0.035	0.001	0.995
8	5	0.744	0.794	0.911
8	10	0.686	0.829	0.855
8	15	0.606	0.855	0.790
8	20	0.550	0.864	0.732
8	25	0.505	0.872	0.673
8	30	0.458	0.875	0.620
8	35	0.423	0.877	0.577
8	40	0.386	0.887	0.542

where $\mathbf{R}(k_n, \lambda_n) = \mathbf{R}$. The optimal pair of (k_n, λ_n) is the minimizer of $\mathrm{GCV}(k_n, \lambda_n)$ and can be approximated using a grid search on both k_n and λ_n. We present coverage rates and mean-square errors for sample sizes of $n = 1000$, 500, and 100, using the same data-generating process as used above (see Tables 34 and 35).

As seen in these results, generalized cross-validation usually performs well in that it produces small mean-square and average mean-square errors. However, when both λ_n and k_n are allowed to vary, and when k_n is fixed and λ_n varies,

Table 28
Generalized cross-validation, univariate model, $n = 1000$.

$k_n - r + 1$	λ_n ($\times 10^{-6}$)	%, $g(0.67)$	%, θ	AMSE, $\hat{g}$ ($\times 10^{-2}$)	MSE, $\hat{\theta}$ ($\times 10^{-4}$)
10	38.88	0.945	0.946	0.013	0.101
50	4.533	0.937	0.945	0.039	0.105
100	0.996	0.911	0.939	0.059	0.107
200	0.215	0.900	0.939	0.087	0.110
400	0.051	0.879	0.930	0.126	0.114
800	0.014	0.840	0.921	0.177	0.121

Table 29
Generalized cross-validation, bivariate model, $n = 1000$.

$k_n - r + 1$	λ_n ($\times 10^{-4}$)	%, $g(0.67, 0.33)$	%, θ	AMSE, $\hat{g}$ ($\times 10^{-2}$)	MSE, $\hat{\theta}$ ($\times 10^{-4}$)
10	3.543	0.915	0.933	0.106	0.112
18	1.529	0.853	0.920	0.189	0.123
26	0.812	0.797	0.900	0.264	0.137

Table 30
Generalized cross-validation, univariate model, $n = 500$.

$k_n - r + 1$	λ_n ($\times 10^{-6}$)	%, $g(0.67)$	%, θ	AMSE, $\hat{g}$ ($\times 10^{-2}$)	MSE, $\hat{\theta}$ ($\times 10^{-4}$)
10	53.560	0.953	0.934	0.026	0.226
50	4.277	0.941	0.920	0.069	0.239
100	0.942	0.910	0.922	0.102	0.246
200	0.221	0.884	0.913	0.148	0.259
300	0.114	0.866	0.904	0.181	0.268

Table 31
Generalized cross-validation, bivariate model, $n = 500$.

$k_n - r + 1$	λ_n ($\times 10^{-4}$)	%, $g(0.67, 0.33)$	%, θ	AMSE, $\hat{g}$ ($\times 10^{-2}$)	MSE, $\hat{\theta}$ ($\times 10^{-4}$)
10	3.118	0.901	0.916	0.172	0.267
15	1.702	0.840	0.879	0.249	0.297
20	1.049	0.791	0.856	0.317	0.331

Table 32
Generalized cross-validation, univariate model, $n = 100$.

$k_n - r + 1$	λ_n ($\times 10^{-6}$)	%, $g(0.67)$	%, θ	AMSE, $\hat{g}$ ($\times 10^{-2}$)	MSE, $\hat{\theta}$ ($\times 10^{-4}$)
10	74.970	0.890	0.926	0.112	1.124
20	25.480	0.853	0.918	0.155	1.185
30	10.880	0.833	0.909	0.189	1.234
40	5.836	0.817	0.901	0.216	1.273
50	3.737	0.794	0.893	0.242	1.314

Table 33
Generalized cross-validation, bivariate model, $n = 100$.

$k_n - r + 1$	λ_n ($\times 10^{-4}$)	%, $g(0.67, 0.33)$	%, θ	AMSE, $\hat{g}$ ($\times 10^{-2}$)	MSE, $\hat{\theta}$ ($\times 10^{-4}$)
2	0.079	0.289	0.916	0.450	1.535
4	2.500	0.791	0.922	0.284	1.261
6	3.117	0.805	0.891	0.349	1.443

Table 34
Generalized cross-validation, univariate model.

n	$k_n - r + 1$	λ_n ($\times 10^{-6}$)	%, $g(0.67)$	%, θ	AMSE, $\hat{g}$ ($\times 10^{-2}$)	MSE, $\hat{\theta}$ ($\times 10^{-4}$)
1000	7.546	40.970	0.925	0.947	0.013	0.101
500	7.188	47.120	0.921	0.931	0.024	0.226
100	7.160	53.390	0.819	0.914	0.111	1.130

Table 35
Generalized cross-validation, bivariate model.

n	$k_n - r + 1$	λ_n ($\times 10^{-4}$)	%, $g(0.67, 0.33)$	%, θ	AMSE, $\hat{g}$ ($\times 10^{-2}$)	MSE, $\hat{\theta}$ ($\times 10^{-4}$)
1000	4.396	1.263	0.706	0.945	0.053	0.105
500	4.122	1.352	0.756	0.931	0.085	0.234
100	4.880	1.864	0.604	0.727	3.681	243.429

Table 36
Robinson's kernel estimator, univariate model.

n	a	b	%, θ	AMSE, $\hat{g}$ ($\times 10^{-2}$)	MSE, $\hat{\theta}$ ($\times 10^{-4}$)
1000	0.030	0.001	0.971	0.028	0.103
500	0.035	0.001	0.926	0.052	0.237
100	0.049	0.001	0.940	0.212	1.255

Table 37
Robinson's kernel estimator, bivariate model.

n	a	b	%, θ	AMSE, $\hat{g}$ ($\times 10^{-2}$)	MSE, $\hat{\theta}$ ($\times 10^{-4}$)
1000	0.064	0.001	0.974	0.199	0.121
500	0.075	0.001	0.969	0.337	0.282
100	0.104	0.001	0.946	1.099	9.773

generalized cross-validation does not achieve the optimal coverage rate for $\hat{g}$. Also, for 100 observations, it does not choose the optimal parameter values for our purposes, as seen when comparing Tables 33 and 35. Using other data-driven methods in this model is a topic of future research.

The third part of the simulation study evaluated the results from another type of semiparametric estimation, as a way to evaluate and compare the performance of the estimators proposed herein. Specifically, we used Robinson's kernel estimators [37] with the same data generating process used above. In their notation, we used $k(u) = \frac{1}{2}\mathbf{1}(|u| \leq 1)$ with $K(\mathbf{u}) = \Pi_{j=1}^{d_z} k(u_j)$. We choose the bandwidth a and the parameter b, which is used to trim out any small values of the estimate of $f(\mathbf{z})$ via the indicator function $\mathbf{1}(|\hat{f}_i| > b)$, using generalized cross-validation with criterion function

$$\mathrm{GCV}(a, b) = \frac{\frac{1}{n}\sum_{i=1}^{n}\{\mathbf{x}_i^\top \hat{\boldsymbol{\theta}} + \tilde{g}(\mathbf{z}_i) - y_i\}^2}{\left(1 - \frac{1}{n}\sum_{i=1}^{n} L_i\right)^2},$$

with $L_i \equiv L_i(a) = K(0)/\sum_{j=1}^{n} K\{(\mathbf{z}_i - \mathbf{z}_j)/a\}$ and $\tilde{g}(\mathbf{z}_i) = \hat{y}_i - \hat{\mathbf{x}}_i^\top \hat{\boldsymbol{\theta}}$, on an appropriate grid. (The dependence on b is through $\hat{\boldsymbol{\theta}}$.) We noted that the value of b apparently did not change the cross-validation score, implying that none of the estimates of $f(\mathbf{z})$ were small enough to be trimmed out in our study. To the knowledge of the author, Robinson [37] did not propose an estimator for $g(\mathbf{z})$ for $\mathbf{z}$ not in the set of observations, so we did not report the coverage rates of $g(0.67)$ and $g(0.67, 0.33)$ here (see Tables 36 and 37).

In the univariate case using cross-validation, the results are very similar for penalized spline estimation and kernel estimation. The mean-square error of $\hat{\theta}$ and average mean-square error of $\hat{g}$ are slightly smaller with spline estimation, whereas the coverage rates of θ are slightly better with kernel estimation. In the bivariate case with cross-validation, penalized spline estimation performs slightly better for 500 and 1000 observations. For 100 observations (which is quite small for $d_z = 2$), kernel estimation performs better. We suspect that better results could be achieved for spline estimation with a different parameter selector.

Acknowledgments

The author gratefully acknowledges the consistently insightful guidance and mentorship of Matias Cattaneo with this project and paper. The author also wishes to thank Volker Elling for invaluable contributions to this project and paper, along with Mouli Banerjee, Kristen Moore, and Virginia Young, who all provided comments on early versions of the manuscript. The author gratefully acknowledges financial support from the National Science Foundation (SES 1122994).

Appendix A

Let $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ denote the minimum and maximum eigenvalues of the square matrix $\mathbf{A}$. We also introduce the usual regression B-spline estimator $\hat{s}(\mathbf{z}) = \mathbf{p}_n(\mathbf{z})^\top(\mathbf{P}^\top\mathbf{P})^{-}\mathbf{P}^\top\mathbf{y}$, $\hat{s} \in \mathcal{S}_{n,r}$, which equals $\hat{g}(\mathbf{z}; \mathbf{0}_{d_x})$ when $\lambda_n = 0$.

A.1. Preliminary lemmas

We present six preliminary technical lemmas and discuss their novelty relative to the literature on regression and penalized B-splines. The first lemma is a well-known result from the regression B-splines literature [25], which gives (asymptotic) control of the eigenvalues of the B-splines (random) design matrix.

Lemma 1. *Suppose Assumptions 1 and 2 hold. Then, for all sufficiently large n,*

$$\lambda_{\min}(\mathbf{P}^\top\mathbf{P}/n) \asymp_p 1, \qquad \lambda_{\max}(\mathbf{P}^\top\mathbf{P}/n) \asymp_p 1, \quad \text{and} \quad \lambda_{\max}\{\mathrm{E}(\mathbf{P}^\top\mathbf{P}/n) - \mathbf{P}^\top\mathbf{P}/n\} = \mathrm{o}_p(1).$$

The following lemma gives the best L_∞ approximation rate to derivatives of g over $\mathcal{S}_{n,r}$. (Recall that $\partial^{\mathbf{0}_{d_z}} g(\mathbf{z}) \equiv g(\mathbf{z})$.)

Lemma 2. *Suppose Assumptions 1 and 2 hold. Then, for $g \in \mathcal{C}^{\alpha_g}(\mathscr{Z})$ and $\boldsymbol{\ell} = (\ell_1, \ldots, \ell_{d_z})^\top \in \mathbb{Z}_*^{d_z}$ with $\max_{1\le j\le d_z} \ell_j \le r_g$, there exists $s_g \in \mathcal{S}_{n,r}$ such that*

$$\sup_{\mathbf{z}\in\mathscr{Z}} |\partial^{\boldsymbol{\ell}} s_g(\mathbf{z}) - \partial^{\boldsymbol{\ell}} g(\mathbf{z})| \lesssim k_n^{-(r_g - |\boldsymbol{\ell}|)}, \qquad \sup_{\mathbf{z}\in\mathscr{Z}} |s_g(\mathbf{z}) - g(\mathbf{z})| \lesssim k_n^{-r_g}.$$

The next lemma gives simple sufficient conditions ensuring that a particular band matrix has its minimum eigenvalue bounded away from zero. This result will be used in the proof of Lemma 4.

Lemma 3. *Let $\mathbf{A}_N$ be a band matrix of the form*

$$\mathbf{A}_N = \begin{pmatrix} a_1^2 & a_1b_1 & 0 & 0 & 0 & 0 \\ a_1b_1 & a_2^2+b_1^2 & a_2b_2 & 0 & 0 & 0 \\ 0 & a_2b_2 & a_3^2+b_2^2 & \ddots & 0 & 0 \\ 0 & 0 & a_3b_3 & \ddots & 0 & 0 \\ 0 & 0 & 0 & \ddots & a_{N-2}b_{N-2} & 0 \\ 0 & 0 & 0 & 0 & a_{N-1}^2+b_{N-2}^2 & a_{N-1}b_{N-1} \\ 0 & 0 & 0 & 0 & a_{N-1}b_{N-1} & a_N^2+b_{N-1}^2 \end{pmatrix}.$$

Suppose that $a_i \ge c > 0$ for all $i = 1, \ldots, N$, where c is independent of N. Then $\lambda_{\min}(\mathbf{A}_N) \ge c$.

The following lemma describes the rate of approximation of the derivative of a regression B-spline to the unknown regression function for each evaluation point not taking a knots value. We conjecture that this latter restriction, which is not crucial for our purposes, is in fact an artifact of our method of proof. This result extends Lemma 5.1 of Huang [25] to multivariate derivatives.

Lemma 4. *Suppose Assumptions 1 and 2 hold. Then,*

$$|\partial^{\boldsymbol{\ell}}\mathrm{E}\{\hat{s}(\mathbf{z})|\mathbf{X},\mathbf{Z}\} - \partial^{\boldsymbol{\ell}} g(\mathbf{z})| \lesssim_p k_n^{-(r_g - |\boldsymbol{\ell}|)}$$

for any $\mathbf{z} = (z_1, \ldots, z_{d_z})^\top \in \mathscr{Z}$ such that z_j is not equal to a knot value, $j = 1, \ldots, d_z$, and for any $\boldsymbol{\ell} = (\ell_1, \ldots, \ell_{d_z})^\top \in \mathbb{Z}_*^{d_z}$ with $\max_{1\le j\le d_z} \ell_j \le r_g$.

The final two technical lemmas are used to characterize (and control) asymptotically the properly standardized penalization matrix $\mathbf{D}$. More specifically, Lemma 5 finds a simple representation of an appropriately scaled version of $\mathbf{D}$ based on its eigenvalues. Then, Lemma 6 characterizes asymptotically the rate of the eigenvalues.

For any $u, v \in H$, let $\tilde{a}(u,v)$, $b(u,v)$, and $a(u,v)$ be the bilinear forms:

$$\tilde{a}(u,v) = \int_{\mathscr{Z}} \sum_{|\boldsymbol{\ell}|=m} \{\partial^{\boldsymbol{\ell}} u(\mathbf{z})\}\{\partial^{\boldsymbol{\ell}} v(\mathbf{z})\}\mathrm{d}\mathbf{z}, \qquad b(u,v) = \int_{\mathscr{Z}} u(\mathbf{z})v(\mathbf{z})f(\mathbf{z})\mathrm{d}\mathbf{z}, \qquad a(u,v) = \tilde{a}(u,v) + b(u,v).$$

We consider the eigenvalues of the differential equation

$$a(u,v) = \mu b(u,v), \quad \text{for all } v \in \mathcal{S}_{n,r} \text{ and for some } u \in \mathcal{S}_{n,r}, \tag{3}$$

which we denote by $\hat{\mu}_1 \le \cdots \le \hat{\mu}_{k_n^{d_z}}$.

Lemma 5. *Suppose Assumptions 1 and 2 hold. Then, for all n large enough,*

$$\{\mathrm{E}(\mathbf{P}^\top\mathbf{P}/n)\}^{-1/2}\mathbf{D}\{\mathrm{E}(\mathbf{P}^\top\mathbf{P}/n)\}^{-1/2} = \mathbf{U}\mathbf{M}\mathbf{U}^\top,$$

where $\mathbf{M}$ is the $(k_n^{d_z} \times k_n^{d_z})$ diagonal matrix of eigenvalues $\hat{\mu}_1, \ldots, \hat{\mu}_{k_n^{d_z}}$ and $\mathbf{U}$ is an orthogonal matrix of eigenvectors.

Lemma 6. *Let $\{\hat{\mu}_k : k = 1, \ldots, k_n^{d_z}\}$ be the eigenvalues associated with (3). Then, $\hat{\mu}_1 = \cdots = \hat{\mu}_m = 0$, and for $k = m+1, \ldots, k_n^{d_z}$, $c_1 k^{2m/d_z} \le \hat{\mu}_k \le c_2 k^{2m/d_z}$, where $c_1 > 0$ and $c_2 < \infty$ are independent of k and k_n.*

A.2. Proof of theorems

Proof of Theorem 1. Let $\mathbb{1}_n$ equal 1 if $\lambda_{\min}(\mathbf{P}^\top\mathbf{P}/n)$ is bounded away from 0 and $\lambda_{\max}(\mathbf{P}^\top\mathbf{P}/n)$ and $\lambda_{\max}\{\mathrm{E}(\mathbf{P}^\top\mathbf{P}/n) - \mathbf{P}^\top\mathbf{P}/n\}$ are bounded away from infinity, and equal 0 otherwise. Note that $\mathbb{1}_n \to_p 1$ by Lemma 1. Recall $\boldsymbol{\theta} = \mathbf{0}_{d_x}$ in this theorem. Define $\hat{\mathbf{G}} = (\hat{g}(\mathbf{z}_1), \ldots, \hat{g}(\mathbf{z}_n))^\top$ and $\hat{\mathbf{S}} = (\hat{s}(\mathbf{z}_1), \ldots, \hat{s}(\mathbf{z}_n))^\top$.

Observe that

$$\mathrm{E}(\mathbb{1}_n\|\hat{g} - g\|^2_{\hat{F},2,0}|\mathbf{Z}) \asymp \mathrm{E}\{\mathbb{1}_n\|\hat{\mathbf{G}} - \mathrm{E}(\hat{\mathbf{G}}|\mathbf{Z})\|^2|\mathbf{Z}\}/n + \|\mathrm{E}\{\mathbb{1}_n(\hat{\mathbf{G}} - \hat{\mathbf{S}})|\mathbf{Z}\}\|^2/n + \|\mathrm{E}\{\mathbb{1}_n(\hat{\mathbf{S}} - \mathbf{G})|\mathbf{Z}\}\|^2/n.$$

We first study each of these three terms. Let $\mathbf{B} = \mathbf{P}(\mathbf{P}^\top\mathbf{P})^{-1/2}\mathbf{U}$ and $\mathbf{M}$ as given in Lemma 5. We omit qualifiers such as "for all n large enough" or "with probability approaching 1" whenever possible to simplify the exposition.

For the first term (variance) we have:

$$\begin{aligned}
\mathrm{E}\{\mathbb{1}_n\|\hat{\mathbf{G}} - \mathrm{E}(\hat{\mathbf{G}}|\mathbf{Z})\|^2|\mathbf{Z}\}/n &= \mathrm{E}\{\mathbb{1}_n\|(\mathbf{I}_n + \lambda_n\mathbf{M}/n)^{-1}\mathbf{B}^\top(\mathbf{y} - \mathbf{G})\|^2|\mathbf{Z}\}/n \\
&\lesssim \mathbb{1}_n\frac{1}{n}\sum_{k=1}^{k_n^{d_z}}\frac{1}{\left(1 + \frac{\lambda_n}{n}\hat{\mu}_k\right)^2} \quad \text{(using } \mathrm{E}(y_i^2|\mathbf{z}_i) \text{ bounded)} \\
&\asymp \mathbb{1}_n\frac{1}{n}\sum_{k=1}^{k_n^{d_z}}\frac{1}{\left(1 + \frac{\lambda_n}{n}k^{2m/d_z}\right)^2} \quad \text{(using Lemma 6)},
\end{aligned}$$

where the first inequality is an equality under homoscedasticity (and the numerator would be replaced with σ^2_ε).

For the second term (penalization bias), letting $(b_1, \ldots, b_{k_n^{d_z}})^\top = \mathbf{B}^\top\mathbf{G}$, we have:

$$\begin{aligned}
\|\mathrm{E}\{\mathbb{1}_n(\hat{\mathbf{G}} - \hat{\mathbf{S}})|\mathbf{Z}\}\|^2/n &= \mathrm{E}[\mathbb{1}_n\|\{\mathbf{I}_n - (\mathbf{I}_n + \lambda_n\mathbf{M}/n)^{-1}\}\mathbf{B}^\top\mathbf{G}\|^2|\mathbf{Z}]/n \\
&\lesssim \mathbb{1}_n\frac{1}{n}\sum_{k=1}^{k_n^{d_z}} b_k^2\left(\frac{\frac{\lambda_n}{n}\hat{\mu}_k}{1 + \frac{\lambda_n}{n}\hat{\mu}_k}\right)^2 \\
&\asymp \mathbb{1}_n\frac{\lambda_n^2}{n}\sum_{k=1}^{k_n^{d_z}} b_k^2\frac{\left(\frac{1}{n}k_n^{2m/d_z}\right)}{\left(1 + \frac{\lambda_n}{n}k_n^{2m/d_z}\right)^2} \quad \text{(using Lemma 6)}.
\end{aligned}$$

For the third term (smoothing bias) we have $\|\mathrm{E}\{\mathbb{1}_n(\hat{\mathbf{S}} - \mathbf{G})|\mathbf{Z}\}\|^2/n \lesssim_p k_n^{-2r_g}$, using Lemmas 2 and 4.

Now, if $\lambda_n k_n^{2m}/n < 1$, we have:

$$\frac{1}{n}\sum_{k=1}^{k_n^{d_z}}\frac{1}{\left(1 + \frac{\lambda_n}{n}k_n^{2m/d_z}\right)^2} \leq \frac{k_n^{d_z}}{n},$$

and

$$\mathrm{E}\left\{\mathbb{1}_n\frac{\lambda_n^2}{n}b_k^2\sum_{k=1}^{k_n^{d_z}}\frac{\left(\frac{\hat{\mu}_k}{n}\right)\left(\frac{1}{n}k^{2m/d_z}\right)}{\left(1 + \frac{\lambda_n}{n}k^{2m/d_z}\right)^2}\right\} \leq \mathrm{E}\left(\mathbb{1}_n\frac{\lambda_n^2k_n^{2m}}{n^2}\sum_{k=1}^{k_n^{d_z}}b_k^2\frac{\hat{\mu}_k}{n}\right) \lesssim \frac{\lambda_n^2k_n^{2m}}{n^2},$$

because

$$\mathrm{E}\left(\mathbb{1}_n\sum_{k=1}^{k_n^{d_z}}b_k^2\hat{\mu}_k/n\right) = \mathrm{E}(\mathbb{1}_n\mathbf{G}^\top\mathbf{P}(\mathbf{P}^\top\mathbf{P})^{-1/2}\mathbf{U}\mathbf{M}\mathbf{U}^\top(\mathbf{P}^\top\mathbf{P})^{-1/2}\mathbf{P}^\top\mathbf{G}/n) \lesssim 1.$$

Therefore, in this case, we conclude that $\|\hat{g} - g\|^2_{\hat{F},2,0} \lesssim_p k_n^{d_z}/n + \lambda_n^2k_n^{2m}/n^2 + k_n^{-2r_g}$.

Next, if $\lambda_n k_n^{2m}/n \geq 1$ for all sufficiently large n, we have:

$$\begin{aligned}\frac{1}{n}\sum_{k=1}^{k_n^{d_z}} \frac{1}{\left(1+\frac{\lambda_n}{n}k^{2m/d_z}\right)^2} &= \frac{1}{n}\int_0^{k_n^{d_z}} \frac{\mathrm{d}v}{\left(1+\frac{\lambda_n}{n}v^{2m/d_z}\right)^2} + r_m \\ &= \frac{1}{n}\left(\frac{\lambda_n}{n}\right)^{-d_z/2m} \int_0^{(\lambda_n/n)^{d_z/2m}k_n^{d_z}} \frac{\mathrm{d}u}{(1+u^{2m/d_z})^2} + r_m \\ &\lesssim \frac{1}{n}\left(\frac{\lambda_n}{n}\right)^{-d_z/2m}\end{aligned}$$

with r_m the remainder term from the Euler–Maclaurin formula, using the substitution $u = (\lambda_n/n)^{d_z/2m}v$ (the integral is finite for $m > d_z/4$, even if $\lambda_n k_n^{2m}/n$ is unbounded), and

$$\mathrm{E}\left\{\frac{\lambda_n}{n}\sum_{k=1}^{k_n^{d_z}} b_k^2 \frac{\left(\frac{\hat{\mu}_k}{n}\right)\left(\frac{\lambda_n}{n}k^{2m/d_z}\right)}{\left(1+\frac{\lambda_n}{n}k^{2m/d_z}\right)^2}\right\} \leq \mathrm{E}\left(\frac{\lambda_n}{4n}\sum_{k=1}^{k_n^{d_z}} b_k^2\frac{\hat{\mu}_k}{n}\right) \lesssim \frac{\lambda_n}{n}$$

because $x(1+x)^{-2} \leq 0.25$ for $x \geq 1$. Therefore, in this case, $\|\hat{g}-g\|_{\hat{F},2,0}^2 \lesssim_p n^{(d_z-2m)/2m}/\lambda_n^{d_z/2m} + \lambda_n/n + k_n^{-2r_g}$.

This completes the proof for the case $|\boldsymbol{\ell}| = 0$, so we consider next the case $|\boldsymbol{\ell}| > 0$. Let $\boldsymbol{\ell} = (\ell_1, \ldots, \ell_{d_z})^\top \in \mathbb{Z}_*^{d_z}$, and define $\boldsymbol{\Delta}_{j,(\ell)} = \boldsymbol{\Delta}_{j,1}^\top \cdots \boldsymbol{\Delta}_{j,\ell}^\top$, where $\ell \in \mathbb{Z}_+$ and

$$\boldsymbol{\Delta}_{j,\eta} = (r-\eta)\begin{pmatrix} \frac{-1}{\kappa_{j,r+1}-\kappa_{j,\eta+1}} & 0 & 0 & \cdots & 0 \\ \frac{1}{\kappa_{j,r+1}-\kappa_{j,\eta+1}} & \frac{-1}{\kappa_{j,r+2}-\kappa_{j,\eta+2}} & 0 & \cdots & 0 \\ 0 & \frac{1}{\kappa_{j,r+2}-\kappa_{j,\eta+2}} & \frac{-1}{\kappa_{j,r+3}-\kappa_{j,\eta+3}} & \cdots & 0 \\ 0 & 0 & \frac{1}{\kappa_{j,r+3}-\kappa_{j,\eta+3}} & \ddots & 0 \\ \vdots & \vdots & \vdots & \ddots & \frac{-1}{\kappa_{j,k_n+r-\eta}-\kappa_{j,k_n}} \\ 0 & 0 & 0 & \cdots & \frac{1}{\kappa_{j,k_n+r-\eta}-\kappa_{j,k_n}} \end{pmatrix}$$

for $\eta = 1, \ldots, \ell$. As shown in [52] using [18] when $d_z = 1$,

$$\hat{s}^{(\ell)}(\mathbf{z}) = \partial^\ell \hat{s}(\mathbf{z}) = \partial^\ell\{\mathbf{p}_n(\mathbf{z})^\top\hat{\boldsymbol{\beta}}\} = \mathbf{p}_{n,(\ell)}(\mathbf{z})^\top\boldsymbol{\Delta}_{1,(\ell)}\hat{\boldsymbol{\beta}},$$

where $\mathbf{p}_{n,(\ell)}(\mathbf{z})$ is the vector of B-spline basis functions of order $r-\ell$. Thus, let $\mathbf{p}_{n,j,(\ell_j)}$ be the vector of B-spline basis functions in direction j of degree $r-\ell_j$, and let $\boldsymbol{\Delta}_{j,(\ell_j)}$ be the matrix given above using the knots in direction j. Also, let $\mathbf{p}_{n,(\boldsymbol{\ell})}(\mathbf{z}) = \mathbf{p}_{n,1,(\ell_1)}(\mathbf{z}) \otimes \cdots \otimes \mathbf{p}_{n,d_z,(\ell_{d_z})}(\mathbf{z})$, and $\mathbf{P}_{(\boldsymbol{\ell})}$ be the B-spline design matrix using B-splines of order $r-\ell_j$ in direction j and $\boldsymbol{\Delta}_{(\boldsymbol{\ell})} = \boldsymbol{\Delta}_{1,(\ell_1)} \otimes \cdots \otimes \boldsymbol{\Delta}_{d_z,(\ell_{d_z})}$.

Using this notation and for $\bar{\mathbf{s}}_g(\mathbf{z}) = \mathbf{p}_n(\mathbf{z})^\top\bar{\boldsymbol{\beta}}$ in Lemma 2, we have $\partial^{\boldsymbol{\ell}}\hat{g}(\mathbf{z}) - \partial^{\boldsymbol{\ell}}\bar{\mathbf{s}}_g(\mathbf{z}) = \mathbf{p}_{n,(\boldsymbol{\ell})}(\mathbf{z})^\top\boldsymbol{\Delta}_{(\boldsymbol{\ell})}(\hat{\boldsymbol{\beta}}-\bar{\boldsymbol{\beta}})$ and therefore, using Lemma 2,

$$\mathbb{1}_n\|\hat{g}-g\|_{\hat{F},2,\ell}^2 \lesssim_p \mathbb{1}_n \sup_{|\boldsymbol{\ell}|\leq\ell} \|(\hat{\boldsymbol{\beta}}-\bar{\boldsymbol{\beta}})^\top\boldsymbol{\Delta}_{(\boldsymbol{\ell})}^\top\mathbf{P}_{(\boldsymbol{\ell})}^\top\mathbf{P}_{(\boldsymbol{\ell})}\boldsymbol{\Delta}_{(\boldsymbol{\ell})}(\hat{\boldsymbol{\beta}}-\bar{\boldsymbol{\beta}})/n\|^2 + k_n^{-2(r_g-\ell)}.$$

Finally, using Lemma 1, $\mathbb{1}_n\|\hat{\boldsymbol{\beta}}-\bar{\boldsymbol{\beta}}\|\mathbb{1}_n \lesssim_p \|\hat{g}-\bar{s}_g\|_{\hat{F},2,0}$ and $\lambda_{\max}(\mathbf{P}_{(\boldsymbol{\ell})}^\top\mathbf{P}_{(\boldsymbol{\ell})}/n)$ bounded in probability for all $\boldsymbol{\ell} \in \mathbb{Z}_+^{d_z}$. Therefore, because $\sup_{|\boldsymbol{\ell}|\leq\ell}\|\mathbb{1}_n\boldsymbol{\Delta}_{(\boldsymbol{\ell})}\| \lesssim k_n^\ell$, collecting the results above the proof is complete.

Proof of Theorem 2. Observe that $\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}) = (\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}/n)^{-1}\tilde{\mathbf{X}}^\top(\mathbf{I}_n-\mathbf{R})\boldsymbol{\varepsilon}/\sqrt{n} + (\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}/n)^{-1}\tilde{\mathbf{X}}^\top(\mathbf{I}_n-\mathbf{R})\mathbf{G}/\sqrt{n}$, and let $\boldsymbol{\nu} = (\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_n)^\top$ and $\mathbf{H} = (\mathbf{h}(\mathbf{z}_1), \ldots, \mathbf{h}(\mathbf{z}_n))^\top$. Also, let $\boldsymbol{\nu}^{j\top}$ be the jth row of $\boldsymbol{\nu}$ and $\mathbf{h}^{j\top}$ be the jth row of $\mathbf{H}$, and let $\mathbf{R}_0$ equal $\mathbf{R}$ when $\lambda_n = 0$.

First, we show that $\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}/n = \mathrm{E}(\boldsymbol{\nu}_i\boldsymbol{\nu}_i^\top) + \mathrm{o_p}(1)$. Note that

$$\begin{aligned}\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}/n = {} & \boldsymbol{\nu}^\top(\mathbf{I}_n-\mathbf{R})^\top(\mathbf{I}_n-\mathbf{R})\boldsymbol{\nu}/n + \mathbf{H}^\top(\mathbf{I}_n-\mathbf{R})^\top(\mathbf{I}_n-\mathbf{R})\boldsymbol{\nu}/n \\ & + \boldsymbol{\nu}^\top(\mathbf{I}_n-\mathbf{R})^\top(\mathbf{I}_n-\mathbf{R})\mathbf{H}/n + \mathbf{H}^\top(\mathbf{I}_n-\mathbf{R})^\top(\mathbf{I}_n-\mathbf{R})\mathbf{H}/n,\end{aligned}$$

where (i) $\boldsymbol{\nu}^\top(\mathbf{I}_n - \mathbf{R})^\top(\mathbf{I}_n - \mathbf{R})\boldsymbol{\nu}/n = \boldsymbol{\nu}^\top\boldsymbol{\nu} + \mathrm{o_p}(1) = \mathrm{E}(\boldsymbol{\nu}_i\boldsymbol{\nu}_i^\top) + \mathrm{o_p}(1)$, (ii) $\mathbf{H}^\top(\mathbf{I}_n - \mathbf{R})^\top(\mathbf{I}_n - \mathbf{R})\boldsymbol{\nu}/n = \mathrm{o_p}(1)$, (iii) $\boldsymbol{\nu}^\top(\mathbf{I}_n - \mathbf{R})^\top(\mathbf{I}_n - \mathbf{R})\mathbf{H}/n = \mathrm{o_p}(1)$, and (iv) $\mathbf{H}^\top(\mathbf{I}_n - \mathbf{R})^\top(\mathbf{I}_n - \mathbf{R})\mathbf{H}/n = \mathrm{o_p}(1)$. More specifically, for (i), note that

$$\mathrm{E}(\mathbb{1}_n\boldsymbol{\nu}^{j\top}\mathbf{R}\boldsymbol{\nu}^j)/n = \mathrm{E}[\mathbb{1}_n\mathrm{tr}\{\mathbf{R}\mathrm{E}(\boldsymbol{\nu}^j\boldsymbol{\nu}^{j\top}|\mathbf{Z})\}]/n = \frac{\mathbb{1}_n}{n}\sum_{k=1}^{k_n^{d_z}}\frac{1}{1+\frac{\lambda_n}{n}\hat{\mu}_k} \leq k_n^{d_z}/n,$$

and

$$\mathrm{E}(\mathbb{1}_n\boldsymbol{\nu}^{j\top}\mathbf{R}\mathbf{R}^\top\boldsymbol{\nu}^j)/n = \mathrm{E}[\mathbb{1}_n\mathrm{Tr}\{\mathbf{R}\mathbf{R}^\top\mathrm{E}(\boldsymbol{\nu}^j\boldsymbol{\nu}^{j\top}|\mathbf{Z})\}]/n = \frac{\mathbb{1}_n}{n}\sum_{k=1}^{k_n^{d_z}}\frac{1}{\left(1+\frac{\lambda_n}{n}\hat{\mu}_k\right)^2} \leq k_n^{d_z}/n.$$

So $\boldsymbol{\nu}^\top\mathbf{R}\boldsymbol{\nu}/n = \mathrm{O_p}(k_n^{d_z}/n)$ and $\boldsymbol{\nu}^\top\mathbf{R}\mathbf{R}^\top\boldsymbol{\nu}/n = \mathrm{O_p}(k_n^{d_z}/n)$, and thus $\boldsymbol{\nu}^\top(\mathbf{I}_n-\mathbf{R})^\top(\mathbf{I}_n-\mathbf{R})\boldsymbol{\nu}/n = \mathrm{E}(\upsilon_i\upsilon_i^\top|z_i) + \mathrm{o_p}(1)$. For (iv), define $\hat{h}_0^j = \hat{h}^j$ when $\lambda_n = 0$. Then

$$\mathbf{h}^{j\top}\{(\mathbf{I}_n - \mathbf{R}) - (\mathbf{I}_n - \mathbf{R}_0)\}^\top\{(\mathbf{I}_n - \mathbf{R}) - (\mathbf{I}_n - \mathbf{R}_0)\}\mathbf{h}^j/n = \sum_{i=1}^n\{\mathrm{E}(\hat{h}_0^j(\mathbf{z}_i) - \hat{h}_j(\mathbf{z}_i)|\mathbf{Z})\}^2/n,$$

which is $\mathrm{O_p}(\lambda_n k_n^m/n)$ if $\lambda_n k_n^{2m}/n < 1$ and $\mathrm{O_p}(\sqrt{\lambda_n/n})$ if $\lambda_n k_n^{2m}/n \geq 1$. Also, $\mathbf{h}^{j\top}(\mathbf{I}_n - \mathbf{R}_0)(\mathbf{I}_n - \mathbf{R}_0)^\top\mathbf{h}^j/n = \mathrm{O_p}(k_n^{-2r_h})$, so $\mathbf{H}^\top(\mathbf{I}_n - \mathbf{R})(\mathbf{I}_n - \mathbf{R})^\top\mathbf{H}/n = \mathrm{o_p}(1)$. For (ii) and (iii),

$$\begin{aligned}\mathbf{h}^{j\top}(\mathbf{I}_n - \mathbf{R})(\mathbf{I}_n - \mathbf{R})^\top\boldsymbol{\nu}^j/n &= \boldsymbol{\nu}^{j\top}(\mathbf{I}_n - \mathbf{R})(\mathbf{I}_n - \mathbf{R})^\top\mathbf{h}^j/n \\ &\leq \{\mathbf{h}^{j\top}(\mathbf{I}_n - \mathbf{R})(\mathbf{I}_n - \mathbf{R})^\top\mathbf{h}^j/n \cdot \boldsymbol{\nu}^{j\top}(\mathbf{I}_n - \mathbf{R})(\mathbf{I}_n - \mathbf{R})^\top\boldsymbol{\nu}^j/n\}^{1/2} \\ &= \mathrm{o_p}(1).\end{aligned}$$

Second, we show that $\tilde{\mathbf{X}}^\top(\mathbf{I}_n - \mathbf{R})\boldsymbol{\varepsilon}/\sqrt{n} = \boldsymbol{\nu}^\top\boldsymbol{\varepsilon}/\sqrt{n} + \mathrm{o_p}(1)$. Note that

$$\tilde{\mathbf{X}}^\top(\mathbf{I}_n - \mathbf{R})\boldsymbol{\varepsilon}/\sqrt{n} = \boldsymbol{\nu}^\top(\mathbf{I}_n - \mathbf{R})^\top(\mathbf{I}_n - \mathbf{R})\boldsymbol{\varepsilon}/\sqrt{n} + \mathbf{H}^\top(\mathbf{I}_n - \mathbf{R})^\top(\mathbf{I}_n - \mathbf{R})\boldsymbol{\varepsilon}/\sqrt{n},$$

where (i) $\boldsymbol{\nu}^\top(\mathbf{I}_n - \mathbf{R})^\top(\mathbf{I}_n - \mathbf{R})\boldsymbol{\varepsilon}/\sqrt{n} = \boldsymbol{\nu}^\top\boldsymbol{\varepsilon}/\sqrt{n} + \mathrm{o_p}(1)$ and (ii) $\mathbf{H}^\top(\mathbf{I}_n - \mathbf{R})^\top(\mathbf{I}_n - \mathbf{R})\boldsymbol{\varepsilon}/n = \mathrm{o_p}(1)$. For (i),

$$\mathrm{E}(\mathbb{1}_n\boldsymbol{\nu}^{j\top}\mathbf{R}\boldsymbol{\varepsilon}/\sqrt{n})^2 \lesssim \mathrm{E}(\mathbb{1}_n\boldsymbol{\nu}^{j\top}\mathbf{R}^\top\mathbf{R}\boldsymbol{\nu}^j)/n \lesssim \lambda_{\min}^{-2}(\mathbf{I} + \lambda_n\mathbf{M}/n)\mathrm{E}(\mathbb{1}_n\boldsymbol{\nu}^{j\top}\mathbf{R}\boldsymbol{\nu}^j)/n \lesssim k_n^{d_z}/n.$$

Also, since $\lambda_{\min}(\mathbf{I} + \lambda_n\mathbf{M}/n) \geq 1$, $\boldsymbol{\nu}^\top\mathbf{R}\mathbf{R}^\top\boldsymbol{\varepsilon}/\sqrt{n} \lesssim \boldsymbol{\nu}^\top\mathbf{R}\boldsymbol{\varepsilon}/\sqrt{n} = \mathrm{o_p}(1)$. For (ii), since the kth diagonal element of $\{\mathbf{I} - (\mathbf{I} + \lambda_n\mathbf{M}/n)^{-1}\}^2$ is $[(\lambda\hat{\mu}_k/n)/\{1 - (\lambda\hat{\mu}_k/n)\}]^2 \leq 1$,

$$\mathrm{E}[\mathbb{1}_n[\mathbf{h}^{j\top}\{(\mathbf{I}_n - \mathbf{R}) - (\mathbf{I}_n - \mathbf{R}_0)\}\{(\mathbf{I}_n - \mathbf{R}) - (\mathbf{I}_n - \mathbf{R}_0)\}^\top\boldsymbol{\varepsilon}/\sqrt{n}]^2|\mathbf{X}, \mathbf{Z}] \lesssim \mathrm{E}(\hat{\mathbf{h}}_0^j - \hat{\mathbf{h}}^j|\mathbf{Z})^\top\mathrm{E}(\hat{\mathbf{h}}_0^j - \hat{\mathbf{h}}^j|\mathbf{Z}),$$

so $[\mathbf{H}^\top\{(\mathbf{I}_n - \mathbf{R}) - (\mathbf{I}_n - \mathbf{R}_0)\}\{(\mathbf{I}_n - \mathbf{R}) - (\mathbf{I}_n - \mathbf{R}_0)\}^\top\boldsymbol{\varepsilon}/\sqrt{n}]^2$ is $\mathrm{O_p}(\lambda_n k_n^m/d)$ if $\lambda k^{2m}/d_z < 1$ and $\mathrm{O_p}(\sqrt{\lambda_n/n})$ is $\lambda k_n^{2m}/d_z \geq 1$. Also, $\mathrm{E}[\mathbb{1}_n\{\mathbf{h}^{j\top}(\mathbf{I}_n - \mathbf{R}_0)(\mathbf{I}_n - \mathbf{R}_0)^\top\boldsymbol{\varepsilon}/\sqrt{n}\}^2|\mathbf{X}, \mathbf{Z}] \lesssim \mathbf{h}^{j\top}(\mathbf{I}_n - \mathbf{R}_0)(\mathbf{I}_n - \mathbf{R}_0)^\top\mathbf{h}^j/n$, so $\mathbf{H}^\top(\mathbf{I}_n - \mathbf{R})(\mathbf{I}_n - \mathbf{R})^\top\boldsymbol{\varepsilon} = \mathrm{O_p}(k_n^{-2r_h})$.

Third, we show that $\tilde{\mathbf{X}}^\top(\mathbf{I}_n - \mathbf{R})\mathbf{G}/\sqrt{n} = \mathrm{o_p}(1)$ because (i) $\boldsymbol{\nu}^\top(\mathbf{I}_n - \mathbf{R})^\top(\mathbf{I}_n - \mathbf{R})\mathbf{G}/\sqrt{n} = \mathrm{o_p}(1)$ and (ii) $\mathbf{H}^\top(\mathbf{I}_n - \mathbf{R})^\top(\mathbf{I}_n - \mathbf{R})\mathbf{G}/\sqrt{n} = \mathrm{o_p}(1)$. Specifically, for (i), we have $\boldsymbol{\nu}^{j\top}(\mathbf{I}_n - \mathbf{R})(\mathbf{I}_n - \mathbf{R})^\top\mathbf{G}/\sqrt{n} \leq \sqrt{n}\{\boldsymbol{\nu}^{j\top}(\mathbf{I}_n - \mathbf{R})(\mathbf{I}_n - \mathbf{R})^\top\boldsymbol{\nu}^j/n\}^{1/2}\{\mathbf{G}^\top(\mathbf{I}_n - \mathbf{R})(\mathbf{I}_n - \mathbf{R})^\top\mathbf{G}/n\}^{1/2} = \mathrm{o_p}(1)$, and for (ii), $\mathbf{h}^{j\top}(\mathbf{I}_n - \mathbf{R})^\top(\mathbf{I}_n - \mathbf{R})\mathbf{G}/\sqrt{n} = \sqrt{n}\{\mathbf{h}^{j\top}(\mathbf{I}_n - \mathbf{R})^\top(\mathbf{I}_n - \mathbf{R})\mathbf{h}^j/n \cdot \mathbf{G}^\top(\mathbf{I}_n - \mathbf{R})^\top(\mathbf{I}_n - \mathbf{R})\mathbf{G}/n\}^{1/2}$, using the assumptions in the statement of Theorem 2.

Finally, note that $\mathrm{E}(\boldsymbol{\nu}_i\boldsymbol{\varepsilon}_i) = \mathrm{E}\{\boldsymbol{\nu}_i\mathrm{E}(\boldsymbol{\varepsilon}_i|\mathbf{x}_i, \mathbf{z}_i)\} = 0$. Then $\boldsymbol{\Omega} \equiv \mathrm{V}(\boldsymbol{\nu}^\top\boldsymbol{\varepsilon}/\sqrt{n}) = \mathrm{E}(\boldsymbol{\nu}_i\boldsymbol{\nu}_i^\top\varepsilon_i^2)$, so by the Central Limit Theorem, $\boldsymbol{\Omega}^{-1/2}\boldsymbol{\nu}^\top\boldsymbol{\varepsilon}/\sqrt{n} \to_d \mathcal{N}(0, 1)$. So $\mathbf{V}_n^{-1/2}\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \to_d \mathcal{N}(0, 1)$.

Last, we show that $\hat{\boldsymbol{\Omega}}_n = \boldsymbol{\Omega} + \mathrm{o_p}(1)$. Note that

$$\begin{aligned}\mathbb{1}_n\sum_{j=1}^n(\mathbf{R})_{ij}^2 &= \mathbb{1}_n\sum_{j=1}^n\mathbf{p}_n(\mathbf{z}_i)^\top(\mathbf{P}^\top\mathbf{P} + \lambda_n\mathbf{D})^{-1}\mathbf{p}_n(\mathbf{z}_j)\mathbf{p}_n(\mathbf{z}_j)^\top(\mathbf{P}^\top\mathbf{P} + \lambda_n\mathbf{D})^{-1}\mathbf{p}_n(\mathbf{z}_i) \\ &= \mathbb{1}_n\mathbf{p}_n(\mathbf{z}_i)^\top(\mathbf{P}^\top\mathbf{P})^{-1/2}\mathbf{U}(\mathbf{I} + \lambda_n\mathbf{M}/n)^{-2}\mathbf{U}^\top(\mathbf{P}^\top\mathbf{P})^{-1/2}\mathbf{p}_n(\mathbf{z}_i) \\ &\leq \mathbb{1}_n\lambda_{\min}^{-1}(\mathbf{P}^\top\mathbf{P}/n)\lambda_{\min}^{-2}(\mathbf{I} + \lambda_n\mathbf{M}/n)\mathbf{p}_n(\mathbf{z}_i)^\top\mathbf{p}_n(\mathbf{z}_i)/n \\ &\lesssim k_n^{d_z}/n.\end{aligned}$$

Let N_δ be the number of observations lying in a hyper-interval δ, then $\mathrm{E}(N_\delta) = n\int_\delta f(\mathbf{z})d\mathbf{z}/\int_{[0,1]^d} f(\mathbf{z})d\mathbf{z} \lesssim n/k_n^{d_z}$. So by Markov's inequality, $N_\delta = \mathrm{O_p}(n/k_n^{d_z})$ for all δ, and thus $\sum_{i=1}^n p_k(\mathbf{z}_i) \lesssim_p n/k_n^{d_z/2}$. So

$$\begin{aligned}\mathbb{1}_n\sum_{j=1}^n(\mathbf{R})_{ij} &\le \mathbb{1}_n\lambda_{\min}^{-1}(\mathbf{P}^\top\mathbf{P}/n)\lambda_{\min}^{-1}(\mathbf{I}+\lambda_n\mathbf{M}/n)\mathbf{p}_n(\mathbf{z}_i)^\top\sum_{j=1}^n\mathbf{p}_n(\mathbf{z}_j)/n\\ &\asymp_p k_n^{-d_z/2}\sum_{k=1}^{k_n^{d_z}}p_k(\mathbf{z}_i)\asymp_p 1.\end{aligned}$$

Also, $\mathbb{1}_n(\mathbf{R})_{ii} \ge \mathbb{1}_n\lambda_{\max}^{-1}(\mathbf{P}^\top\mathbf{P}/n)\lambda_{\max}^{-1}(\mathbf{I}+\lambda_n\mathbf{M}/n)\mathbf{p}_n(\mathbf{z}_i)^\top\mathbf{p}_n(\mathbf{z}_i)/n \ge 0$, so $\mathbb{1}_n|(\mathbf{R})_{ii}| = \mathbb{1}_n(\mathbf{R})_{ii}$, and $\mathbb{1}_n(\mathbf{R})_{ii} \lesssim \mathbf{p}_n(\mathbf{z}_i)^\top\mathbf{p}_n(\mathbf{z}_i)/n = k_n^{d_z}/n$.

Then defining $\mathbf{T} = \mathbf{I}_n - \mathbf{R}$,

$$\sum_{j=1}^n(\mathbf{T})_{ij}^2 = \sum_{j=1,j\ne i}^n(\mathbf{R})_{ij}^2 + \{1-(\mathbf{R})_{ii}\}^2 \le \sum_{j=1}^n(\mathbf{R})_{ij}^2 + (\mathbf{R})_{ii}^2 + 1 \lesssim_p 1,$$

and

$$\begin{aligned}\Big\{\sum_{j=1}^n(\mathbf{T})_{ij}\Big\}^2 &= \Big\{\sum_{j=1,j\ne i}^n(\mathbf{R})_{ij} + 1 - (\mathbf{R})_{ii}\Big\}^2\\ &\lesssim \Big\{\sum_{j=1}^n(\mathbf{R})_{ij}\Big\}^2 + \{1-(\mathbf{R})_{ii}\}^2 \lesssim_p 1.\end{aligned}$$

Now letting $\bar{\boldsymbol{\varepsilon}} = \mathbf{T}\boldsymbol{\varepsilon}$ and similarly for $\bar{\mathbf{X}}$ and $\bar{\mathbf{y}}$,

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}} - \hat{\mathbf{G}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}} - \mathbf{P}(\mathbf{P}^\top\mathbf{P}+\lambda_n\mathbf{D})^{-1}\mathbf{P}^\top(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\theta}}) = \mathbf{T}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\theta}}) = \bar{\boldsymbol{\varepsilon}} + \bar{\mathbf{X}}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}) + \bar{\mathbf{G}}.$$

So $\hat{\boldsymbol{\varepsilon}}_i = \bar{\boldsymbol{\varepsilon}}_i + \bar{\mathbf{X}}_i^\top(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}) + \bar{\mathbf{G}}_i$, where $\bar{\mathbf{G}}_i$ is the ith element of $\bar{G}$ and $\bar{\mathbf{X}}_i^\top$ is the ith row of $\bar{\mathbf{X}}$. Then $\hat{\boldsymbol{\varepsilon}}_i^2 = \bar{\boldsymbol{\varepsilon}}_i^2 + 2\bar{\boldsymbol{\varepsilon}}_i(\hat{\boldsymbol{\varepsilon}}_i - \bar{\boldsymbol{\varepsilon}}_i) + (\hat{\boldsymbol{\varepsilon}}_i - \bar{\boldsymbol{\varepsilon}}_i)^2 = \bar{\boldsymbol{\varepsilon}}_i^2 + 2\bar{\boldsymbol{\varepsilon}}_i\{\bar{\mathbf{X}}_i^\top(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}) + \bar{\mathbf{G}}_i\} + \{\bar{\mathbf{X}}_i^\top(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}) + \bar{\mathbf{G}}_i\}^2$. Consider $\bar{\mathbf{X}}_i^\top(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})$, and note that $\sum_{j=1}^n\mathbf{T}_{ij}\mathbf{X}_{\ell j} \lesssim_p \sum_{j=1}^n\mathbf{T}_{ij} \lesssim_p 1$. Since $\hat{\boldsymbol{\theta}}-\boldsymbol{\theta} = \mathrm{O_p}(n^{-1/2})$,

$$\bar{\mathbf{X}}_i^\top(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}) = \mathbf{T}_i^\top\mathbf{X}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}) = \sum_{\ell=1}^d\sum_{j=1}^n\mathbf{T}_{ij}\mathbf{X}_{\ell j}(\hat{\boldsymbol{\theta}}_\ell - \boldsymbol{\theta}_\ell) \lesssim_p n^{-1/2},$$

where $\mathbf{T}_i^\top$ is the ith row of $\mathbf{T}$. Also,

$$\bar{\mathbf{G}}_i = \mathbf{T}_i^\top\mathbf{G} = \sum_{j=1}^n\mathbf{T}_{ij}g(\mathbf{z}_j) \le \sup_{z\in[0,1]^d}g(\mathbf{z})\sum_{j=1}^n\mathbf{T}_{ij} \lesssim_p 1.$$

So $\bar{\mathbf{X}}_i^\top(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}) + \bar{\mathbf{G}}_i \lesssim_p 1$. Finally, $\mathrm{E}(\bar{\varepsilon}_i^2|\mathbf{X},\mathbf{Z}) = \sum_{j=1}^n T_{ij}^2\mathrm{E}(\varepsilon_j^2|\mathbf{X},\mathbf{Z}) \lesssim 1$. So $\bar{\boldsymbol{\varepsilon}}_i^2 \lesssim_p 1$ and thus $\hat{\boldsymbol{\varepsilon}}_i^2 \lesssim_p 1$.

Now letting $\hat{\boldsymbol{\Omega}}_0 = \hat{\boldsymbol{\Omega}}$ and $\mathbf{T}_0 = \mathbf{T}$ when $\lambda = 0$, consider

$$\begin{aligned}\hat{\boldsymbol{\Omega}} - \hat{\boldsymbol{\Omega}}_0 &= (\mathbf{X}^\top\mathbf{T}\mathbf{T}^\top\hat{\boldsymbol{\Sigma}}\mathbf{T}\mathbf{T}^\top\mathbf{X} - \mathbf{X}^\top\mathbf{T}_0\mathbf{T}_0^\top\hat{\boldsymbol{\Sigma}}\mathbf{T}_0\mathbf{T}_0^\top\mathbf{X})/n\\ &= \mathbf{X}^\top(\mathbf{T}-\mathbf{T}_0)(\mathbf{T}-\mathbf{T}_0)^\top\hat{\boldsymbol{\Sigma}}(\mathbf{T}-\mathbf{T}_0)(\mathbf{T}-\mathbf{T}_0)^\top\mathbf{X}/n + \mathbf{X}^\top(\mathbf{T}-\mathbf{T}_0)(\mathbf{T}-\mathbf{T}_0)^\top\hat{\boldsymbol{\Sigma}}(\mathbf{T}-\mathbf{T}_0)\mathbf{T}_0^\top\mathbf{X}/n\\ &\quad + \mathbf{X}^\top(\mathbf{T}-\mathbf{T}_0)(\mathbf{T}-\mathbf{T}_0)^\top\hat{\boldsymbol{\Sigma}}\mathbf{T}_0(\mathbf{T}-\mathbf{T}_0)^\top\mathbf{X}/n + \mathbf{X}^\top(\mathbf{T}-\mathbf{T}_0)(\mathbf{T}-\mathbf{T}_0)^\top\hat{\boldsymbol{\Sigma}}\mathbf{T}_0\mathbf{T}_0^\top\mathbf{X}/n\\ &\quad + \mathbf{X}^\top(\mathbf{T}-\mathbf{T}_0)\mathbf{T}_0^\top\hat{\boldsymbol{\Sigma}}(\mathbf{T}-\mathbf{T}_0)(\mathbf{T}-\mathbf{T}_0)^\top\mathbf{X}/n + \mathbf{X}^\top(\mathbf{T}-\mathbf{T}_0)\mathbf{T}_0^\top\hat{\boldsymbol{\Sigma}}(\mathbf{T}-\mathbf{T}_0)\mathbf{T}_0^\top\mathbf{X}/n\\ &\quad + \mathbf{X}^\top(\mathbf{T}-\mathbf{T}_0)\mathbf{T}_0^\top\hat{\boldsymbol{\Sigma}}\mathbf{T}_0(\mathbf{T}-\mathbf{T}_0)^\top\mathbf{X}/n + \mathbf{X}^\top(\mathbf{T}-\mathbf{T}_0)\mathbf{T}_0^\top\hat{\boldsymbol{\Sigma}}\mathbf{T}_0\mathbf{T}_0^\top\mathbf{X}/n\\ &\quad + \mathbf{X}^\top\mathbf{T}_0(\mathbf{T}-\mathbf{T}_0)^\top\hat{\boldsymbol{\Sigma}}(\mathbf{T}-\mathbf{T}_0)(\mathbf{T}-\mathbf{T}_0)^\top\mathbf{X}/n + \mathbf{X}^\top\mathbf{T}_0(\mathbf{T}-\mathbf{T}_0)^\top\hat{\boldsymbol{\Sigma}}(\mathbf{T}-\mathbf{T}_0)\mathbf{T}_0^\top\mathbf{X}/n\\ &\quad + \mathbf{X}^\top\mathbf{T}_0(\mathbf{T}-\mathbf{T}_0)^\top\hat{\boldsymbol{\Sigma}}\mathbf{T}_0(\mathbf{T}-\mathbf{T}_0)^\top\mathbf{X}/n + \mathbf{X}^\top\mathbf{T}_0(\mathbf{T}-\mathbf{T}_0)^\top\hat{\boldsymbol{\Sigma}}\mathbf{T}_0\mathbf{T}_0^\top\mathbf{X}/n\\ &\quad + \mathbf{X}^\top\mathbf{T}_0\mathbf{T}_0^\top\hat{\boldsymbol{\Sigma}}(\mathbf{T}-\mathbf{T}_0)(\mathbf{T}-\mathbf{T}_0)^\top\mathbf{X}/n + \mathbf{X}^\top\mathbf{T}_0\mathbf{T}_0^\top\hat{\boldsymbol{\Sigma}}(\mathbf{T}-\mathbf{T}_0)\mathbf{T}_0^\top\mathbf{X}/n + \mathbf{X}^\top\mathbf{T}_0\mathbf{T}_0^\top\hat{\boldsymbol{\Sigma}}\mathbf{T}_0(\mathbf{T}-\mathbf{T}_0)^\top\mathbf{X}/n.\end{aligned}$$

As above, since the elements of the diagonal matrix $\mathbf{T}-\mathbf{T}_0$ have absolute value less than one, the elements of $(\mathbf{T}-\mathbf{T}_0)(\mathbf{T}-\mathbf{T}_0)^\top$ are less than (the absolute value of) the elements of $\mathbf{T}-\mathbf{T}_0$. Since $\hat{\boldsymbol{\varepsilon}}_i^2 = \mathrm{O_p}(1)$, $\mathbf{X}^{j\top}(\mathbf{T}-\mathbf{T}_0)\hat{\boldsymbol{\Sigma}}(\mathbf{T}-\mathbf{T}_0)\mathbf{X}^j/n \lesssim_p n^{-1}\sum_{i=1}^n\{\hat{\mathbf{h}}_0^j(\mathbf{z}_i) - \hat{\mathbf{h}}^j(\mathbf{z}_i)\}^2 = \mathrm{o_p}(1)$. So $\mathbf{X}^\top(\mathbf{T}-\mathbf{T}_0)(\mathbf{T}-\mathbf{T}_0)^\top\hat{\boldsymbol{\Sigma}}(\mathbf{T}-\mathbf{T}_0)^\top(\mathbf{T}-\mathbf{T}_0)\mathbf{X}/n = \mathrm{o_p}(1)$. Then using the Cauchy–Schwarz inequality and the fact that $\hat{\boldsymbol{\Omega}}_0 = \mathrm{O_p}(1)$, we see that each term above is $\mathrm{o_p}(1)$, and thus $\hat{\boldsymbol{\Omega}} = \hat{\boldsymbol{\Omega}}_0 + \mathrm{o_p}(1)$. Then since $\hat{\boldsymbol{\Omega}}_0 = \boldsymbol{\Omega} + \mathrm{o_p}(1)$ as shown in [10], we obtain $\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega} = (\hat{\boldsymbol{\Omega}} - \hat{\boldsymbol{\Omega}}_0) + (\hat{\boldsymbol{\Omega}}_0 - \boldsymbol{\Omega}) = \mathrm{o_p}(1)$.

Proof of Theorem 3. First, note that in case (ii), since $r_g \geq m+1$, $\lambda_n k_n^{2r_g}/n \geq 1$. So

$$nk_n^{-d_z-2r_g} = \frac{\lambda_n/k_n^{d_z}}{\lambda_n k_n^{2r_g}/n} \to 0.$$

In case (iii),

$$\frac{\lambda_n^2 k_n^{4m-d_z-2r_g}}{n} = \frac{\lambda_n^3 k_n^{4m-d_z}/n^2}{\lambda_n k_n^{2r_g}/n} \to 0.$$

Consider the variance of $\partial^{\boldsymbol{\ell}}\hat{g}(\mathbf{z})$, and observe that $\boldsymbol{\Delta}_{j,(\ell)}$ is an upper triangular band matrix with the last η rows missing and that the kth diagonal entry of $\boldsymbol{\Delta}_{j,(\ell)}$ is $\prod_{s=1}^{\ell}\frac{s-r}{\kappa_{j,k}-\kappa_{j,k-r+s}}$. Define $\mathbf{p}_{j,n,(\ell_j)}(\mathbf{z})$ to be the vector of B-spline basis functions in direction j of order $r-\ell_j$. Then

$$\begin{aligned}
\mathbf{p}_{n,(\boldsymbol{\ell})}(\mathbf{z})^\top\boldsymbol{\Delta}_{(\boldsymbol{\ell})}\boldsymbol{\Delta}_{(\boldsymbol{\ell})}\mathbf{p}_{n,(\boldsymbol{\ell})}(\mathbf{z}) &= \prod_{j=1}^{d_z}\mathbf{p}_{j,n,(\ell_j)}(z_j)^\top\boldsymbol{\Delta}_{j,(\ell_j)}\boldsymbol{\Delta}_{j,(\ell_j)}^\top\mathbf{p}_{j,n,(\ell_j)}(z_j)\\
&\gtrsim \prod_{j=1}^{d_z}k_n\left(\prod_{s=1}^{\ell_j}\frac{1}{\kappa_{j,r}-\kappa_{j,s}}\right)^2\\
&\asymp k_n^{2|\boldsymbol{\ell}|+d_z}.
\end{aligned}$$

So defining $\mathbf{Q} = \mathrm{E}(\mathbf{P}^\top\mathbf{P}/n)$,

$$\begin{aligned}
\mathbf{W}_{n,\boldsymbol{\ell}}(\mathbf{z}) &= \mathrm{V}\{\partial^{\boldsymbol{\ell}}\mathbf{p}_n(\mathbf{z})^\top(\mathbf{P}^\top\mathbf{P}+\lambda_n\mathbf{D})^{-1}\mathbf{P}^\top\boldsymbol{\varepsilon}|\mathbf{X},\mathbf{Z}\}\\
&= \mathbf{p}_{n,(\boldsymbol{\ell})}(\mathbf{z})^\top\boldsymbol{\Delta}_{(\boldsymbol{\ell})}(\mathbf{Q}+\lambda_n\mathbf{D}/n)^{-1}\mathbf{P}^\top\mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top|\mathbf{X},\mathbf{Z})\mathbf{P}(\mathbf{Q}+\lambda_n\mathbf{D}/n)^{-1}\boldsymbol{\Delta}_{(\boldsymbol{\ell})}^\top\mathbf{p}_{n,(\boldsymbol{\ell})}(\mathbf{z})/n^2 + \mathrm{o_p}(1)\\
&\gtrsim_p \lambda_{\max}^{-2}(\mathbf{I}+\lambda\mathbf{M}/n)\mathbf{p}_{n,(\boldsymbol{\ell})}(\mathbf{z})^\top\boldsymbol{\Delta}_{(\boldsymbol{\ell})}\boldsymbol{\Delta}_{(\boldsymbol{\ell})}^\top\mathbf{p}_{n,(\boldsymbol{\ell})}(\mathbf{z})/n\\
&\asymp (1+\lambda_n k_n^{2m}/n)^{-2}k_n^{2|\boldsymbol{\ell}|+d_z}/n.
\end{aligned}$$

If $\lambda_n k_n^{2m}/n < 1$, then $(1+\lambda_n k_n^{2m}/n)^{-2} \asymp 1$, so $\mathbf{W}_{n,\boldsymbol{\ell}}(\mathbf{z}) \gtrsim_p k_n^{2|\boldsymbol{\ell}|+d_z}/n$. If $1 \leq \lambda_n k_n^{2m}/n < \infty$, then $\mathbf{W}_{n,\boldsymbol{\ell}}(\mathbf{z}) \gtrsim_p k_n^{2|\boldsymbol{\ell}|+d_z}/n \asymp k_n^{2|\boldsymbol{\ell}|}n^{(d_z-2m)/2m}/\lambda^{d_z/2m}$; and if $\lambda_n k_n^{2m}/n$ is unbounded, then $(1+\lambda_n k_n^{2m}/n)^{-2} \asymp \lambda_n k_n^{2m}/n^{-2}$, so $\mathbf{W}_{n,\boldsymbol{\ell}}(\mathbf{z}) \gtrsim_p k_n^{2|\boldsymbol{\ell}|}(n/\lambda_n k_n^{2m})^2(k_n^{d_z}/n) \gtrsim k_n^{2|\boldsymbol{\ell}|}(n/\lambda_n k_n^{2m})^2\{n^{(d_z-2m)/2m}/\lambda_n^{d_z/2m}\}$.

Second, consider the bias of $\partial^{\boldsymbol{\ell}}\hat{g}(\mathbf{z})$, and note that

$$\begin{aligned}
\partial^{\boldsymbol{\ell}}\hat{g}(\mathbf{z}) - \partial^{\boldsymbol{\ell}}g(\mathbf{z}) = {}& \{\partial^{\boldsymbol{\ell}}\mathbf{p}_n(\mathbf{z})^\top(\mathbf{P}^\top\mathbf{P}+\lambda_n\mathbf{D})^{-1}\mathbf{P}^\top\mathbf{G} - \partial^{\boldsymbol{\ell}}g(\mathbf{z})\} - \partial^{\boldsymbol{\ell}}\mathbf{p}_n(\mathbf{z})^\top(\mathbf{P}^\top\mathbf{P}+\lambda_n\mathbf{D})^{-1}\mathbf{P}^\top\mathbf{X}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})\\
&+ \partial^{\boldsymbol{\ell}}\mathbf{p}_n(\mathbf{z})^\top(\mathbf{P}^\top\mathbf{P}+\lambda_n\mathbf{D})^{-1}\mathbf{P}^\top\boldsymbol{\varepsilon},
\end{aligned}$$

where (i) $\partial^{\boldsymbol{\ell}}\mathbf{p}_n(\mathbf{z})^\top(\mathbf{P}^\top\mathbf{P}+\lambda_n\mathbf{D})^{-1}\mathbf{P}^\top\mathbf{G} - \partial^{\boldsymbol{\ell}}g(\mathbf{z})$ is $\mathrm{o_p}[k_n^{\boldsymbol{\ell}}\{(\lambda_n k^m/n) + k_n^{-r_g}\}]$ if $\lambda_n k_n^{2m}/n < 1$ and is $\mathrm{o_p}\{k_n^{\boldsymbol{\ell}}(\sqrt{\lambda_n/n} + k_n^{-r_g})\}$ if $\lambda_n k_n^{2m}/n \geq 1$, and (ii) $\partial^{\boldsymbol{\ell}}\mathbf{p}_n(\mathbf{z})^\top(\mathbf{P}^\top\mathbf{P}+\lambda_n\mathbf{D})^{-1}\mathbf{P}^\top\mathbf{X}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})$ is $\mathrm{O_p}[k_n^{\boldsymbol{\ell}}\{(\lambda_n k_n^m/n) + n^{-1/2}\}]$ if $\lambda_n k_n^{2m}/n < 1$ and is $\mathrm{O_p}\{k_n^{\boldsymbol{\ell}}(\sqrt{\lambda_n/n} + n^{-1/2})\}$ if $\lambda_n k_n^{2m}/n \geq 1$. Specifically, for (i),

$$\begin{aligned}
\mathrm{E}\{\partial^{\boldsymbol{\ell}}\hat{g}(\mathbf{z}) - \partial^{\boldsymbol{\ell}}\hat{s}(\mathbf{z})|\mathbf{X},\mathbf{Z}\} = {}& -\mathbb{1}_n\{\lambda_n\mathbf{p}_n(\mathbf{z})^\top\boldsymbol{\Delta}_{(\boldsymbol{\ell})}(\mathbf{P}^\top\mathbf{P}+\lambda_n\mathbf{D})^{-1}\mathbf{D}(\mathbf{P}^\top\mathbf{P})^{-1}\mathbf{P}^\top\bar{\mathbf{S}}_g\\
&- \lambda_n\mathbf{p}_n(\mathbf{z})^\top\boldsymbol{\Delta}_{(\boldsymbol{\ell})}(\mathbf{P}^\top\mathbf{P}+\lambda_n\mathbf{D})^{-1}\mathbf{D}(\mathbf{P}^\top\mathbf{P})^{-1}\mathbf{P}^\top(\mathbf{G}-\bar{\mathbf{S}}_g)\}.
\end{aligned}$$

For the first term,

$$\begin{aligned}
\mathbb{1}_n\lambda_n\mathbf{p}_n(\mathbf{z})^\top(\mathbf{P}^\top\mathbf{P}+\lambda_n\mathbf{D})^{-1}\mathbf{D}(\mathbf{P}^\top\mathbf{P})^{-1}\mathbf{P}^\top\bar{\mathbf{S}}_g &\lesssim \mathbb{1}_n\lambda_n\mathbf{p}_n(\mathbf{z})^\top(\mathbf{I}+\lambda_n\mathbf{M}/n)^{-1}(\mathbf{M}/n)\mathbf{P}^\top\bar{\mathbf{S}}_g/n\\
&\lesssim \mathbb{1}_n\frac{\lambda_n}{n}\sum_{i=1}^n\sum_{k=1}^{k_n^{d_z}}\frac{\frac{\hat{\mu}_k}{n}}{1+\frac{\lambda_n}{n}\hat{\mu}_k}p_k(\mathbf{z})p_k(\mathbf{z}_i).
\end{aligned}$$

If $\lambda_n k_n^{2m}/n < 1$, since the number of observations for which p_k is nonzero is $\mathrm{O_p}(n/k_n^{d_z})$,

$$\begin{aligned}\mathbb{1}_n \frac{\lambda_n}{n}\sum_{i=1}^{n}\sum_{k=1}^{k_n^{d_z}}\frac{\frac{\hat\mu_k}{n}}{1+\frac{\lambda_n}{n}\hat\mu_k}p_k(\mathbf{z})p_k(\mathbf{z}_i) &\lesssim \mathbb{1}_n\frac{\lambda_n}{n}\sum_{i=1}^{n}\sum_{k=1}^{k_n^{d_z}}\frac{k^{2m/d}}{n}p_k(\mathbf{z})p_k(\mathbf{z}_i)\\ &\lesssim_p \frac{\lambda_n}{n}k_n^m k_n^{d_z/2}\frac{1}{n}k_n^{d_z/2}\frac{n}{k_n^{d_z}}\\ &\asymp \lambda_n k_n^m/n,\end{aligned}$$

If $\lambda_n k_n^{2m}/n \geq 1$, $\sqrt{\lambda_n k^{2m/d}/n}/(1+\lambda_n k^{2m/d}/n) \leq 1/2$. Then letting S_z be the set of all k such that $p_k(\mathbf{z}) \geq 0$,

$$\begin{aligned}\mathbb{1}_n \frac{\lambda_n}{n}\sum_{i=1}^{n}\sum_{k=1}^{k_n^{d_z}}\frac{\frac{\hat\mu_k}{n}}{1+\frac{\lambda_n}{n}\hat\mu_k}p_k(\mathbf{z})p_k(\mathbf{z}_i) &\asymp \mathbb{1}_n\sqrt{\frac{\lambda_n}{n}}\frac{1}{n}\sum_{i=1}^{n}\sum_{k\in S_z}\frac{\sqrt{\frac{\lambda_n k^{2m/d}}{n}}}{1+\frac{\lambda_n}{n}k^{2m/d}}\mathbf{p}_k(\mathbf{z})\mathbf{p}_k(\mathbf{z}_i)\\ &\lesssim_p \sqrt{\frac{\lambda_n}{n}}k_n^{d_z/2}\frac{1}{n}k_n^{d_z/2}\frac{n}{k_n^{d_z}}\\ &\asymp \sqrt{\lambda_n/n}.\end{aligned}$$

So using the structure of $\mathbf{\Delta}_{(\ell)}$ as in the proof of Theorem 1, we obtain the rates $k_n^\ell\lambda_n k^{2m}/n$, $k_n^\ell\sqrt{\lambda_n/n}$, and $k_n^\ell\sqrt{\lambda_n k^{2m/d}/n}\sqrt{\lambda_n/n}$ for $\mathrm{E}[\mathbb{1}_n\{\partial^\ell\hat g(\mathbf{z}) - \partial^\ell\hat s(\mathbf{z})\}|\mathbf{X},\mathbf{Z}]$. Then using Lemmas 3 and 4, if $\lambda_n k_n^{2m}/n \leq 1$, $\partial^\ell\mathbf{p}_n(\mathbf{z})'(\mathbf{P}^\top\mathbf{P}+\lambda_n\mathbf{D})^{-1}\mathbf{P}^\top\mathbf{G}-\partial^\ell g(\mathbf{z}) \lesssim_p k_n^\ell\{(\lambda_n k_n^m/n)+k_n^{-r_g}\}$, and if $\lambda_n k_n^{2m}/n \geq 1$, $\partial^\ell\mathbf{p}_n(\mathbf{z})'(\mathbf{P}^\top\mathbf{P}+\lambda_n\mathbf{D})^{-1}\mathbf{P}^\top\mathbf{G}-\partial^\ell g(\mathbf{z}) \lesssim_p k_n^\ell(\sqrt{\lambda_n/n}+k_n^{-r_g})$.

For (ii), let b_n equal $\lambda_n k_n^m/n$ if $\lambda_n k^{2m}/n < 1$ and λ_n/n if $\lambda_n k^{2m}/n \geq 1$. As shown in the proof of Theorem 2, if $nk_n^{-2r_g-2r_h} \to 0$ and $b_n \to 0$, then $\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}/n = \mathbf{\Sigma}_v(\mathbf{z}) + \mathrm{o_p}(1)$. So under these conditions, $(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}/n)^{-1} = \mathrm{O_p}(1)$ since $\mathbf{\Sigma}_v(\mathbf{z}) > 0$. Also, using the same steps as in the proof of Theorem 2, we see that under the same conditions, $\tilde{\mathbf{X}}^\top(\tilde{\mathbf{y}}-\tilde{\mathbf{X}}\boldsymbol{\theta})/n = \boldsymbol{v}^\top\boldsymbol{\varepsilon}/n + \mathrm{O_p}(b_n)$. We also showed that $\mathbf{\Omega}^{-1/2}\boldsymbol{v}^\top\boldsymbol{\varepsilon}/\sqrt{n} \to_d \mathcal{N}(0,1)$. Since $\mathrm{E}(\boldsymbol{v}_i\boldsymbol{v}_i^\top\varepsilon_i^2|\mathbf{x}_i,\mathbf{z}_i) \leq \{\mathrm{E}(\|\boldsymbol{v}_i\|^4|\mathbf{x}_i\mathbf{z}_i)\}^{1/2}\{\mathrm{E}(\varepsilon_i^4|\mathbf{z}_i)\}^{1/2}$ is bounded above, $\mathbf{\Omega}$ is bounded above, so $\boldsymbol{v}^\top\boldsymbol{\varepsilon}/n = \mathrm{O_p}(1/\sqrt{n})$. Then $\hat{\boldsymbol{\theta}}-\boldsymbol{\theta} = (\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}/n)^{-1}\tilde{\mathbf{X}}^\top(\tilde{\mathbf{y}}-\tilde{\mathbf{X}}\boldsymbol{\theta})/n = \mathrm{O_p}(n^{-1/2}+b_n)$, so

$$\begin{aligned}\mathbf{p}_n(\mathbf{z})^\top(\mathbf{P}^\top\mathbf{P}+\lambda_n\mathbf{D})^{-1}\mathbf{P}^\top\mathbf{X}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}) &= \mathbf{p}_n(\mathbf{z})^\top(\mathbf{Q}+\lambda_n\mathbf{D}/n)^{-1}\mathbf{P}^\top\mathbf{X}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})/n + \mathrm{o_p}(1)\\ &\lesssim_p \frac{1}{n}\sum_{k=1}^{k_n^{d_z}}\frac{p_k(\mathbf{z})}{1+\frac{\lambda_n}{n}\hat\mu_k}\sum_{i=1}^{n}p_k(\mathbf{z}_i)\sum_{j=1}^{d}\mathbf{X}_{ji}(\hat\theta_j-\theta_j)\\ &\lesssim_p \frac{1}{n}k_n^{d_z/2}\times k_n^{d_z/2}\times\frac{n}{k_n^{d_z}}(n^{-1/2}+b_n)\\ &= n^{-1/2}+b_n.\end{aligned}$$

Then using the structure of $\mathbf{\Delta}_{(\ell)}$, we obtain $\partial^\ell\mathbf{p}_n(\mathbf{z})^\top(\mathbf{P}^\top\mathbf{P}+\lambda_n\mathbf{D})^{-}\mathbf{P}^\top\mathbf{X}^\top(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}) \lesssim_p k_n^\ell(n^{-1/2}+b_n)$.

Finally, consider $\{\partial^\ell\hat g(\mathbf{z}) - \partial^\ell g(\mathbf{z})\}/\sqrt{\mathbf{W}_{n,\ell}(\mathbf{z})}$. If $\lambda_n k_n^{2m}/n < 1$,

$$\frac{\partial^\ell\mathbf{p}_n(\mathbf{z})'(\mathbf{P}^\top\mathbf{P}+\lambda_n\mathbf{D})^{-1}\mathbf{P}^\top\mathbf{G}-\partial^\ell g(\mathbf{z})}{\sqrt{\mathbf{W}_{n,\ell}(\mathbf{z})}} = \mathrm{O_p}\left(\frac{\lambda_n k_n^m/n + k_n^{-r_g}}{\sqrt{k_n^{d_z}/n}}\right) = \mathrm{o_p}(1);$$

if $\lambda_n k_n^{2m}/n \geq 1$ is bounded above,

$$\frac{\partial^\ell\mathbf{p}_n(\mathbf{z})'(\mathbf{P}^\top\mathbf{P}+\lambda_n\mathbf{D})^{-1}\mathbf{P}^\top\mathbf{G}-\partial^\ell g(\mathbf{z})}{\sqrt{\mathbf{W}_{n,\ell}(\mathbf{z})}} = \mathrm{O_p}\left(\frac{\sqrt{\lambda_n/n} + k_n^{-r_g}}{\sqrt{k_n^{d_z}/n}}\right) = \mathrm{o_p}(1);$$

and if $\lambda_n k_n^{2m}/n$ is unbounded,

$$\frac{\partial^\ell\mathbf{p}_n(\mathbf{z})'(\mathbf{P}^\top\mathbf{P}+\lambda_n\mathbf{D})^{-1}\mathbf{P}^\top\mathbf{G}-\partial^\ell g(\mathbf{z})}{\sqrt{\mathbf{W}_{n,\ell}(\mathbf{z})}} = \mathrm{O_p}\left\{\frac{\lambda_n k_n^{2m}}{n}\left(\frac{\sqrt{\lambda_n/n} + k_n^{-r_g}}{\sqrt{k_n^{d_z}/n}}\right)\right\} = \mathrm{o_p}(1).$$

Similarly, if $\lambda_n k_n^{2m}/n < 1$,

$$\frac{\partial^\ell \mathbf{p}_n(\mathbf{z})'(\mathbf{P}^\top\mathbf{P}+\lambda_n\mathbf{D})^{-1}\mathbf{P}^\top\mathbf{X}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})}{\sqrt{\mathbf{W}_{n,\ell}(\mathbf{z})}} = \mathrm{O_p}\left(\frac{\sqrt{1/n}+\lambda k_n^m/n}{\sqrt{k_n^{d_z}/n}}\right) = \mathrm{o_p}(1);$$

if $\lambda_n k_n^{2m}/n \geq 1$ is bounded above,

$$\frac{\partial^\ell \mathbf{p}_n(\mathbf{z})'(\mathbf{P}^\top\mathbf{P}+\lambda_n\mathbf{D})^{-1}\mathbf{P}^\top\mathbf{X}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})}{\sqrt{\mathbf{W}_{n,\ell}(\mathbf{z})}} = \mathrm{O_p}\left(\frac{\sqrt{1/n}+\sqrt{\lambda_n/n}}{\sqrt{k_n^{d_z}/n}}\right) = \mathrm{o_p}(1);$$

and if $\lambda_n k_n^{2m}/n$ is unbounded,

$$\frac{\partial^\ell \mathbf{p}_n(\mathbf{z})'(\mathbf{P}^\top\mathbf{P}+\lambda_n\mathbf{D})^{-1}\mathbf{P}^\top\mathbf{X}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})}{\sqrt{\mathbf{W}_{n,\ell}(\mathbf{z})}} = \mathrm{O_p}\left(\frac{\lambda_n k_n^{2m}}{n}\frac{\sqrt{1/n}+\sqrt{\lambda_n/n}}{\sqrt{k_n^{d_z}/n}}\right) = \mathrm{o_p}(1).$$

Note that $\mathbf{W}_{n,\ell}(\mathbf{z}) \gtrsim \mathbf{p}_{n,\ell}(\mathbf{z})^\top\boldsymbol{\Delta}_{(\ell)}(\mathbf{P}^\top\mathbf{P}+\lambda_n\mathbf{D})^{-1}\mathbf{P}^\top\mathbf{P}(\mathbf{P}^\top\mathbf{P}+\lambda_n\mathbf{D})^{-1}\boldsymbol{\Delta}_{(\ell)}{}'\mathbf{p}_{n,\ell}(\mathbf{z})$. Define $d_i = \partial^\ell\mathbf{p}_n(\mathbf{z})^\top(\mathbf{P}^\top\mathbf{P}+\lambda_n\mathbf{D})^{-1}\mathbf{p}_n(\mathbf{z}_i)$, then since $\mathbf{p}_n(\mathbf{z})'\mathbf{p}_n(\mathbf{z}) \leq k_n^{d_z}$,

$$\begin{aligned}
\mathbb{1}_n d_i^2 &= \mathbb{1}_n \mathbf{p}_{k_n^{d_z}-\ell}(\mathbf{z})^\top\boldsymbol{\Delta}_{(\ell)}(\mathbf{P}^\top\mathbf{P}+\lambda_n\mathbf{D})^{-1}\mathbf{p}_n(\mathbf{z}_i)\mathbf{p}_n(\mathbf{z}_i)^\top(\mathbf{P}^\top\mathbf{P}+\lambda_n\mathbf{D})^{-1}\boldsymbol{\Delta}_{(\ell)}{}'\mathbf{p}_{k_n^{d_z}-\ell}(\mathbf{z}) \\
&\leq \mathbb{1}_n\lambda_{\max}\{\mathbf{p}_n(\mathbf{z}_i)\mathbf{p}_n(\mathbf{z}_i)^\top\}\mathbf{p}_{k_n^{d_z}-\ell}(\mathbf{z})^\top\boldsymbol{\Delta}_{(\ell)}(\mathbf{P}^\top\mathbf{P}+\lambda_n\mathbf{D})^{-1}(\mathbf{P}^\top\mathbf{P})^{-1}\mathbf{P}^\top\mathbf{P}(\mathbf{P}^\top\mathbf{P}+\lambda_n\mathbf{D})^{-1}\boldsymbol{\Delta}_{(\ell)}{}'\mathbf{p}_{k_n^{d_z}-\ell}(\mathbf{z}) \\
&\lesssim k_n^{d_z}\mathbf{W}_{n,\ell}/n \\
&= o(\mathbf{W}_{n,\ell}).
\end{aligned}$$

So since $\sum_{i=1}^n d_i^2 \asymp \mathbf{W}_{n,\ell}(\mathbf{z})$, $\max_{1\leq i\leq n} d_i^2 = o\left(\sum_{i=1}^n d_i^2\right) = o(\mathbf{W}_{n,\ell}(\mathbf{z}))$, and by the Lindeberg–Feller Central Limit Theorem, $\mathbb{1}_n\{\partial^\ell\hat{g}(\mathbf{z}) - \partial^\ell g(\mathbf{z})\}/\sqrt{\mathbf{W}_{n,\ell}(\mathbf{z})} \to_d \mathcal{N}(0,1)$. Then since $(\mathbb{1}_n - 1)\{\partial^\ell\hat{g}(\mathbf{z}) - \partial^\ell g(\mathbf{z})\}/\sqrt{\mathbf{W}_{n,\ell}(\mathbf{z})} \to_p 0$, we have $\{\partial^\ell\hat{g}(\mathbf{z}) - \partial^\ell g(\mathbf{z})\}/\sqrt{\mathbf{W}_{n,\ell}(\mathbf{z})} \to_d \mathcal{N}(0,1)$.

Last, we show that $\hat{\mathbf{W}}_{n,\ell}(\mathbf{z}) = \mathbf{W}_{n,\ell}(\mathbf{z}) + \mathrm{o_p}(1)$. Note that $\mathbf{X}_i^\top(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}) \lesssim_p \sum_{j=1}^d(\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j) \lesssim_p n^{-1/2} + b_n$. Then

$$\begin{aligned}
\mathbb{1}_n\{\hat{\mathbf{W}}_{n,\ell}(\mathbf{z}) - \mathbf{W}_{n,\ell}(\mathbf{z})\} &\lesssim_p \mathbb{1}_n\mathbf{p}_n(\mathbf{z})^\top\Delta_\ell(\mathbf{P}^\top\mathbf{P}+\lambda_n\mathbf{D})^{-1}\mathbf{P}^\top\mathbf{P}(\mathbf{P}^\top\mathbf{P}+\lambda_n\mathbf{D})^{-1}\Delta_\ell{}^\top\mathbf{p}_n(\mathbf{z}) \\
&= \mathbb{1}_n\mathbf{p}_n(\mathbf{z})^\top\Delta_\ell\Delta_\ell\mathbf{p}_n(\mathbf{z})/n + \mathrm{o_p}(1) \\
&\lesssim_p k_n^{2|\ell|}k_n^{d_z}/n \\
&\lesssim 1
\end{aligned}$$

So $\hat{\mathbf{W}}_{n,\ell}(\mathbf{z}) - \mathbf{W}_{n,\ell}(\mathbf{z}) = \mathrm{o_p}(1)$, as desired.

Appendix B. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.jmva.2016.10.001.

References

[1] R. Adams, J. Fournier, Sobolev Spaces. Vol. 140, Academic Press, 2003.
[2] M. Aerts, G. Claeskens, M.P. Wand, Some theory for penalized spline generalized additive models, J. Statist. Plann. Inference 103 (2002) 455–470. http://dx.doi.org/10.1016/S0378-3758(01)00237-3.
[3] S. Agmon, Lectures on Elliptic Boundary Value Problems, American Mathematical Society (RI), 2010.
[4] G. Aneiros, P. Vieu, Variable selection in infinite-dimensional problems, Statist. Probab. Lett. 94 (2014) 12–20. http://dx.doi.org/10.1016/j.spl.2014.06.025.
[5] G. Aneiros-Pérez, P. Vieu, Semi-functional partial linear regression, Statist. Probab. Lett. 76 (2006) 1102–1110. http://dx.doi.org/10.1016/j.spl.2005.12.007.
[6] P. Avramidis, Two-step cross-validation selection method for partially linear models, Statist. Sinica 15 (2005) 1033–1048.
[7] A.K. Aziz, The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations, Academic Press, 2014.
[8] A. Belloni, V. Chernozhukov, D. Chetverikov, K. Kato, Some new asymptotic theory for least squares series: Pointwise and uniform results, J. Econometrics 186 (2015) 345–366. http://dx.doi.org/10.1016/j.jeconom.2015.02.014.
[9] M.D. Cattaneo, M.H. Farrell, Optimal convergence rates, Bahadur representation, and asymptotic normality of partitioning estimators, J. Econometrics 174 (2013) 127–143. http://dx.doi.org/10.1016/j.jeconom.2013.02.002.
[10] M.D. Cattaneo, M. Jansson, W.K. Newey, Alternative asymptotics and the partially linear model with many regressors, 2015. arXiv preprint arXiv:1505.08120.
[11] M.D. Cattaneo, M. Jansson, W.K. Newey, Treatment effects with many covariates and heteroskedasticity, 2015. arXiv preprint arXiv:1507.02493.
[12] X.W. Chang, L. Qu, Wavelet estimation of partially linear models, Comput. Statist. Data Anal. 47 (2004) 31–48. http://dx.doi.org/10.1016/j.csda.2003.10.018.

[13] X. Chen, Large sample sieve estimation of semi-nonparametric models, Handbook Econom. 6 (2007) 5549–5632. http://dx.doi.org/10.1016/S1573-4412(07)06076-X.
[14] X. Chen, T.M. Christensen, Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions, J. Econometrics 188 (2015) 447–465.
[15] G. Claeskens, T. Krivobokova, J. Opsomer, Asymptotic properties of penalized spline estimators, Biometrika 96 (2009) 529–544. http://dx.doi.org/10.1093/biomet/asp035.
[16] D. Cox, Multivariate smoothing spline functions, SIAM J. Numer. Anal. (1984) 789–813. http://dx.doi.org/10.1137/0721053.
[17] P. Craven, Smoothing noisy data with spline functions, Numer. Math. 31 (1978) 377–403. http://dx.doi.org/10.1007/bf01404567.
[18] C. De Boor, A Practical Guide to Splines, Springer, Berlin, 1978, http://dx.doi.org/10.1007/978-1-4612-6333-3.
[19] S.G. Donald, W.K. Newey, Series estimation of semilinear models, J. Multivariate Anal. 50 (1994) 30–40. http://dx.doi.org/10.1006/jmva.1994.1032.
[20] P.H.C. Eilers, B.D. Marx, Flexible smoothing with b-splines and penalties, Statist. Sci. 11 (1996) 89–121. http://dx.doi.org/10.1214/ss/1038425655.
[21] R. Haberman, Applied partial differential equations with Fourier series and boundary value problems, Recherche 67 (2003) 02.
[22] P. Hall, J.D. Opsomer, Theory for penalized spline regression, Biometrika 92 (2005) 105–118. http://dx.doi.org/10.1093/biomet/92.1.105.
[23] W.K. Härdle, M. Müller, S. Sperlich, A. Werwatz, Nonparametric and Semiparametric Models, Springer, Berlin, 2004, http://dx.doi.org/10.1007/978-3-642-17146-8.
[24] J. Huang, Asymptotics for polynomial spline regression under weak conditions, Statist. Probab. Lett. 65 (2003) 207–216. http://dx.doi.org/10.1016/j.spl.2003.09.003.
[25] J. Huang, Local asymptotics for polynomial spline regression, Ann. Statist. 31 (2003) 1600–1635. http://dx.doi.org/10.1214/aos/1065705120.
[26] G.W. Imbens, J.M. Wooldridge, Recent developments in the econometrics of program evaluation, Journal of Economic Literature 47 (2009) 5–86. http://dx.doi.org/10.1257/jel.47.1.5.
[27] G. Kauermann, T. Krivobokova, L. Fahmeir, Some asymptotic results on generalized penalized spline smoothing, J. R. Stat. Soc. Ser. B 71 (2009) 487–503. http://dx.doi.org/10.1111/j.1467-9868.2008.00691.x.
[28] T. Krivobokova, T. Kneib, G. Claeskens, Simultaneous confidence bands for penalized spline estimators, J. Amer. Statist. Assoc. 105 (2010) 852–863. http://dx.doi.org/10.1198/jasa.2010.tm09165.
[29] Q. Li, Efficient estimation of additive partially linear models, Internat. Econom. Rev. 41 (2000) 1073–1092. http://dx.doi.org/10.1111/1468-2354.00096.
[30] Y. Li, D. Ruppert, On the asymptotics of penalized splines, Biometrika 95 (2008) 415–436. http://dx.doi.org/10.1093/biomet/asn010.
[31] G. Li, L. Zhu, L. Xue, S. Feng, Empirical likelihood inference in partially linear single-index models for longitudinal data, J. Multivariate Anal. 101 (2010) 718–732. http://dx.doi.org/10.1016/j.jmva.2009.08.006.
[32] S. Manzan, D. Zerom, Kernel estimation of a partially linear additive model, Statist. Probab. Lett. 72 (2005) 313–322. http://dx.doi.org/10.1016/j.spl.2005.02.005.
[33] M. Naimark, W. Everett, Linear Differential Operators, Harrap, 1968.
[34] W.K. Newey, Convergence rates and asymptotic normality for series estimators, J. Econometrics 79 (1997) 147–168. http://dx.doi.org/10.1016/S0304-4076(97)00011-0.
[35] F. O'Sullivan, A statistical perspective on ill-posed inverse problems (with discussion), Statist. Sci. 1 (1986) 502–527. http://dx.doi.org/10.1214/ss/1177013525.
[36] J. Póo, Constrained smoothing splines, Econom. Theory 15 (1999) 114–138. http://dx.doi.org/10.1017/S0266466699151065.
[37] P.M. Robinson, Root-*n*-consistent semiparametric regression, Econometrica 56 (1988) 931–954. http://dx.doi.org/10.2307/1912705.
[38] D. Ruppert, M. Wand, R. Carroll, Semiparametric Regression, Cambridge University Press, 2003, http://dx.doi.org/10.1017/CBO9780511755453.
[39] L. Schumaker, Spline Functions: Basic Theory, Wiley, New York, 1981.
[40] J. Shi, T.S. Lau, Empirical likelihood for partially linear models, J. Multivariate Anal. 72 (2000) 132–148. http://dx.doi.org/10.1006/jmva.1999.1866.
[41] B.W. Silverman, Some aspects of the spline smoothing approach to non-parametric regression curve fitting, J. R. Stat. Soc. Ser. B 47 (1985) 1–52. http://dx.doi.org/10.2307/2345542.
[42] P. Speckman, Spline smoothing and optimal rates of convergence in nonparametric regression models, Ann. Statist. 13 (1985) 970–983. http://dx.doi.org/10.1214/aos/1176349650.
[43] J.H. Stock, Nonparametric policy analysis, J. Amer. Statist. Assoc. 84 (1989) 567–575. http://dx.doi.org/10.1080/01621459.1989.10478805.
[44] C.J. Stone, The use of polynomial splines and their tensor products in multivariate function estimation, Ann. Statist. 22 (1994) 118–171. http://dx.doi.org/10.1214/aos/1176325361.
[45] F. Utreras, Cross-validation techniques for smoothing spline functions in one or two dimensions, in: Smoothing Techniques for Curve Estimation, 1979, pp. 196–232. http://dx.doi.org/10.1007/BFb0098498.
[46] G. Wahba, Spline Models for Observational Data, Society for Industrial Mathematics, 1990, http://dx.doi.org/10.1137/1.9781611970128.
[47] M.P. Wand, On the optimal amount of smoothing in penalized spline regression, Biometrika 86 (1999) 936–940. http://dx.doi.org/10.1093/biomet/86.4.936.
[48] M.P. Wand, J.T. Ormerod, On semiparametric regression with O'Sullivan penalized splines, Aust. N. Z. J. Stat. 50 (2008) 179–198. http://dx.doi.org/10.1111/j.1467-842X.2008.00507.x.
[49] W.L. Xu, A note on the optimality of generalized cross-validation bandwidth selection in partially linear models with kernel smoothing estimator, Acta Math. Appl. Sin. 22 (2006) 345–352. http://dx.doi.org/10.1007/s10255-006-0310-y.
[50] Y. Yu, D. Ruppert, Penalized spline estimation for partially linear single-index models, J. Amer. Statist. Assoc. 97 (2002) 1042–1054. http://dx.doi.org/10.1198/016214502388618861.
[51] S. Zhou, X. Shen, D.A. Wolfe, Local asymptotics for regression splines and confidence regions, Ann. Statist. 26 (1998) 1760–1782. http://dx.doi.org/10.1214/aos/1024691356.
[52] S. Zhou, D. Wolfe, On derivative estimation in spline regression, Statist. Sinica 10 (2000) 93–108.