On Nonparametric and Semiparametric Partitioning-Based Methods in Applied Microeconomics

by

Yingjie Feng

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Economics)
in The University of Michigan
2019

Doctoral Committee:

Professor Matias D. Cattaneo, Chair
Assistant Professor Andreas Hagemann
Professor Lutz Kilian
Professor Rocío Titiunik

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

Yingjie Feng
yjfeng@umich.edu
ORCID iD: 0000-0002-9413-3239

Acknowledgments

First and foremost, I would like to thank my beloved wife, Wuyue You. She has made many sacrifices in the past five years, giving me full support for this research. Without her love, understanding, and encouragement, this dissertation would not exist at all.

I am deeply grateful to my advisor, mentor, and friend, Matias Cattaneo. It has been my good fortune to work with him during my Ph.D. program. Without his encouragement, I would not have discovered my interest in econometrics and entered this field. His continued guidance and seemingly endless patience helped me to overcome the difficulties in preparing this dissertation. His invaluable support gave me the courage to confront many challenges in the past few years.

I am indebted to Andreas Hagemann, Lutz Kilian, and Rocío Titiunik, who sat through multiple presentations, provided insightful feedback and valuable career advice, and gave strong support in many aspects. I also thank my collaborators, Max Farrell and Richard Crump. I have truly benefited from their wisdom and experience.

Finally, I wish to thank my friends, Xing Guo, Ting Lan, Xinwei Ma, Kenichi Nagasawa, Huayu Xu, and Guang Zeng, who helped me in many ways, and made these years truly enjoyable.

Table of Contents

List of Tables

Table

List of Figures

Figure

List of Appendices

Appendix

Abstract

This dissertation concerns estimation and inference using partitioning-based least squares estimators in nonparametric and semiparametric models.

Chapter II studies the large sample properties of partitioning-based estimators in a standard nonparametric regression model. First, a general characterization of their leading asymptotic bias is obtained, based on which several bias-corrected estimators are proposed. Second, integrated mean squared error (IMSE) approximations for the point estimator are established for principled tuning parameter selection. Third, pointwise and uniform inference methods are developed with and without bias correction techniques. In particular, the uniform inference results rely on novel uniform distributional approximations for the undersmoothed and robust bias-corrected t-statistic processes. In the univariate case, they require seemingly minimal rate restrictions and improve on the approximation rates known in the literature.

Chapter III examines *binscatter*, a particular application of partitioning-based methods to semiparametric partial linear models. An array of theoretical and practical results is offered, including principled number of bins selection, confidence intervals and bands, hypothesis tests for parametric and shape restrictions of the regression function, and several other new methods applicable to canonical binscatter and higher-order polynomial, covariate-adjusted, and smoothness-restricted extensions.

Chapter IV concerns the methodology for implementing these results. I first discuss several commonly used basis expansions. Their leading approximation errors are presented, which can be used for tuning parameter selection and bias-corrected inference. Subsequently, I give a more detailed IMSE approximation for the special case of a tensor-product partition. Using these results, I propose two data-driven procedures (rule-of-thumb and direct plug-in) for tuning parameter selection. Finally, an empirical example and simulation evidence are provided.

Chapter I
Introduction

Nonparametric and semiparametric methods are important tools for researchers in economics and many other disciplines. Compared to classical parametric approaches, they allow for more flexible functional form assumptions. This dissertation focuses on the popular class of partitioning-based least squares estimators, including splines, piecewise polynomials, and compactly supported wavelets. Such methods have been widely used in the study of treatment effects (Cattaneo and Farrell, 2011a), empirical finance (Cattaneo, Crump, Farrell, and Schaumburg, 2019a), and binscatter analysis (Chetty, Friedman, Hilger, Saez, Schanzenbach, and Yagan, 2011), among many other applications. From a theoretical perspective, they form the basis of many classical nonparametric series methods (Newey, 1997; Belloni, Chernozhukov, Chetverikov, and Kato, 2015; Chen, 2007) and relate to other modern machine learning techniques, such as regression trees (Breiman, Friedman, Stone, and Olshen, 1984; Hastie, Tibshirani, and Friedman, 2009) and trend filtering (Tibshirani, 2014).

The partitioning-based least squares methods are characterized by two features. First, the support of covariates is partitioned into non-overlapping cells and a set of local basis functions are constructed on top of the partition. Second, the final fit is determined by a least squares regression using these bases. The key distinguishing characteristic is that each basis function is nonzero on only a small, contiguous set of cells of the partition. This contrasts with, for example, global polynomial approximations.

This dissertation aims to offer valid, easy-to-implement estimation and inference procedures using partitioning-based estimators in nonparametric and semiparametric models. Chapter II, joint with Matias Cattaneo and Max Farrell, studies the large sample properties of such estimators in a standard nonparametric regression setup. First, we obtain a general characterization of their leading asymptotic bias, based on which several bias-corrected estimators are proposed. Second, we establish integrated mean squared error (IMSE) approximations for the point estimator, which can be used for principled tuning parameter selection. Third, we develop pointwise inference methods based on undersmoothing and robust bias correction. Fourth, by employing different coupling approaches, we develop uniform distributional approximations for the undersmoothed and robust bias-corrected t-statistic processes and construct valid confidence bands. In the univariate case, the uniform distributional approximations require seemingly minimal rate restrictions and improve on the approximation rates known in the literature.

Completely nonparametric models are subject to the curse of dimensionality when there are multiple covariates, and semiparametric models are often employed in such cases. Chapter III, joint with Matias Cattaneo, Richard Crump and Max Farrell, focuses on *binscatter*, a very popular tool in applied microeconomics that can be viewed as a semiparametric partial linear regression with a nonparametric component estimated using partitioning-based methods. We offer several theoretical and practical results that aid both in understanding current practices (i.e., their validity or lack thereof) and in offering theory-based guidance for future applications. The main results include principled number of bins selection, confidence intervals and bands, hypothesis tests for parametric and shape restrictions of the regression function, and several other new methods that are applicable to canonical binscatter and higher-order polynomial, covariate-adjusted, and smoothness-restricted extensions. In particular, we highlight important methodological problems related to covariate adjustment methods used in current practice.

Chapter IV concerns several methodological issues related to the implementation of these methods. First, I discuss three commonly used basis expansions. In particular, the leading approximation errors are presented, which can be used for tuning parameter selection and bias-corrected inference. Second, I specialize the general IMSE approximation given in Chapter II to a more detailed result for partitioning schemes formed via tensor products of intervals, which is usually of more practical interest. Third, using these results, I propose two data-driven procedures (rule-of-thumb and direct plug-in) for tuning parameter selection. Finally, an empirical example and simulation evidence are provided.

Chapter II
General Large Sample Properties

2.1 Introduction

This chapter studies the standard nonparametric regression setup, where $\{(y_i, \mathbf{x}_i'), i = 1, \dots n\}$ is a random sample from the model

$$y_i = \theta(\mathbf{x}_i) + \varepsilon_i, \qquad \mathbb{E}[\varepsilon_i|\mathbf{x}_i] = 0, \qquad \mathbb{E}[\varepsilon_i^2|\mathbf{x}_i] = \sigma^2(\mathbf{x}_i), \tag{2.1}$$

for a scalar response y_i and a d-vector of continuously distributed covariates $\mathbf{x}_i = (x_{1,i}, \dots, x_{d,i})'$ with compact support $\mathcal{X}$. The object of interest is the unknown regression function $\theta(\cdot)$ and its derivatives. We focus on general large sample properties of *partitioning-based*, or locally-supported, series (linear sieve) least squares regression estimators. These methods first partition the support of covariates, and then construct a set of local basis functions on top of it, each of which is nonzero on only a small number of cells of the partition. The final fit is determined by a least squares regression using these bases. Concrete examples are splines, piecewise polynomials and compactly supported wavelets. For this class of estimators, we develop novel bias approximations and pointwise and uniform estimation and inference results, with and without bias correction techniques.

A partitioning-based estimator is made precise by the partition of $\mathcal{X}$ and basis expansion used. Let $\Delta = \{\delta_l \subset \mathcal{X} : 1 \leq l \leq \bar{\kappa}\}$ be a collection of $\bar{\kappa}$ open and disjoint sets, the closure of whose union is $\mathcal{X}$ (or, more generally, covers $\mathcal{X}$). δ_l is restricted to be polyhedral, which allows for tensor products of (marginally-formed) intervals as well as other popular partitioning shapes. Based on this partition, the dictionary of K basis functions, each of order m (e.g., $m = 4$ for cubic splines) is denoted by $\mathbf{x}_i \mapsto \mathbf{p}(\mathbf{x}_i) := \mathbf{p}(\mathbf{x}_i; \Delta, m) = (p_1(\mathbf{x}_i; \Delta, m), \cdots, p_K(\mathbf{x}_i; \Delta, m))'$. For $\mathbf{x} \in \mathcal{X}$ and $\mathbf{q} = (q_1, \dots, q_d)' \in \mathbb{Z}_+^d$, the partial derivative $\partial^{\mathbf{q}}\theta(\mathbf{x})$ is estimated by least squares regression

$$\widehat{\partial^{\mathbf{q}}\theta}(\mathbf{x}) = \partial^{\mathbf{q}}\mathbf{p}(\mathbf{x})'\widehat{\boldsymbol{\beta}}, \qquad\qquad \widehat{\boldsymbol{\beta}} \in \operatorname*{arg\,min}_{\mathbf{b}\in\mathbb{R}^K} \sum_{i=1}^{n}(y_i - \mathbf{p}(\mathbf{x})'\mathbf{b})^2, \tag{2.2}$$

where $\partial^{\mathbf{q}}\theta(\mathbf{x}) = \partial^{q_1+\cdots+q_d}\theta(\mathbf{x})/\partial^{q_1}x_1\cdots\partial^{q_d}x_d$ (and for boundary points defined from the interior of $\mathcal{X}$ as usual), and $\theta(\mathbf{x}) := \partial^{\mathbf{0}}\theta(\mathbf{x})$.

The approximation power of this class of estimators comes from two user-specified parameters: the granularity of the partition Δ and the order $m \in \mathbb{Z}_+$ of the basis. The choice m is often fixed in practice, and hence we regard Δ as the tuning parameter for this class of nonparametric estimators. Under the assumptions given later, $\bar{\kappa} \to \infty$ as the sample size $n \to \infty$, and the volume of each δ_l shrinks proportionally to h^d, where $h = \max\{\text{diam}(\delta) : \delta \in \Delta\}$ serves as a universal measure of the granularity. Thus, as $\bar{\kappa} \to \infty$, h^d vanishes at the same rate, and with each basis being supported only on a finite number of cells, K diverges proportionally as well.

The first contribution, in Section 2.3, is a general characterization of the bias of partitioning-based estimators, which is later used for robust bias corrected inference and for tuning parameter selection. The generic bias approximation will be specialized to splines, wavelets, and piecewise polynomials in Chapter IV, leading to novel bias representations.

The second contribution, in Section 2.4, is a general integrated mean squared error (IMSE) expansion for partitioning-based estimators. These results lead to IMSE-optimal partitioning choices, and hence deliver IMSE-optimal point estimators of the regression function and its derivatives. We show that the IMSE-optimal choice of partition granularity obeys $h_{\texttt{IMSE}} \asymp n^{-1/(2m+d)}$, which translates to the familiar $K_{\texttt{IMSE}} \asymp n^{-d/(2m+d)}$, and give a precise characterization of the leading constant. For simple cases on tensor-product partitions, some results exist for splines (Agarwal and Studden, 1980; Zhou, Shen, and Wolfe, 1998; Zhou and Wolfe, 2000) and piecewise polynomials (Cattaneo and Farrell, 2013). In addition to generalizing these results substantially (e.g., allowing for more general support and partitioning schemes), our characterization for compactly supported wavelets (discussed in Chapter IV) appears to be new.

The IMSE-optimal partition scheme, and consistent implementations thereof, can not be used directly to form valid pointwise or uniform (in $\mathbf{x} \in \mathcal{X}$) inference procedures. From a nonparametric inference perspective, undersmoothing is a theoretically valid approach (i.e., employing a finer partition than the IMSE-optimal one), but it is difficult to implement in a principled way. Inspired by results proving that undersmoothing is never optimal relative to bias correction for kernel-based nonparametrics (Calonico, Cattaneo, and Farrell, 2018b), we develop three robust bias-corrected inference procedures using our new bias characterizations of partitioning-based estimators. These methods are more involved than their kernel-based counterparts, but are still based on least-squares regression using partitioning-based estimation. Specifically, we show that the conventional partitioning-based estimator $\widehat{\partial^{\mathbf{q}}\theta}(\mathbf{x})$ and the three bias-corrected estimators we propose have a common structure, which we exploit to obtain general pointwise and uniform distributional approximations under weak (sometimes minimal) conditions. These robust bias correction results for partitioning-based estimators, both pointwise and uniform in $\mathbf{x}$, appear to be new to the literature. They are practically useful because they allow for mean squared error minimizing tuning parameter choices (e.g., "rule-of-thumb", "plug-in", or "cross-validation" methods), thus offering a data-driven method combining optimal point estimation and valid inference, both employing the same partitioning scheme.

Section 2.5 establishes pointwise in $\mathbf{x} \in \mathcal{X}$ distributional approximations for both conventional and robust bias-corrected t-statistics based on partitioning-based estimators. These pointwise distributional results are made uniform in Section 2.6, where we establish a strong approximation for the whole t-statistic processes, indexed by the point $\mathbf{x} \in \mathcal{X}$, covering both conventional and robust bias-corrected inference. To illustrate, Section 2.6.3 constructs valid confidence bands for (derivatives of) the regression function using our uniform distributional approximations. When compared to the current literature, we obtain a strong approximation to the *entire* t-statistic process under either weaker or seemingly minimal conditions on the tuning parameter h (i.e., on K or $\bar{\kappa}$), depending on the case under consideration.

Finally, Section 2.7 concludes. Appendix A gives proofs of our main results.

2.1.1 Related Literature

This paper contributes primarily to two literatures, nonparametric regression and strong approximations. There is a vast literature on nonparametric regression, summarized in many textbook treatments (e.g., Fan and Gijbels, 1996; Györfi, Kohler, Krzyżak, and Walk, 2002; Wasserman, 2006; Horowitz, 2009; Ruppert, Wand, and Carroll, 2009, and references therein). Of particular relevance are treatments of series (linear sieve) methods in general, and while some results concerning partitioning-based estimators exist, they are mainly limited to splines, wavelets, or piecewise polynomials, considered separately (Newey, 1997; Huang, 1998; Zhou, Shen, and Wolfe, 1998; Huang, 2003; Chen, 2007; Cattaneo and Farrell, 2013; Belloni, Chernozhukov, Chetverikov, and Kato, 2015; Chen and Christensen, 2015; Belloni, Chernozhukov, Chetverikov, and Fernandez-Val, 2018). Piecewise polynomial fits on partitions have a long and ongoing tradition in statistics, dating at least to the regressogram of Tukey (Tukey, 1961a), continuing through Stone (1982) (named local polynomial regression therein) and Györfi, Kohler, Krzyżak, and Walk (2002); Cattaneo and Farrell (2013), and up to modern, data-driven partitioning techniques such as regression trees (Breiman, Friedman, Stone, and Olshen, 1984; Hastie, Tibshirani, and Friedman, 2009), trend filtering (Tibshirani, 2014), and related methods (Zhang and Singer, 2010). Partitioning-based methods have also featured as inputs or preprocessing in treatment effects (Cattaneo and Farrell, 2011a; Calonico, Cattaneo, and Titiunik, 2015), empirical finance (Cattaneo, Crump, Farrell, and Schaumburg, 2019b), "binscatter" analysis (Cattaneo, Crump, Farrell, and Feng, 2019b), and other settings. The bias corrections we develop for series estimation and uniform inference follow recent work in kernel-based nonparametric inference (Calonico, Cattaneo, and Titiunik, 2014; Calonico, Cattaneo, and Farrell, 2018b,a). Our coupling and strong approximation results relate to early work discussed in Eggermont and LaRiccia (2009, Chapter 22) and the more recent work in Chernozhukov, Lee, and Rosen (2013), Chernozhukov, Chetverikov, and Kato (2014a,b, 2015, 2016) and Zhai (2018), as well as with the results for series estimators in Belloni, Chernozhukov, Chetverikov, and Kato (2015) and Belloni, Chernozhukov, Chetverikov, and Fernandez-Val (2018). See also Zaitsev (2013) for a review on strong approximation methods, and background references. Finally, see Hall and Horowitz (2013), and references therein, for related work on valid confidence bands for (derivatives of) the regression function.

2.1.2 Notation

For a d-tuple $\mathbf{q} = (q_1, \ldots, q_d) \in \mathbb{Z}_+^d$, define $[\mathbf{q}] = \sum_{j=1}^d q_j$, $\mathbf{x}^{\mathbf{q}} = x_1^{q_1} x_2^{q_2} \cdots x_d^{q_d}$ and $\partial^{\mathbf{q}}\theta(\mathbf{x}) = \partial^{[\mathbf{q}]}\theta(\mathbf{x})/\partial x_1^{q_1} \ldots \partial x_d^{q_d}$. Unless explicitly stated otherwise, whenever $\mathbf{x}$ is a boundary point of some closed set, the partial derivative is understood as the limit with $\mathbf{x}$ ranging within it. Let $\mathbf{0} = (0, \cdots, 0)'$ be the length-d zero vector. We set $\theta(\mathbf{x}) := \partial^{\mathbf{0}}\theta(\mathbf{x})$ and $\widehat{\theta}_j(\mathbf{x}) := \widehat{\partial^{\mathbf{0}}\theta}_j(\mathbf{x})$ for $j = 0, 1, 2, 3$ and collect the covariates as $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]'$. The tensor product or Kronecker product operator is $\otimes$. The smallest integer greater than or equal to u is $\lceil u \rceil$. For two random variables X and Y, $X =_d Y$ denotes that they have the same probability law.

We use several norms. For a vector $\mathbf{v} = (v_1, \ldots, v_M) \in \mathbb{R}^M$, we write

$\|\mathbf{v}\| = (\sum_{i=1}^{M} v_i^2)^{1/2}$ and $\dim(\mathbf{v}) = M$. For a matrix $\mathbf{A} \in \mathbb{R}^{M\times N}$, $\|\mathbf{A}\| = \max_i \sigma_i(\mathbf{A})$ and $\|\mathbf{A}\|_\infty = \max_{1\leq i\leq M} \sum_{j=1}^{N} |a_{ij}|$ for operator norms induced by L_2 and L_∞ norms, where $\sigma_i(\mathbf{A})$ is the i-th singular value of $\mathbf{A}$, and $\lambda_{\min}(\mathbf{A})$ is the minimum eigenvalue of $\mathbf{A}$.

We use the usual empirical process notation:

$$\mathbb{E}_n[g(\mathbf{x}_i)] = \frac{1}{n}\sum_{i=1}^{n} g(\mathbf{x}_i), \quad \text{and} \quad \mathbb{G}_n[g(\mathbf{x}_i)] = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(g(\mathbf{x}_i) - \mathbb{E}[g(\mathbf{x}_i)]).$$

For sequences of numbers or random variables: $a_n \lesssim b_n$ denotes that $\limsup_n |a_n/b_n|$ is finite; $a_n = O_{\mathbb{P}}(b_n)$ denotes $\limsup_{\epsilon\to\infty} \limsup_n \mathbb{P}[|a_n/b_n| \geq \epsilon] = 0$; $a_n = o(b_n)$ denotes $a_n/b_n \to 0$; $a_n = o_{\mathbb{P}}(b_n)$ denotes $a_n/b_n \to_{\mathbb{P}} 0$, where $\to_{\mathbb{P}}$ is convergence in probability; $a_n \asymp b_n$ denotes $a_n \lesssim b_n$ and $b_n \lesssim a_n$. Limits are taken as $n \to \infty$ (and $h \to 0$, $K \to \infty$, when appropriate), unless stated otherwise.

Finally, throughout the paper, $r_n > 0$ denotes a non-vanishing sequence and $\nu > 0$ denotes a fixed constant used to characterize moment bounds.

2.2 Setup

We first make precise our setup and assumptions. Our first assumption restricts the data generating process.

Assumption II.1 (Data Generating Process)**.**

1. *$\{(y_i, \mathbf{x}_i') : 1 \leq i \leq n\}$ are i.i.d. satisfying (2.1), where $\mathbf{x}_i$ has compact connected support $\mathcal{X} \subset \mathbb{R}^d$ and an absolutely continuous distribution function. The density of $\mathbf{x}_i$, $f(\cdot)$, and the conditional variance of y_i given $\mathbf{x}_i$, $\sigma^2(\cdot)$, are bounded away from zero and continuous.*

2. *$\theta(\cdot)$ is S-times continuously differentiable, for $S > [\mathbf{q}]$, and all $\partial^{\varsigma}\theta(\cdot)$, $[\varsigma] = S$, are Hölder continuous with exponent $\varrho > 0$.*

The next two assumptions specify a set of high-level conditions on the partition and basis: we require that the partition is "quasi-uniform" and the basis is "locally" supported.

Assumption II.2 (Quasi-Uniform Partition)**.** *The ratio of the sizes of inscribed and circumscribed balls of each $\delta \in \Delta$ is bounded away from zero uniformly in $\delta \in \Delta$, and*

$$\frac{\max\{\operatorname{diam}(\delta) : \delta \in \Delta\}}{\min\{\operatorname{diam}(\delta) : \delta \in \Delta\}} \lesssim 1,$$

where $\operatorname{diam}(\delta)$ *denotes the diameter of* δ.

This condition implies that the size of each $\delta \in \Delta$ can be well characterized by the diameter of δ and that we can use $h = \max\{\operatorname{diam}(\delta) : \delta \in \Delta\}$ as a universal measure of mesh sizes of elements in Δ. In the univariate case, it reduces to a bounded mesh ratio. A special case of a quasi-uniform partition is one formed via a tensor product of univariate marginal partitions on each dimension of $\mathbf{x} \in \mathcal{X}$, with appropriately chosen knot positions. If Δ covers only strict subset of $\mathcal{X}$, then our results hold on that subset.

We focus on nonrandom partitions in this chapter. Data-dependent partitioning could be accommodated by sample splitting: estimating the partition configuration in one subsample and performing inference in the other. In this way, quite general partitions can be used with our results, including data-driven methods such as regression trees and other modern machine learning techniques. In fact, these modern methods would typically generate non-tensor-product partitioning schemes. In general, treating data-dependent partitioning would require non-trivial additional technical work and further assumptions. We defer general discussion of this to future study, but we note that Chapter III will provide formal results treating random partitions based empirical quantiles, and some other specific results are also available in the literature (Breiman, Friedman, Stone, and Olshen, 1984; Nobel, 1996; Calonico, Cattaneo, and Titiunik, 2015).

The second assumption on the partitioning-based estimators employs generalized notions of *stable local basis* (Davydov, 2001) and *active basis* (Huang, 2003). We say a function $p(\cdot)$ on $\mathcal{X}$ is *active* on $\delta \in \Delta$ if it is not identically zero on δ.

Assumption II.3 (Local Basis)**.**

1. *For each basis function* p_k*,* $k = 1, \dots, K$*, the union of elements of* Δ *on which* p_k *is active is a connected set, denoted by* $\mathcal{H}_k$*. For all* $k = 1, \dots, K$*, both the number of elements of* $\mathcal{H}_k$ *and the number of basis functions which are active on* $\mathcal{H}_k$ *are bounded by a constant.*

2. *For any* $\mathbf{a} = (a_1, \cdots, a_K)' \in \mathbb{R}^K$*,*

$$\mathbf{a}' \int_{\mathcal{H}_k} \mathbf{p}(\mathbf{x}; \Delta, m)\mathbf{p}(\mathbf{x}; \Delta, m)' \, d\mathbf{x}\, \mathbf{a} \gtrsim a_k^2 h^d, \qquad k = 1, \dots, K.$$

3. *For an integer* $\varsigma \in [[\mathbf{q}], m)$*, for all* $\boldsymbol{\varsigma}, [\boldsymbol{\varsigma}] \leq \varsigma$*,*

$$h^{-[\boldsymbol{\varsigma}]} \lesssim \inf_{\delta \in \Delta} \inf_{\mathbf{x} \in \operatorname{clo}(\delta)} \|\partial^{\boldsymbol{\varsigma}} \mathbf{p}(\mathbf{x}; \Delta, m)\| \leq \sup_{\delta \in \Delta} \sup_{\mathbf{x} \in \operatorname{clo}(\delta)} \|\partial^{\boldsymbol{\varsigma}} \mathbf{p}(\mathbf{x}; \Delta, m)\| \lesssim h^{-[\boldsymbol{\varsigma}]}$$

where $\mathrm{clo}(\delta)$ *is the closure of* δ*, and for* $[\boldsymbol{\varsigma}] = \varsigma + 1$*,*

$$\sup_{\delta\in\Delta} \sup_{\mathbf{x}\in\mathrm{clo}(\delta)} \|\partial^{\boldsymbol{\varsigma}}\mathbf{p}(\mathbf{x};\Delta,m)\| \lesssim h^{-\varsigma-1}.$$

Assumption II.3 imposes conditions ensuring the stability of the L_2 projection operator onto the approximating space. Condition II.3(1) requires that each basis function in $\mathbf{p}(\mathbf{x};\Delta,m)$ be supported by a region consisting of a finite number of cells in Δ. Therefore, as $\bar{\kappa}\to\infty$ (and $h\to 0$), each element of Δ shrinks and all the basis functions are "locally supported" relative to the whole support of the data. Another common assumption in least squares regression is that the regressors are not too co-linear: the minimum eigenvalue of $\mathbb{E}[\mathbf{p}(\mathbf{x}_i)\mathbf{p}(\mathbf{x}_i)']$ is usually assumed to be bounded away from zero. Since the local support condition in Assumption II.3(1) implies a banded structure for this matrix, it suffices to require that the basis functions are not too co-linear locally, as stated in Assumption II.3(2). These two assumptions are very similar to Conditions A.2 and Conditions A.3 in the Appendix of Huang (2003), and therefore they could also be used to establish theoretical results analogous to those discussed in that appendix (those results are not explicitly needed in our paper because our proofs are different). Finally, Assumption II.3(3) controls the magnitude of the local basis in a uniform sense.

Assumptions II.2 and II.3 implicitly relate the number of approximating series terms, the number of knots used and the maximum mesh size: $K \asymp \bar{\kappa} \asymp h^{-d}$. By restricting the growth rate of these tuning parameters, the least squares partitioning-based estimator satisfying the above conditions is well-defined in large samples. We next state a high-level requirement that gives explicit expression of the leading approximation error. For each $\mathbf{x}\in\mathcal{X}$, let $\delta_{\mathbf{x}}$ be the element of Δ whose closure contains $\mathbf{x}$ and $h_{\mathbf{x}}$ for the diameter of this $\delta_{\mathbf{x}}$.

Assumption II.4 (Approximation Error)**.** *For all* $\boldsymbol{\varsigma}$ *satisfying* $[\boldsymbol{\varsigma}]\le\varsigma$*, given in Assumption II.3, there exists* $s^*\in\mathcal{S}_{\Delta,m}$*, the linear span of* $\mathbf{p}(\mathbf{x};\Delta,m)$*, and*

$$\mathscr{B}_{m,\boldsymbol{\varsigma}}(\mathbf{x}) = -\sum_{\boldsymbol{u}\in\Lambda_m} \partial^{\boldsymbol{u}}\theta(\mathbf{x})h_{\mathbf{x}}^{m-[\boldsymbol{\varsigma}]}B_{\boldsymbol{u},\boldsymbol{\varsigma}}(\mathbf{x})$$

such that

$$\sup_{\mathbf{x}\in\mathcal{X}} |\partial^{\boldsymbol{\varsigma}}\theta(\mathbf{x}) - \partial^{\boldsymbol{\varsigma}}s^*(\mathbf{x}) + \mathscr{B}_{m,\boldsymbol{\varsigma}}(\mathbf{x})| \lesssim h^{m+\varrho-[\boldsymbol{\varsigma}]} \tag{2.3}$$

and

$$\sup_{\delta\in\Delta}\sup_{\mathbf{x}_1,\mathbf{x}_2\in\mathrm{clo}(\delta)}\frac{|B_{\boldsymbol{u},\boldsymbol{\varsigma}}(\mathbf{x}_1)-B_{\boldsymbol{u},\boldsymbol{\varsigma}}(\mathbf{x}_2)|}{\|\mathbf{x}_1-\mathbf{x}_2\|}\lesssim h^{-1} \qquad (2.4)$$

where $B_{\boldsymbol{u},\boldsymbol{\varsigma}}(\cdot)$ is a known function that is bounded uniformly over n, and Λ_m is a multi-index set, which depends on the basis, with $[\boldsymbol{u}]=m$ for $\boldsymbol{u}\in\Lambda_m$.

More common, nonspecific rate assumptions such as $\sup_{\mathbf{x}\in\mathcal{X}}|\partial^{\mathbf{q}}\theta(\mathbf{x})-\partial^{\mathbf{q}}s^*(\mathbf{x})|\lesssim h^{m-[\mathbf{q}]}$ will not suffice for our bias correction and IMSE expansion results; (2.3) is needed. The rate-only version is implied by our assumptions. The terms $B_{\boldsymbol{u},\boldsymbol{\varsigma}}(\cdot)$ in $\mathscr{B}_{m,\boldsymbol{\varsigma}}(\cdot)$ are known functions of the point $\mathbf{x}$ which depend on the particular partitioning scheme and bases used. The only unknowns in the approximation error $\mathscr{B}_{m,\boldsymbol{\varsigma}}(\cdot)$ are the higher-order derivatives of $\theta(\cdot)$. In Chapter IV we verify this (and the other assumptions) for splines, wavelets, and piecewise polynomials, including explicit formulas for the leading error in (2.3) and give precise characterizations of Λ_m. We assume sufficient smoothness exists to characterize these terms: see Calonico, Cattaneo, and Farrell (2018b) for a discussion when smoothness constrains inference.

The function $\mathscr{B}_{m,\boldsymbol{\varsigma}}(\cdot)$ is understood as the approximation error in L_∞ norm, and is not in general the misspecification (or smoothing) bias of a series estimator. In least squares series regression settings, the leading smoothing bias is described by two terms in general: $\mathscr{B}_{m,\boldsymbol{\varsigma}}(\cdot)$ and the accompanying error from the linear projection of $\mathscr{B}_{m,\mathbf{0}}(\cdot)$ onto $\mathcal{S}_{\Delta,m}$. We formalize this result in Lemma II.1 below. The second bias term is often ignored in the literature because in several cases the leading approximation error $\mathscr{B}_{m,\mathbf{0}}(\cdot)$ is *approximately orthogonal* to $\mathbf{p}(\cdot)$ with respect to the Lebesgue measure, that is, if

$$\max_{1\leq k\leq K}\int_{\mathcal{H}_k}p_k(\mathbf{x};\Delta,m)\mathscr{B}_{m,\mathbf{0}}(\mathbf{x})\,d\mathbf{x}=o(h^{m+d}), \qquad (2.5)$$

under Assumptions II.1–II.4. In some simple cases, (2.5) is automatically satisfied if one constructs the leading error based on a basis representing the orthogonal complement of $\mathcal{S}_{\Delta,m}$. When (2.5) holds, the leading term in L_∞ approximation error coincides with the leading misspecification (or smoothing) bias of a partitioning-based series estimator. When a stronger quasi-uniformity condition holds (i.e., neighboring cells are of the same size asymptotically), a sufficient condition for (2.5) is simply the orthogonality between $B_{\boldsymbol{u},\mathbf{0}}(\cdot)$ and $\mathbf{p}(\cdot)$ in L_2 with respect to the Lebesgue measure, for all $\boldsymbol{u}\in\Lambda_m$.

For general partitioning-based estimators this orthogonality need not hold. For example, (2.5) is hard to verify when the partitioning employed is sufficiently uneven, as is usually the case when employing machine learning methods. All our main results hold when this orthogonality fails, and importantly, our bias correction methods and IMSE expansion explicitly account for the L_2 projection of $\mathscr{B}_{m,\mathbf{0}}(\cdot)$ onto the approximating space spanned by $\mathbf{p}(\cdot)$.

2.3 Characterization and Correction of Bias

We now precisely characterize the bias of $\widehat{\partial^{\mathbf{q}}\theta}(\mathbf{x})$ under Assumptions II.1–II.4, but not assuming (2.5). Then, using this result, we develop valid IMSE expansions and three robust bias-corrected inference procedures. This section focuses on bias correction, and Section 2.5 presents the associated robust Studentization adjustments for inference, following the ideas in Calonico, Cattaneo, and Farrell (2018b) for kernel-based nonparametrics.

Given our assumptions, the estimator $\widehat{\partial^{\mathbf{q}}\theta}(\mathbf{x})$ of (2.2) can be written as

$$\widehat{\partial^{\mathbf{q}}\theta}_0(\mathbf{x}) := \widehat{\boldsymbol{\gamma}}_{\mathbf{q},0}(\mathbf{x})'\mathbb{E}_n[\boldsymbol{\Pi}_0(\mathbf{x}_i)y_i], \tag{2.6}$$

where

$$\widehat{\boldsymbol{\gamma}}_{\mathbf{q},0}(\mathbf{x})' := \partial^{\mathbf{q}}\mathbf{p}(\mathbf{x})'\mathbb{E}_n[\mathbf{p}(\mathbf{x}_i)\mathbf{p}(\mathbf{x}_i)']^{-1} \quad \text{and} \quad \boldsymbol{\Pi}_0(\mathbf{x}_i) := \mathbf{p}(\mathbf{x}_i).$$

The subscript of “0” will differentiate this estimator from the bias-corrected versions below. We first give a preliminary result.

Lemma II.1 (Conditional Bias). *Let Assumptions II.1, II.2, II.3, and II.4 hold. If* $\frac{\log n}{nh^d} = o(1)$, *then*

$$\begin{aligned}
&\mathbb{E}[\widehat{\partial^{\mathbf{q}}\theta}_0(\mathbf{x})|\mathbf{X}] - \partial^{\mathbf{q}}\theta(\mathbf{x}) \\
&\quad = \widehat{\boldsymbol{\gamma}}_{\mathbf{q},0}(\mathbf{x})'\mathbb{E}_n[\boldsymbol{\Pi}_0(\mathbf{x}_i)\theta(\mathbf{x}_i)] - \partial^{\mathbf{q}}\theta(\mathbf{x}) && (2.7)\\
&\quad = \mathscr{B}_{m,\mathbf{q}}(\mathbf{x}) - \widehat{\boldsymbol{\gamma}}_{\mathbf{q},0}(\mathbf{x})'\mathbb{E}_n[\boldsymbol{\Pi}_0(\mathbf{x}_i)\mathscr{B}_{m,\mathbf{0}}(\mathbf{x}_i)] + O_{\mathbb{P}}(h^{m+\varrho-[\mathbf{q}]}). && (2.8)
\end{aligned}$$

The proof of this lemma generalizes an idea in Zhou, Shen, and Wolfe (1998, Theorem 2.2) to handle partitioning-based series estimators beyond the specific example of B-Splines on tensor-product partitions. The first component $\mathscr{B}_{m,\mathbf{q}}(\mathbf{x})$ is the leading term in the asymptotic error expansion and depends on the function space generated by the series employed. The second component comes from the least squares regression, and it can be interpreted as the projection of the leading approximation error onto the space spanned by the basis employed. Because the approximating basis $\mathbf{p}(\mathbf{x})$ is locally supported (Assumption II.3), the orthogonality condition in (2.5), when it holds, suffices to guarantee that the projection of leading error is of smaller order (such as for B-splines on a tensor-product partition). In general the bias will be $O(h^{m-[\mathbf{q}]})$ and further, in finite samples both terms may be important even if (2.5) holds.

We consider three bias correction methods to remove the leading bias terms of Lemma II.1. All three methods rely, in one way or another, on a higher order basis: for some $\tilde{m} > m$, let $\tilde{\mathbf{p}}(\mathbf{x}) := \tilde{\mathbf{p}}(\mathbf{x}; \tilde{\Delta}, \tilde{m})$ be a basis of order $\tilde{m}$ defined on partition $\tilde{\Delta}$ which has maximum mesh $\tilde{h}$. Objects accented with a tilde always pertain to this secondary basis and partition for bias correction. In practice, a simple choice is $\tilde{m} = m + 1$ and $\tilde{\Delta} = \Delta$.

The first, and most obvious approach, is simply to use the higher order basis in place of the original basis (c.f., Huang, 2003, Section 5.3). This is thus named *higher-order-basis bias correction* and numbered as approach $j = 1$. In complete parallel to (2.6) define

$$\widehat{\partial^{\mathbf{q}}\theta}_1(\mathbf{x}) := \widehat{\boldsymbol{\gamma}}_{\mathbf{q},1}(\mathbf{x})'\mathbb{E}_n[\boldsymbol{\Pi}_1(\mathbf{x}_i)y_i], \tag{2.9}$$

where

$$\widehat{\boldsymbol{\gamma}}_{\mathbf{q},1}(\mathbf{x})' := \partial^{\mathbf{q}}\tilde{\mathbf{p}}(\mathbf{x})'\mathbb{E}_n[\tilde{\mathbf{p}}(\mathbf{x}_i)\tilde{\mathbf{p}}(\mathbf{x}_i)']^{-1} \quad \text{and} \quad \boldsymbol{\Pi}_1(\mathbf{x}_i) := \tilde{\mathbf{p}}(\mathbf{x}_i).$$

This approach can be viewed as a bias correction of the original point estimator because, trivially, $\widehat{\partial^{\mathbf{q}}\theta}_1(\mathbf{x}) = \widehat{\partial^{\mathbf{q}}\theta}_0(\mathbf{x}) - (\widehat{\partial^{\mathbf{q}}\theta}_0(\mathbf{x}) - \widehat{\partial^{\mathbf{q}}\theta}_1(\mathbf{x}))$. Valid inference based on $\widehat{\partial^{\mathbf{q}}\theta}_1(\mathbf{x})$ can be viewed as "undersmoothing" applied to the higher-order point estimator, but is distinct from undersmoothing $\widehat{\partial^{\mathbf{q}}\theta}_0(\mathbf{x})$ (i.e., using a finer partition Δ and keeping the order fixed). Huang (2003) used this idea to remove the asymptotic bias of splines estimators.

Our second approach makes use of the generic expression of the least squares bias in (2.7). The unknown objects in this expression are θ and $\partial^{\mathbf{q}}\theta$, both of which can be estimated using the higher-order estimator (2.9). By plugging these into (2.7) and subtracting the result from $\widehat{\partial^{\mathbf{q}}\theta}_0(\mathbf{x})$, we obtain the *least-squares bias correction*, numbered as approach 2:

$$\begin{aligned}\widehat{\partial^{\mathbf{q}}\theta}_2(\mathbf{x}) &:= \widehat{\partial^{\mathbf{q}}\theta}_0(\mathbf{x}) - \Big(\widehat{\boldsymbol{\gamma}}_{\mathbf{q},0}(\mathbf{x})'\mathbb{E}_n[\boldsymbol{\Pi}_0(\mathbf{x}_i)\widehat{\theta}_1(\mathbf{x}_i)] - \widehat{\partial^{\mathbf{q}}\theta}_1(\mathbf{x})\Big) \qquad (2.10)\\ &:= \widehat{\boldsymbol{\gamma}}_{\mathbf{q},2}(\mathbf{x})'\mathbb{E}_n[\boldsymbol{\Pi}_2(\mathbf{x}_i)y_i]\end{aligned}$$

where

$$\begin{aligned}
\widehat{\gamma}_{\mathbf{q},2}(\mathbf{x})' &:= \Big(\widehat{\gamma}_{\mathbf{q},0}(\mathbf{x})', -\widehat{\gamma}_{\mathbf{q},0}(\mathbf{x})'\mathbb{E}_n[\mathbf{p}(\mathbf{x}_i)\tilde{\mathbf{p}}(\mathbf{x}_i)']\mathbb{E}_n[\tilde{\mathbf{p}}(\mathbf{x}_i)\tilde{\mathbf{p}}(\mathbf{x}_i)']^{-1} + \widehat{\gamma}_{\mathbf{q},1}(\mathbf{x})'\Big) \\
\text{and} \qquad \mathbf{\Pi}_2(\mathbf{x}_i) &:= (\mathbf{p}(\mathbf{x}_i)', \tilde{\mathbf{p}}(\mathbf{x}_i)')',
\end{aligned}$$

which is exactly of the same form as $\widehat{\partial^{\mathbf{q}}\theta}_0(\mathbf{x})$ and $\widehat{\partial^{\mathbf{q}}\theta}_1(\mathbf{x})$ (cf., (2.6) and (2.9)), except for the change in $\widehat{\gamma}_{\mathbf{q},j}(\mathbf{x})$ and $\mathbf{\Pi}_j(\mathbf{x}_i)$.

Finally, approach number 3 targets the leading terms identified in Equation (2.8). We dub this approach *plug-in bias correction*, as it specifically estimates the leading bias terms, in fixed-n form, of $\widehat{\partial^{\mathbf{q}}\theta}_0(\mathbf{x})$ according to Assumption II.4. To be precise, we employ the explicit plug-in bias estimator

$$\widehat{\mathscr{B}}_{m,\mathbf{q}}(\mathbf{x}) = -\sum_{\boldsymbol{u}\in\Lambda_m}\Big(\partial^{\boldsymbol{u}}\widehat{\theta}_1(\mathbf{x})\Big)h_{\mathbf{x}}^{m-[\mathbf{q}]}B_{\boldsymbol{u},\mathbf{q}}(\mathbf{x}),$$

with $[\mathbf{q}] < m$ and Λ_m as in Assumption II.4, leading to

$$\begin{aligned}
\widehat{\partial^{\mathbf{q}}\theta}_3(\mathbf{x}) &:= \widehat{\partial^{\mathbf{q}}\theta}_0(\mathbf{x}) - \Big(\widehat{\mathscr{B}}_{m,\mathbf{q}}(\mathbf{x}) - \widehat{\gamma}_{\mathbf{q},0}(\mathbf{x})'\mathbb{E}_n[\mathbf{\Pi}_0(\mathbf{x}_i)\widehat{\mathscr{B}}_{m,\mathbf{0}}(\mathbf{x}_i)]\Big) \qquad (2.11) \\
&:= \widehat{\gamma}_{\mathbf{q},3}(\mathbf{x})'\mathbb{E}_n[\mathbf{\Pi}_3(\mathbf{x}_i)y_i]
\end{aligned}$$

where

$$\begin{aligned}
\widehat{\gamma}_{\mathbf{q},3}(\mathbf{x})' = \Big(\widehat{\gamma}_{\mathbf{q},0}(\mathbf{x})', \sum_{\boldsymbol{u}\in\Lambda_m}\Big\{&\widehat{\gamma}_{\boldsymbol{u},1}(\mathbf{x})'h_{\mathbf{x}}^{m-[\mathbf{q}]}B_{\boldsymbol{u},\mathbf{q}}(\mathbf{x}) \\
&- \widehat{\gamma}_{\mathbf{q},0}(\mathbf{x})'\mathbb{E}_n[\mathbf{p}(\mathbf{x}_i)h_{\mathbf{x}_i}^{m}B_{\boldsymbol{u},\mathbf{0}}(\mathbf{x}_i)\widehat{\gamma}_{\boldsymbol{u},1}(\mathbf{x}_i)']\Big\}\Big), \\
\text{and} \qquad \mathbf{\Pi}_3(\mathbf{x}_i) := (\mathbf{p}(\mathbf{x}_i)', \tilde{\mathbf{p}}(\mathbf{x}_i)')'.&
\end{aligned}$$

When the orthogonality condition (2.5) holds, the second correction term in $\widehat{\partial^{\mathbf{q}}\theta}_3(\mathbf{x})$ is asymptotically negligible relative to the first. However, in finite samples both terms can be important, so we consider the general case.

Our results employing bias correction will require the following conditions on the higher-order basis used for bias estimation.

Assumption II.5 (Bias Correction)**.** *The partition $\tilde{\Delta}$ satisfies Assumption II.2, with maximum mesh $\tilde{h}$ and the basis $\tilde{\mathbf{p}}(\mathbf{x}; \tilde{\Delta}, \tilde{m})$ satisfies Assumptions II.3 and II.4 with $\tilde{\varsigma} = \tilde{\varsigma}(\tilde{m}) \geq m$ in place of ς. Let $\rho := h/\tilde{h}$, which obeys $\rho \to \rho_0 \in (0, \infty)$. In addition, for $j = 3$, either (i) $\tilde{\mathbf{p}}(\mathbf{x}; \tilde{\Delta}, \tilde{m})$ spans a space containing the span of $\mathbf{p}(\mathbf{x}; \Delta, m)$, and for all $\boldsymbol{u} \in \Lambda_m$, $\partial^{\boldsymbol{u}}\mathbf{p}(\mathbf{x}; \Delta, m) = \mathbf{0}$; or (ii) both $\mathbf{p}(\mathbf{x}; \Delta, m)$ and $\tilde{\mathbf{p}}(\mathbf{x}; \tilde{\Delta}, \widetilde{m})$ reproduce polynomials of degree $[\mathbf{q}]$.*

In addition to removing the leading bias, the conditions in Assumption II.5 require that the asymptotic variance of bias-corrected estimators is properly bounded from below in a uniform sense, which is critical for inference. Additional conditions are required for plug-in bias correction ($j = 3$) due to the more complicated covariance between $\widehat{\partial^{\mathbf{q}}\theta}_0$ and the estimated leading bias. Orthogonality properties due to the projection structure of the least squares bias correction ($j = 2$) removes these "covariance" components in the variance of $\widehat{\partial^{\mathbf{q}}\theta}_2$. The natural choice of $\tilde{\Delta} = \Delta$ and $\tilde{m} = m + 1$ will satisfy this condition on intuitive conditions.

2.4 IMSE and Convergence Rates

We establish two main results related to the point estimator $\widehat{\partial^{\mathbf{q}}\theta}_0(\mathbf{x})$. First, we obtain valid IMSE expansions for the estimator, which also give as a by-product an estimate of its L_2 convergence rate. Second, we establish the uniform convergence rate of the estimator.

2.4.1 IMSE-Optimal Point Estimation

We first give a very general IMSE approximation, which we will specialize in Chapter IV to a more detailed result for the special case of a tensor-product partition. These expansions are used to obtain optimal choices of partition size from a point estimation perspective, which is important for implementation of partitioning-based nonparametric estimation and inference.

A chief advantage of the robust bias corrected inference methods that we develop in the upcoming sections is that IMSE-optimal tuning parameters (and related choices such as those obtained from cross-validation) are valid for inference, which is not the case for the standard approach unless ad-hoc undersmoothing is used. This allows researchers to combine an optimal estimate of the function, $\widehat{\partial^{\mathbf{q}}\theta}_0(\cdot)$ based on the IMSE-optimal $h_{\mathtt{IMSE}} \asymp n^{-1/(2m+d)}$, and its plug-in or cross-validation implementations thereof, with inference based on the same tuning parameter choices (and hence employing the same partitioning scheme).

Our first result holds for any partition Δ satisfying Assumption II.2.

Theorem II.1 (IMSE)**.** *Let Assumptions II.1, II.2, II.3, and II.4 hold. If* $\frac{\log n}{nh^d} = o(1)$, *then for a weighting function* $w(\mathbf{x})$ *that is continuous and bounded away from zero on*

$\mathcal{X}$,

$$\begin{aligned}\int_{\mathcal{X}} &\mathbb{E}[(\widehat{\partial^{\mathbf{q}}\theta}_0(\mathbf{x}) - \partial^{\mathbf{q}}\theta(\mathbf{x}))^2|\mathbf{X}]w(\mathbf{x})\,d\mathbf{x} \\ &= \frac{1}{n}\Big(\mathscr{V}_{\Delta,\mathbf{q}} + o_{\mathbb{P}}(h^{-d-2[\mathbf{q}]})\Big) + \Big(\mathscr{B}_{\Delta,\mathbf{q}} + o_{\mathbb{P}}(h^{2m-2[\mathbf{q}]})\Big)\end{aligned}$$

where

$$\mathscr{V}_{\Delta,\mathbf{q}} = \mathrm{trace}\left(\boldsymbol{\Sigma}_0 \int_{\mathcal{X}} \boldsymbol{\gamma}_{\mathbf{q},0}(\mathbf{x})\boldsymbol{\gamma}_{\mathbf{q},0}(\mathbf{x})' w(\mathbf{x})d\mathbf{x}\right) \asymp h^{-d-2[\mathbf{q}]},$$

$$\mathscr{B}_{\Delta,\mathbf{q}} = \int_{\mathcal{X}} \Big(\mathscr{B}_{m,\mathbf{q}}(\mathbf{x}) - \boldsymbol{\gamma}_{\mathbf{q},0}(\mathbf{x})'\mathbb{E}[\mathbf{p}(\mathbf{x}_i)\mathscr{B}_{m,\mathbf{0}}(\mathbf{x}_i)]\Big)^2 w(\mathbf{x})d\mathbf{x} \lesssim h^{2m-2[\mathbf{q}]},$$

$\boldsymbol{\Sigma}_0 := \mathbb{E}[\boldsymbol{\Pi}_0(\mathbf{x}_i)\boldsymbol{\Pi}_0(\mathbf{x}_i)'\sigma^2(\mathbf{x}_i)]$, *and* $\boldsymbol{\gamma}_{\mathbf{q},0}(\mathbf{x})' := \partial^{\mathbf{q}}\mathbf{p}(\mathbf{x})'\mathbb{E}[\mathbf{p}(\mathbf{x}_i)\mathbf{p}(\mathbf{x}_i)']^{-1}$.

This theorem shows that the leading term in the integrated (and pointwise) variance of $\widehat{\partial^{\mathbf{q}}\theta}_0(\mathbf{x})$ is of order $n^{-1}h^{-d-2[\mathbf{q}]}$. For the bias term, on the other hand, the theorem only establishes an upper bound: to bound the bias component from below, stronger conditions on the regression function would be needed. It is easy to see that this rate bound is sharp in general.

The quantities $\mathscr{V}_{\Delta,\mathbf{q}}$ and $\mathscr{B}_{\Delta,\mathbf{q}}$ are nonrandom sequences depending on the partitioning scheme Δ in a complicated way, and need not converge as $h \to 0$. Nevertheless, when the integrated squared bias does not vanish ($\mathscr{B}_{\Delta,\mathbf{q}} \neq 0$), Theorem II.1 implies that the IMSE-optimal mesh size $h_{\mathtt{IMSE}}$ is proportional to $n^{-1/(2m+d)}$, or equivalently, the IMSE-optimal number of series terms $K_{\mathtt{IMSE}} \asymp n^{d/(2m+d)}$. Furthermore, because the IMSE expansion is obtained for a given partition scheme, the result in Theorem II.1 can be used to evaluate different partitioning schemes altogether, and to select the "optimal" one in an IMSE sense. We can consider the optimization problem

$$\min_{\Delta\in\mathcal{D}} \left\{\frac{1}{n}\mathscr{V}_{\Delta,\mathbf{q}} + \mathscr{B}_{\Delta,\mathbf{q}}\right\}$$

as a way of selecting an "optimal" partitioning scheme among some class of partitioning schemes $\mathcal{D}$.

Theorem II.1 generalizes prior work substantially. Existing results cover only special cases, such as piecewise polynomials (Cattaneo and Farrell, 2013) or splines (Agarwal and Studden, 1980; Zhou, Shen, and Wolfe, 1998; Zhou and Wolfe, 2000) on tensor-product partitions only, and often restricting to $d = 1$ or $[\mathbf{q}] = 0$. To the best of our knowledge, covering non-tensor-product partitions and other series functions such as wavelets is new to the literature.

The leading constants at this level of generality is quite involved. To illustrate the usefulness of this result in applications, we will consider the special case of a tensor-product partition in Chapter IV.

2.4.2 Convergence Rates

Theorem II.1 immediately delivers the L_2 convergence rate for the point estimator $\widehat{\partial^{\mathbf{q}}\theta}_0(\mathbf{x})$. For completeness, we also establish its uniform convergence rate. Recall that $\nu > 0$.

Theorem II.2 (Convergence Rates)**.** *Let Assumptions II.1, II.2 and II.3 hold. Assume also that* $\sup_{\mathbf{x}\in\mathcal{X}} |\partial^{\mathbf{q}}\theta(\mathbf{x}) - \partial^{\mathbf{q}}s^*(\mathbf{x})| \lesssim h^{m-[\mathbf{q}]}$ *with* s^* *defined in Assumption II.4. Then, if* $\frac{\log n}{nh^d} = o(1)$*,*

$$\int_{\mathcal{X}} \left(\widehat{\partial^{\mathbf{q}}\theta}_0(\mathbf{x}) - \partial^{\mathbf{q}}\theta(\mathbf{x})\right)^2 w(\mathbf{x})\, d\mathbf{x} \lesssim_{\mathbb{P}} \frac{1}{nh^{d+2[\mathbf{q}]}} + h^{2(m-[\mathbf{q}])}$$

If, in addition,

(i) $\sup_{\mathbf{x}\in\mathcal{X}} \mathbb{E}[|\varepsilon_i|^3 \exp(|\varepsilon_i|)] < \infty$ *and* $\frac{(\log n)^3}{nh^d} \lesssim 1$*, or*

(ii) $\sup_{\mathbf{x}\in\mathcal{X}} \mathbb{E}[|\varepsilon_i|^{2+\nu}] < \infty$ *and* $\frac{n^{\frac{2}{2+\nu}}(\log n)^{\frac{2\nu}{4+2\nu}}}{nh^d} \lesssim 1$*,*

then

$$\sup_{\mathbf{x}\in\mathcal{X}} \left|\widehat{\partial^{\mathbf{q}}\theta}_0(\mathbf{x}) - \partial^{\mathbf{q}}\theta(\mathbf{x})\right|^2 \lesssim_{\mathbb{P}} \frac{\log n}{nh^{d+2[\mathbf{q}]}} + h^{2(m-[\mathbf{q}])}.$$

This theorem shows that the partitioning-based estimators can attain the optimal mean-square and uniform convergence rate (Stone, 1982) by proper choice of partitioning scheme, under our high-level assumptions. (The full force of Assumption II.4 is not needed for this result.) Cattaneo and Farrell (2013) were the first to show existence of a series estimator (in particular, piecewise polynomials) attaining the optimal uniform convergence rate, a result that was later generalized to other series estimators in Belloni, Chernozhukov, Chetverikov, and Kato (2015); Chen and Christensen (2015) under various alternative high-level assumptions.

2.5 Pointwise Inference

We give pointwise inference based on classical undersmoothing and all three bias correction methods. All four point estimators take the form

$$\widehat{\partial^{\mathbf{q}}\theta}_j(\mathbf{x}) = \widehat{\gamma}_{\mathbf{q},j}(\mathbf{x})'\mathbb{E}_n[\mathbf{\Pi}_j(\mathbf{x}_i)y_i],$$

where $j = 0$ corresponds to the conventional partitioning estimator, and $j = 1, 2, 3$ refer to the three distinct bias correction strategies. Infeasible inference would be based on the standardized t-statistics

$$T_j(\mathbf{x}) = \frac{\widehat{\partial^{\mathbf{q}}\theta}_j(\mathbf{x}) - \partial^{\mathbf{q}}\theta(\mathbf{x})}{\sqrt{\Omega_j(\mathbf{x})/n}}, \qquad \Omega_j(\mathbf{x}) = \boldsymbol{\gamma}_{\mathbf{q},j}(\mathbf{x})'\boldsymbol{\Sigma}_j\boldsymbol{\gamma}_{\mathbf{q},j}(\mathbf{x}),$$

where, for each $j = 0, 1, 2, 3$, $\boldsymbol{\gamma}_{\mathbf{q},j}(\mathbf{x})$ are defined as $\widehat{\boldsymbol{\gamma}}_{\mathbf{q},j}$ in (2.6), (2.9), (2.10), and (2.11), respectively, but with sample averages and other estimators replaced by their population counterparts, and $\boldsymbol{\Sigma}_j := \mathbb{E}[\boldsymbol{\Pi}_j(\mathbf{x}_i)\boldsymbol{\Pi}_j(\mathbf{x}_i)'\sigma^2(\mathbf{x}_i)]$. These t-statistics are infeasible, but they nonetheless capture the additional variability introduced by the bias correction approach when $j = 1, 2, 3$, the key idea behind robust bias corrected inference (Calonico, Cattaneo, and Titiunik, 2014; Calonico, Cattaneo, and Farrell, 2018b). We also discuss below Studentization, that is, replacing $\Omega_j(\mathbf{x})$ with a consistent estimator.

2.5.1 Distributional Approximation

The first result gives the limiting distribution of the standardized t-statistics $T_j(\mathbf{x})$.

Theorem II.3 (Asymptotic Normality)**.** *Let Assumptions II.1, II.2, II.3, and II.4 hold. Assume* $\sup_{\mathbf{x}\in\mathcal{X}} \mathbb{E}[\varepsilon_i^2 \mathbb{1}\{|\varepsilon_i| > M\}|\mathbf{x}_i = \mathbf{x}] \to 0$ *as* $M \to \infty$*, and* $\frac{\log n}{nh^d} = o(1)$*. Furthermore, for* $j = 0$*, assume* $nh^{2m+d} = o(1)$*; and for* $j = 1, 2, 3$*, assume Assumption II.5 holds and* $nh^{2m+d} \lesssim 1$*.*

Then, for each $j = 0, 1, 2, 3$ *and* $\mathbf{x} \in \mathcal{X}$*,* $\sup_{u\in\mathbb{R}} |\mathbb{P}[T_j(\mathbf{x}) \leq u] - \Phi(u)| = o(1)$*, where* $\Phi(u)$ *denotes the cumulative distribution function of* $\mathsf{N}(0, 1)$*.*

This theorem gives a valid Gaussian approximation for the t-statistics $T_j(\mathbf{x})$, pointwise in $\mathbf{x} \in \mathcal{X}$. The regularity conditions imposed are extremely mild, and in perfect quantitative agreement with those used in Belloni, Chernozhukov, Chetverikov, and Kato (2015) for $j = 0$ (undersmoothing). For $j = 1, 2, 3$ (robust bias correction), the result is new to the literature, and the restrictions are in perfect qualitative agreement with those obtained in Calonico, Cattaneo, and Farrell (2018b) for kernel-based nonparametrics.

2.5.2 Implementation

To make the results in Theorem II.3 feasible, we replace $\Omega_j(\mathbf{x})$ with a consistent estimator. Specifically, we consider the four feasible t-statistics, $j = 0, 1, 2, 3$,

$$\begin{aligned} \widehat{T}_j(\mathbf{x}) &= \frac{\widehat{\partial^{\mathbf{q}}\theta}_j(\mathbf{x}) - \partial^{\mathbf{q}}\theta(\mathbf{x})}{\sqrt{\widehat{\Omega}_j(\mathbf{x})/n}}, & \widehat{\Omega}_j(\mathbf{x}) &= \widehat{\boldsymbol{\gamma}}_{\mathbf{q},j}(\mathbf{x})'\widehat{\boldsymbol{\Sigma}}_j\widehat{\boldsymbol{\gamma}}_{\mathbf{q},j}(\mathbf{x}), \\ \widehat{\boldsymbol{\Sigma}}_j &= \mathbb{E}_n[\boldsymbol{\Pi}_j(\mathbf{x}_i)\boldsymbol{\Pi}_j(\mathbf{x}_i)'\widehat{\varepsilon}^2_{i,j}], & \widehat{\varepsilon}_{i,j} &= y_i - \widehat{\theta}_j(\mathbf{x}_i), \end{aligned} \tag{2.12}$$

Once the basis functions and partitioning schemes are chosen, the statistic $\widehat{T}_j(\mathbf{x})$ is readily implementable. The following theorem gives sufficient conditions for valid pointwise inference.

Theorem II.4 (Variance Consistency)**.** *Let Assumptions II.1, II.2, II.3, and II.4 hold. If* $j = 1, 2, 3$, *also let Assumption II.5 hold. In addition, assume one of the following holds:*

(i) $\sup_{\mathbf{x}\in\mathcal{X}} \mathbb{E}[|\varepsilon_i|^{2+\nu}] < \infty$ *and* $\frac{n^{\frac{2}{2+\nu}}(\log n)^{\frac{2\nu}{4+2\nu}}}{nh^d} = o(1)$, *or*

(ii) $\sup_{\mathbf{x}\in\mathcal{X}} \mathbb{E}[|\varepsilon_i|^3 \exp(|\varepsilon_i|)] < \infty$ *and* $\frac{(\log n)^3}{nh^d} = o(1)$.

Then, for each $j = 0, 1, 2, 3$, $|\widehat{\Omega}_j(\mathbf{x}) - \Omega_j(\mathbf{x})| = o_{\mathbb{P}}(h^{-d-2[\mathbf{q}]})$.

This result together with Theorem II.3, delivers feasible inference. Valid $100(1-\alpha)\%$, $\alpha \in (0, 1)$, confidence intervals for $\partial^{\mathbf{q}}\theta(\mathbf{x})$ are formed in the usual way:

$$\left[\, \widehat{\partial^{\mathbf{q}}\theta}_j(\mathbf{x}) \pm \Phi^{-1}(1-\alpha/2)\cdot\sqrt{\widehat{\Omega}_j(\mathbf{x})/n}\,\right], \qquad j = 0, 1, 2, 3.$$

Importantly, for $j = 1, 2, 3$, the IMSE-optimal partitioning scheme choice derived in Section 2.4 (or related methods like cross-validation) can be used directly, while for $j = 0$ the partitioning has to be undersmoothed (i.e., made finer than the IMSE-optimal choice) in order to obtain valid confidence intervals. See Calonico, Cattaneo, and Farrell (2018b) for more discussion.

2.6 Uniform Inference

We next give a valid distributional approximation for the *whole* process $\{\widehat{T}_j(\mathbf{x}) : \mathbf{x} \in \mathcal{X}\}$, for each $j = 0, 1, 2, 3$. We establish this approximation using two distinct coupling strategies. We then propose a simulation-based feasible implementation of the result. We close by applying our results to construct valid confidence bands for $\partial^{\mathbf{q}}\theta(\cdot)$.

2.6.1 Strong Approximations

The stochastic processes $\{\widehat{T}_j(\mathbf{x}) : \mathbf{x} \in \mathcal{X}\}$ are not asymptotically tight, and therefore do not converge weakly in $\mathcal{L}^\infty(\mathcal{X})$, where $\mathcal{L}^\infty(\mathcal{X})$ denotes the set of all (uniformly) bounded real functions on $\mathcal{X}$ equipped with uniform norm. Nevertheless, their finite sample distribution can be approximated by carefully constructed Gaussian processes (in a possibly enlarged probability space).

We first employ the following lemma to simplify the problem. Recall that r_n is some non-vanishing positive sequence and $\nu > 0$.

Lemma II.2 (Hats Off)**.** *Let Assumptions II.1, II.2, II.3, and II.4 hold. Assume one of the following holds:*

(i) $\sup_{\mathbf{x}\in\mathcal{X}} \mathbb{E}[|\varepsilon_i|^{2+\nu}|\mathbf{x}_i = \mathbf{x}] < \infty$ *and* $\frac{n^{\frac{2}{2+\nu}}(\log n)^{\frac{2+2\nu}{2+\nu}}}{nh^d} = o(r_n^{-2})$*; or*

(ii) $\sup_{\mathbf{x}\in\mathcal{X}} \mathbb{E}[|\varepsilon_i|^3 \exp(|\varepsilon_i|)|\mathbf{x}_i = \mathbf{x}] < \infty$ *and* $\frac{(\log n)^4}{nh^d} = o(r_n^{-2})$.

Furthermore, if $j = 0$, assume $nh^{d+2m} = o(r_n^{-2})$; and, if $j = 1, 2, 3$, assume Assumption II.5 holds and $nh^{d+2m+2\varrho} = o(r_n^{-2})$. Then

$$\sup_{\mathbf{x}\in\mathcal{X}} \left|\widehat{T}_j(\mathbf{x}) - t_j(\mathbf{x})\right| = o_{\mathbb{P}}(r_n^{-1}), \quad t_j(\mathbf{x}) = \frac{\boldsymbol{\gamma}_{\mathbf{q},j}(\mathbf{x})'}{\sqrt{\Omega_j(\mathbf{x})}}\mathbb{G}_n[\mathbf{\Pi}_j(\mathbf{x}_i)\varepsilon_i].$$

Lemma II.2 requires that the estimation and sampling uncertainty of $\widehat{\boldsymbol{\gamma}}_{\mathbf{q},j}$ and $\widehat{\Omega}_j(\mathbf{x})$, as well as the smoothing bias of $\widehat{\partial^{\mathbf{q}}\theta}_j(\mathbf{x})$, be negligible uniformly over $\mathbf{x} \in \mathcal{X}$. Its proof relies on some new technical lemmas, but is otherwise standard. This technical approximation step allows us to focus on developing a distributional approximation for the infeasible stochastic processes $\{t_j(\mathbf{x}) : \mathbf{x} \in \mathcal{X}\}$, $j = 0, 1, 2, 3$. We make precise our uniform distributional approximation in the following definition.

Definition II.1 (Strong Approximation)**.** For each $j = 0, 1, 2, 3$, the law of the stochastic process $\{t_j(\mathbf{x}), \mathbf{x} \in \mathcal{X}\}$ is approximated by that of a Gaussian process $\{Z_j(\mathbf{x}), \mathbf{x} \in \mathcal{X}\}$ in $\mathcal{L}^\infty(\mathcal{X})$ if the following condition holds: in a sufficiently rich probability space, there exists a copy $t_j'(\cdot)$ of $t_j(\cdot)$ and a standard Normal random vector $\mathbf{N}_{K_j} \sim \mathsf{N}(\mathbf{0}, \mathbf{I}_{K_j})$ with $K_j = \dim(\mathbf{\Pi}_j(\mathbf{x}))$ such that

$$\sup_{\mathbf{x}\in\mathcal{X}} \left|t_j'(\mathbf{x}) - Z_j(\mathbf{x})\right| = o_{\mathbb{P}}(r_n^{-1}), \qquad Z_j(\mathbf{x}) = \frac{\boldsymbol{\gamma}_{\mathbf{q},j}(\mathbf{x})'\mathbf{\Sigma}_j^{1/2}}{\sqrt{\Omega_j(\mathbf{x})}}\mathbf{N}_{K_j}.$$

This approximation is denoted by $t_j(\cdot) =_d Z_j(\cdot) + o_{\mathbb{P}}(r_n^{-1})$ in $\mathcal{L}^\infty(\mathcal{X})$.

This definition gives the precise meaning of uniform distributional approximation of $t_j(\cdot)$ by a Gaussian process $Z_j(\cdot)$, and also provides the explicit characterization of such Gaussian process. We establish this strong approximation in two distinct ways. For $d = 1$, we develop a novel two-step coupling approach based on the classical Komlós-Major-Tusnády (KMT) construction (Komlós, Major, and Tusnády, 1975, 1976). For $d > 1$, however, our two-step coupling approach does not generalize easily, and instead we apply an improved version of the classical Yurinskii construction (Yurinskii, 1978). See Zaitsev (2013) for a recent review and background references on strong approximation methods.

Unidimensional Regressor

Let $d = 1$. The following theorem gives a valid distributional approximation for the process $\{\widehat{T}_j(x) : x \in \mathcal{X}\}$ using the Gaussian process $\{Z_j(x) : x \in \mathcal{X}\}$, for $j = 0, 1, 2, 3$, in the sense of Definition II.1.

Theorem II.5 (Strong Approximation: KMT)**.** *Let the assumptions and conditions of Lemma II.2 hold with $d = 1$. If $j = 2, 3$, also assume $\frac{(\log n)^{3/2}}{\sqrt{nh}} = o(r_n^{-2})$. Then, for each $j = 0, 1, 2, 3$, $t_j(\cdot) =_d Z_j(\cdot) + o_{\mathbb{P}}(r_n^{-1})$ in $\mathcal{L}^\infty(\mathcal{X})$, where $Z_j(\cdot)$ is given in Definition II.1.*

The proof of this result employs a two-step coupling approach:

Step 1. On a sufficiently rich probability space, there exists a copy $t_j'(\cdot)$ of $t_j(\cdot)$, and an i.i.d. sequence $\{\zeta_i : 1 \leq i \leq n\}$ of standard Normal random variables, such that

$$\sup_{x \in \mathcal{X}} \left| t_j'(x) - z_j(x) \right| = o_{\mathbb{P}}(r_n^{-1}), \quad z_j(x) = \frac{\boldsymbol{\gamma}_{\mathbf{q},j}(x)'}{\sqrt{\Omega_j(x)}} \mathbb{G}_n[\boldsymbol{\Pi}_j(x_i)\sigma(x_i)\zeta_i].$$

Step 2. On a sufficiently rich probability space, there exists a copy $z_j'(\cdot)$ of $z_j(\cdot)$, and the standard Normal random vector $\mathbf{N}_{K_j}$ from Definition II.1 such that $z_j'(\cdot) =_d \bar{Z}_j(\cdot)$ conditional on $\mathbf{X}$, where

$$\bar{Z}_j(x) = \frac{\boldsymbol{\gamma}_{\mathbf{q},j}(x)' \bar{\boldsymbol{\Sigma}}_j^{1/2}}{\sqrt{\Omega_j(x)}} \mathbf{N}_{K_j}, \quad \bar{\boldsymbol{\Sigma}}_j := \mathbb{E}_n[\boldsymbol{\Pi}_j(x_i)\boldsymbol{\Pi}_j(x_i)'\sigma^2(x_i)],$$

and

$$\sup_{x \in \mathcal{X}} \left| \bar{Z}_j(x) - Z_j(x) \right| = o_{\mathbb{P}}(r_n^{-1}).$$

These two steps summarize our strategy for constructing the unconditionally Gaussian process $\{Z_j(x), x \in \mathcal{X}\}$ approximating the distribution of the whole t-statistic processes $\{t_j(x) : x \in \mathcal{X}\}$: we first couple $t_j(\cdot)$ to the process $z_j(\cdot)$, which is Gaussian only conditionally on $\mathbf{X}$ but not unconditionally (Step 1), and we then show that the unconditionally Gaussian process $Z_j(\cdot)$ approximates the distribution of $z_j(\cdot)$ (Step 2).

To complete the first coupling step, we employ a version of the classical KMT inequalities that applies to independent but non-identically distributed random variables (Sakhanenko, 1985, 1991). We do this because the processes $\{t_j(x) : x \in \mathcal{X}\}$ are characterized by a sum of independent but not identically distributed random variables conditional on $\mathbf{X}$. This part of our proof is inspired by, but is distinct from, the one given in Eggermont and LaRiccia (2009, Chapter 22), where a conditional strong approximation for smoothing splines is established.

The intermediate coupling result in Step 1 has the obvious drawback that the process $\{z_j(x) : x \in \mathcal{X}\}$ is Gaussian only conditionally on $\mathbf{X}$ but not unconditionally. Step 2 addresses this shortcoming by establishing an unconditional coupling, that is, approximating the distribution of the stochastic process $z_j(\cdot)$ by that of the (unconditional) Gaussian process $Z_j(\cdot)$. As shown in Appendix A, verifying the second coupling step boils down to controlling the supremum of a Gaussian random vector of increasing dimension, and in particular the crux is to prove precise (rate) control on $\|\bar{\mathbf{\Sigma}}_j^{1/2} - \mathbf{\Sigma}_j^{1/2}\|$, $j = 0, 1, 2, 3$. Both $\bar{\mathbf{\Sigma}}_j$ and $\mathbf{\Sigma}_j$ are symmetric and positive *semi*-definite. Further, for $j = 0, 1$, $\lambda_{\min}(\mathbf{\Sigma}_j) \gtrsim h^d$ for generic partitioning-based estimators under our assumptions, and therefore we use the bound

$$\|\mathbf{A}_1^{1/2} - \mathbf{A}_2^{1/2}\| \leq \lambda_{\min}(\mathbf{A}_2)^{-1/2} \|\mathbf{A}_1 - \mathbf{A}_2\|, \tag{2.13}$$

which holds for symmetric positive semi-definite $\mathbf{A}_1$ and symmetric positive definite $\mathbf{A}_2$ (Bhatia, 2013, Theorem X.3.8). Using this bound we obtain unconditional coupling from conditional coupling without additional rate restrictions.

However, for $j = 2, 3$ the bound (2.13) cannot be used in general because $\mathbf{p}$ and $\tilde{\mathbf{p}}$ are typically not linearly independent, and hence $\mathbf{\Sigma}_j$ will be singular. To circumvent this problem, we employ the weaker bound (Bhatia, 2013, Theorem X.1.1): if $\mathbf{A}_1$ and $\mathbf{A}_2$ are symmetric positive semi-definite matrices, then

$$\|\mathbf{A}_1^{1/2} - \mathbf{A}_2^{1/2}\| \leq \|\mathbf{A}_1 - \mathbf{A}_2\|^{1/2}. \tag{2.14}$$

This bound can be used for any partitioning-based estimator, with or without bias correction, at the cost of slowing the approximation error rate r_n when constructing the unconditional coupling, and hence leading to the stronger side rate condition as shown in the Theorem II.5 below. When $r_n = 1$, there is no rate penalty, while the penalty is only in terms of $\log n$ terms when $r_n = \sqrt{\log n}$ (as in Theorem II.8 further below). Furthermore, for certain partitioning-based series estimators it is still possible to use (2.13) even when $j = 2, 3$, as the following remark discusses.

Remark II.1 (Square-root Convergence and Improved Rates). The additional restriction imposed in Theorem II.5 for $j = 2, 3$, that $(\log n)^{3/2}/\sqrt{nh} = o(r_n^{-2})$, can be dropped in some special cases. For some bases it is possible to find a transformation matrix $\boldsymbol{\Upsilon}$, with $\|\boldsymbol{\Upsilon}\|_\infty \lesssim 1$, and a basis $\check{\mathbf{p}}$, which obeys Assumption II.3, such that $(\mathbf{p}(\cdot)', \tilde{\mathbf{p}}(\cdot)')' = \boldsymbol{\Upsilon}\check{\mathbf{p}}(\cdot)$. In other words, the two bases $\mathbf{p}$ and $\tilde{\mathbf{p}}$ can be expressed in terms of another basis $\check{\mathbf{p}}$ without linear dependence. Then, a positive lower bound holds for $\lambda_{\min}(\boldsymbol{\Sigma}_j), j = 2, 3$, implying that the bound (2.13) can be used instead of (2.14). For example, for piecewise polynomials and B-splines with equal knot placements for $\mathbf{p}$ and $\tilde{\mathbf{p}}$, a natural choice of $\check{\mathbf{p}}$ is simply a higher-order polynomial basis on the same partition. Since each function in $\mathbf{p}$ and $\check{\mathbf{p}}$ is a polynomial on each $\delta \in \Delta$ and nonzero on a fixed number of cells, the "local representation" condition $\|\boldsymbol{\Upsilon}\|_\infty \lesssim 1$ automatically holds.

The strong approximation results in Theorem II.5 for partitioning-based least squares estimation appear to be new in the literature. An alternative unconditional strong approximation for general series estimators is obtained by Belloni, Chernozhukov, Chetverikov, and Kato (2015) for the case of undersmoothing inference ($j = 0$). Their proof employs the classical Yurinskii's coupling inequality that controls the convergence rate of partial sums in terms of Euclidean norm, leading to the rate restriction $r_n^6 K^5/n \to 0$, up to $\log n$ terms, which does not depend on $\nu > 0$. In contrast, Theorem II.5 employs a (conditional) KMT-type coupling and then a second (unconditional) coupling approximation, and make use of the banded structure of the Gram matrix formed by local bases, to obtain weaker restrictions. Under bounded polynomial moments, we require only $r_n^6 K^3/n^{3\nu/(2+\nu)} \to 0$, up to $\log n$ terms. For example, when $\nu = 2$ and $r_n = \sqrt{\log n}$ this translates to $K^2/n \to 0$, up to $\log n$ terms, which is weaker than previous results in the literature. Under the sub-exponential conditional moment restriction, the restriction can be relaxed all the way to $K/n \to 0$, up to $\log n$ terms, which appears to be a minimal condition. This is for the entire t-statistic process. In addition, Theorem II.5 gives novel strong approximation results for robust bias-corrected t-statistic processes.

Remark II.2 (Strong Approximation: KMT for Haar Basis). Our two-step coupling approach builds on the new coupling Lemma A.3, which appears to be hard to extend to $d > 1$, except for the important special case the undersmoothed ($j = 0$) t-statistic process $\{\widehat{T}_0(x) : x \in \mathcal{X}\}$ constructed using Haar basis, which is a spline, wavelet and piecewise polynomial with $m = 1$. In this case, we establish $t_0(\cdot) =_d Z_0(\cdot) + o_{\mathbb{P}}(r_n^{-1})$ in $\mathcal{L}^\infty(\mathcal{X})$ for any $d \geq 1$ under the same conditions of Lemma II.2.

Multidimensional Regressors

Let $d \geq 1$. The method of proof employed to establish Theorem II.5 does not extend easily to multivariate regressors ($d > 1$) in general. Therefore, we present an alternative strong approximation result based on an improved version of the classical Yurinskii's coupling inequality, recently developed by Belloni, Chernozhukov, Chetverikov, and Fernandez-Val (2018).

Theorem II.6 (Strong Approximation: Yurinskii)**.** *Let the assumptions and conditions of Lemma II.2 hold. Furthermore, assume $\nu \geq 1$ and $\frac{(\log n)^4}{nh^{3d}} = o(r_n^{-6})$. Then, for each $j = 0, 1, 2, 3$, $t_j(\cdot) =_d Z_j(\cdot) + o_{\mathbb{P}}(r_n^{-1})$ in $\mathcal{L}^\infty(\mathcal{X})$, where $Z_j(\cdot)$ is given in Definition II.1.*

This strong approximation result does not have optimal (i.e. minimal) restrictions, but nonetheless improves on previous results by exploiting the specific structure of the partitioning-based estimators, while also allowing for any $d \geq 1$. Specifically, the result sets $\nu = 1$ and requires $r_n^6 K^3/n \to 0$, up to $\log n$ terms, regardless of the moment restriction. While not optimal when $\nu > 3$ (see Remark II.2 for a counterexample), the result still improves on the condition $r_n^6 K^5/n \to 0$, up to $\log n$ terms, mentioned previously. In addition, Theorem II.6 gives novel strong approximation results for robust bias-corrected t-statistic processes and any $d \geq 1$.

2.6.2 Implementation

We present a simple plug-in approach that gives a (feasible) approximation to the infeasible standardized Gaussian processes $\{Z_j(\mathbf{x}) : \mathbf{x} \in \mathcal{X}\}$, in order to conduct inference using the results in Theorem II.5 or Theorem II.6. The following definition gives a precise description of how the approximation works.

Definition II.2 (Simulation-Based Strong Approximation)**.** Let $\mathbb{P}^*[\cdot] = \mathbb{P}[\cdot|\mathbf{y}, \mathbf{X}]$ denote the probability operator conditional on the data. For each $j = 0, 1, 2, 3$, the law of the Gaussian process $\{Z_j(\mathbf{x}) : \mathbf{x} \in \mathcal{X}\}$ is approximated by a (feasible) Gaussian process $\{\widehat{Z}_j(\mathbf{x}) : \mathbf{x} \in \mathcal{X}\}$, with known distribution conditional on the data $(\mathbf{y}, \mathbf{X})$, in $\mathcal{L}^\infty(\mathcal{X})$, if the following condition holds: on a sufficiently rich probability space there exists a copy $\widehat{Z}_j'(\cdot)$ of $\widehat{Z}_j(\cdot)$ such that $\widehat{Z}_j'(\cdot) =_d Z_j(\cdot)$ conditional on the data, and

$$\mathbb{P}^*\left[\sup_{\mathbf{x}\in\mathcal{X}} |\widehat{Z}_j'(\mathbf{x}) - Z_j(\mathbf{x})| \geq \eta r_n^{-1}\right] = o_{\mathbb{P}}(1), \qquad \forall \eta > 0,$$

where, for a $\mathbf{N}_{K_j} \sim \mathsf{N}(\mathbf{0}, \mathbf{I}_{K_j})$ with $K_j = \dim(\mathbf{\Pi}_j(\mathbf{x}))$,

$$\widehat{Z}_j(\mathbf{x}) = \frac{\widehat{\boldsymbol{\gamma}}_{\mathbf{q},j}(\mathbf{x})'\widehat{\mathbf{\Sigma}}_j^{1/2}}{\sqrt{\widehat{\Omega}_j(\mathbf{x})}}\mathbf{N}_{K_j}, \qquad \mathbf{x} \in \mathcal{X}, \qquad j = 0, 1, 2, 3.$$

This approximation is denoted by $\widehat{Z}_j(\cdot) =_{d^*} Z_j(\cdot) + o_{\mathbb{P}^*}(r_n^{-1})$ in $\mathcal{L}^\infty(\mathcal{X})$.

From a practical perspective, Definition II.2 implies that sampling from $\widehat{Z}_j(\cdot)$, conditional on the data, is possible and provides a valid distributional approximation of $Z_j(\cdot)$, for each $j = 0, 1, 2, 3$. The feasible process $\widehat{Z}_j(\mathbf{x})$ given in this definition relies on a direct plug-in approach, where all the unknown quantities in $Z_j(\cdot)$ are replaced by consistent estimators; that is, using the estimators already used in the feasible t-statistics. Resampling is done conditional on the data from a multivariate standard Gaussian of dimension K_j, not n.

Theorem II.7 (Plug-in Approximation)**.** *Let the assumptions and conditions of Lemma II.2 hold. Furthermore, for $j = 2, 3$:*

(i) *when $\sup_{\mathbf{x}\in\mathcal{X}} \mathbb{E}[|\varepsilon_i|^{2+\nu}|\mathbf{x}_i = \mathbf{x}] < \infty$, assume $\frac{n^{\frac{1}{2+\nu}}(\log n)^{\frac{4+3\nu}{4+2\nu}}}{\sqrt{nh^d}} = o(r_n^{-2})$; or*

(ii) *when $\sup_{\mathbf{x}\in\mathcal{X}} \mathbb{E}[|\varepsilon_i|^3 \exp(|\varepsilon_i|)|\mathbf{x}_i = \mathbf{x}] < \infty$, assume $\frac{(\log n)^{5/2}}{\sqrt{nh^d}} = o(r_n^{-2})$.*

Then, for each $j = 0, 1, 2, 3$, $\widehat{Z}_j(\cdot) =_{d^*} Z_j(\cdot) + o_{\mathbb{P}^*}(r_n^{-1})$ in $\mathcal{L}^\infty(\mathcal{X})$, where $\widehat{Z}_j(\cdot)$ is given in Definition II.2.

This result strengthens the rate condition for $j = 2, 3$ compared to Theorems II.5 $(d = 1)$ and II.6 $(d \geq 1)$ only by logarithmic factors when $r_n = \sqrt{\log n}$. Moreover, if the structure discussed in Remark II.1 holds, then this condition can be dropped.

2.6.3 Application: Confidence Bands

A natural application of Theorems II.5, II.6 and II.7 is to construct confidence bands for the regression function or its derivatives. Specifically, for $j = 0, 1, 2, 3$ and $\alpha \in (0, 1)$, we seek a quantile $q_j(\alpha)$ such that

$$\mathbb{P}\left[\sup_{\mathbf{x}\in\mathcal{X}} |\widehat{T}_j(\mathbf{x})| \leq q_j(\alpha)\right] = 1 - \alpha + o(1),$$

which then can be used to construct uniform $100(1-\alpha)$-percent confidence bands for $\partial^{\mathfrak{q}}\theta(\mathbf{x})$ of the form

$$\left[\widehat{\partial^{\mathfrak{q}}\theta}_j(\mathbf{x}) \pm q_j(\alpha)\sqrt{\widehat{\Omega}_j(\mathbf{x})/n} \;:\; \mathbf{x}\in\mathcal{X}\right].$$

The following theorem establishes a valid distributional approximation for the suprema of the t-statistic processes $\{\widehat{T}_j(\mathbf{x}) : \mathbf{x}\in\mathcal{X}\}$ using Chernozhukov, Chetverikov, and Kato (2014b, Lemma 2.4) to convert our strong approximation results into convergence of distribution functions in terms of Kolmogorov distance.

Theorem II.8 (Confidence Bands)**.** *Let the conditions of Theorem II.5 or Theorem II.6 hold with $r_n = \sqrt{\log n}$. If the corresponding conditions of Theorem II.7 hold for each $j = 0, 1, 2, 3$, then*

$$\sup_{u\in\mathbb{R}}\left|\mathbb{P}\left[\sup_{\mathbf{x}\in\mathcal{X}}|\widehat{T}_j(\mathbf{x})|\leq u\right] - \mathbb{P}^*\left[\sup_{\mathbf{x}\in\mathcal{X}}|\widehat{Z}_j(\mathbf{x})|\leq u\right]\right| = o_{\mathbb{P}}(1).$$

Chernozhukov, Chetverikov, and Kato (2014a,b) recently showed that if one is only interested in the supremum of an empirical process rather than the *whole* process, then the sufficient conditions for distributional approximation could be weakened compared to earlier literature. Their result applied Stein's method for Normal approximation to show that suprema of general empirical processes can be approximated by a sequence of suprema of Gaussian processes, under the usual undersmoothing conditions (i.e., $j = 0$). They illustrate their general results by considering t-statistic processes for both kernel-based and series-based nonparametric regression: Chernozhukov, Chetverikov, and Kato (2014b, Remark 3.5) establishes a result analogous to Theorem II.8 under the side rate condition $K/n^{1-2/(2+\nu)} = o(1)$, up to $\log n$ terms (with $q = 2+\nu$ in their notation). In comparison, our result for $j = 0$ and $d = 1$ in Theorem II.8, under the same moment conditions, requires exactly the same side condition, up to $\log n$ terms. However, comparing Theorems II.5 and II.8 shows that the *whole* t-statistic process for partitioning-based series estimators, and not just the suprema thereof, can be approximated under the same weak conditions when $d = 1$. The same result holds for sub-exponential moments, where the rate condition becomes minimal: $K/n = o(1)$, up to $\log n$ factors. We are able to achieve such sharp rate restrictions and approximation rates only via the new two-step coupling approach mentioned above (see Lemma A.3), and by exploiting the specific features of the estimator together with the help of the key anti-concentration idea introduced by Chernozhukov, Chetverikov, and Kato (2014b). In addition, Theorem II.8 gives new inference results for bias-corrected estimators ($j = 1, 2, 3$).

Finally, the strong approximation result for the entire t-statistic processes given in Theorems II.5 and II.6, and related technical results given in Appendix A, can also be used to construct other types of confidence bands for the regression function and its derivatives; e.g., Genovese and Wasserman (2005, 2008). We do not elaborate further on this to conserve space.

2.7 Conclusion

We presented new asymptotic results for partitioning-based least squares regression estimators. The first main contribution gave a general IMSE expansion for the point estimators. The second set of contributions were pointwise and uniform distributional approximations, with and without robust bias correction, for t-statistic processes indexed by $\mathbf{x} \in \mathcal{X}$, with improvements in rate restrictions and convergence rates. For the case of $d = 1$, our uniform approximation results rely on a new coupling approach, which delivered seemingly minimal rate restrictions.

Chapter III

Partial Linear Models: Binscatter

3.1 Introduction

The previous chapter discusses partitioning-based estimators in a general nonparametric regression setup. Such methods perform well when there are only a few covariates. However, as the number of covariates increases, the dimensionality of an approximation basis may grow rapidly, leading to poor finite sample performance of completely nonparametric estimators. In such scenarios researchers often resort to semiparametric methods for dimension reduction at the cost of stronger assumptions on model specifications.

A commonly used semiparametric method in practice is partial linear regression. If we follow the notation in Chapter II, this amounts to assuming that

$$\theta(\mathbf{x}) = \mu(x_1) + \mathbf{w}'\boldsymbol{\gamma}, \quad \mathbf{x} = (x_1, \mathbf{w}')'.$$

Researchers may be interested in the (nonparametric) mean relationship between an outcome y and a scalar independent variable x_1, where $\mathbf{w}$ is a vector of control variables that enter the model linearly. With a little abuse of notation, we will simply denote x_1, the first variable in $\mathbf{x}$, by x throughout this chapter.

The partial linear regression with the nonparametric component $\mu(x)$ estimated using a piecewise constant basis is often referred to as *binscatter*, which is a flexible, yet parsimonious way of visualizing and summarizing large data sets (Chetty and Szeidl, 2006; Chetty, Looney, and Kroft, 2009; Chetty, Friedman, Olsen, and Pistaferri, 2011; Chetty, Friedman, Hilger, Saez, Schanzenbach, and Yagan, 2011). This methodology is also often used for informal (visual) evaluation of substantive hypotheses about shape features of the unknown regression function such as linearity, monotonicity, or concavity. Binscatter has gained immense popularity among empirical researchers and policy makers, and is by now a leading component of the standard applied microeconomics toolkit. However, the remarkable proliferation of binscatter in empirical work has not been accompanied by the development of econometric results guiding its correct use and providing valid statistical procedures. Current practice employing binscatter is usually ad-hoc and undisciplined, which not only hampers replicability across studies, but also has the potential of leading to incorrect empirical conclusions.

This chapter presents the first foundational study of binscatter. The general large sample properties of partitioning-based estimators derived in Chapter II are applied to binscatter, with new theoretical and practical issues resolved. Then we provide several results which aid both in understanding the validity (or lack thereof) of current practices, and in offering principled guidance for future applications. To give a systematic analysis of binscatter, we first recast it as a particular nonparametric estimator of a regression function employing (possibly restricted) piecewise approximations in a semi-linear regression context. Thus, our first main contribution is to set up an econometrics framework to understand and analyze binscatter formally. This framework allows us to obtain an array of theoretical and practical results for canonical binscatter methods, and also to propose new complementary methods delivering more flexible and smooth approximations of the regression function, which still respect the core features of binscatter. The latter methods are particularly well suited for enhanced graphical presentation of estimated regression functions and confidence bands, and for formal testing of substantive hypotheses about the unknown regression function.

Furthermore, using our econometric framework, we highlight important methodological and theoretical problems with the covariate adjustment methods as commonly employed in practice, and propose a new alternative approach that is more generally valid and principled. To be more specific, we discuss the detrimental effects of the widespread practice of first "residualizing" additional covariates and then constructing the binscatter, and show how our proposed alternative covariate-adjusted binscatter circumvents those problems.

The proposed econometric framework is then used to offer several new methodological results for binscatter applications. Specifically, our second main contribution is to develop a valid and optimal selector of the number of bins for binscatter implementation, which is constructed based on an integrated mean square error approximation. Our proposed selector intuitively balances the bias and variance of binscatter, and can contrast sharply with ad-hoc choices encountered in practice: always using 10 or 20 bins. The third main contribution of this paper is to provide valid confidence intervals, confidence bands, and hypothesis testing procedures of both parametric specifications and nonparametric shape restrictions of the unknown regression function. These results not only give principled guidance for current empirical practice, but also offer new methods encoding informal (visual) assessments commonly found in empirical papers using binscatter. The results in this chapter are obtained under random sampling, but our work could be extended to clustered or grouped data with some additional complications. We defer the discussion of this to future research.

The remainder of the paper proceeds as follows. Employing an empirical example throughout, Section 3.2 gives a gentle introduction to binscatter, overviews and illustrates numerically our main methodological results, and discusses related literature. This relatively long section is meant to be not only heuristic and empirically-driven, but also self-contained in terms of reviewing the methodology and contributions offered by our paper. On the other hand, the next three sections are more technical and precise: Section 3.3 introduces and formalizes binscatter, starting with its canonical form, then incorporating covariate-adjustment and within-bin higher-order polynomial fitting, and culminating with a smooth version based on imposing continuity restrictions at the boundaries of the bins; Section 3.4 gives formal results for empirical selection of the number of bins used to implement binscatter; and Section 3.5 presents our main theoretical results for estimation, inference, and graphical presentation, including shape-related testing procedures of substantive interest. Finally, Section 3.6 concludes. Appendix B collects all proofs.

3.2 Overview of Results

In this section we make clear what binned scatter plots are, how they are often used, and how our results can aid empirical practice. The treatment here is informal, but complete, drawing on the formal results presented in the upcoming (more technical) sections. Before detailing our new tools, it is important to define what a binned scatter plot is, and what it is not. See Chetty and Szeidl (2006, Figure 1) for one the very first explicit appearances of a binned scatter plot in the applied microeconomics literature, and see Chetty, Looney, and Kroft (2009), Chetty, Friedman, Olsen, and Pistaferri (2011), and Chetty, Friedman, Hilger, Saez, Schanzenbach, and Yagan (2011) for other early papers using binscatter methods. In addition, see Stepner (2014) for a widely used software implementation of canonical binscatter methods, and see Starr and Goldfarb (2018) for a very recent heuristic overview of canonical binscatter.

We illustrate our methods using data from the American Community Survey. In order to have full control on different features of the statistical model, this section employs a simulated dataset based on a data generating process constructed using the real survey dataset. But, in Chapter IV we return to the original survey dataset to illustrate our main recommendations for practice using the actual data. Appendix B details how the simulated data was constructed.

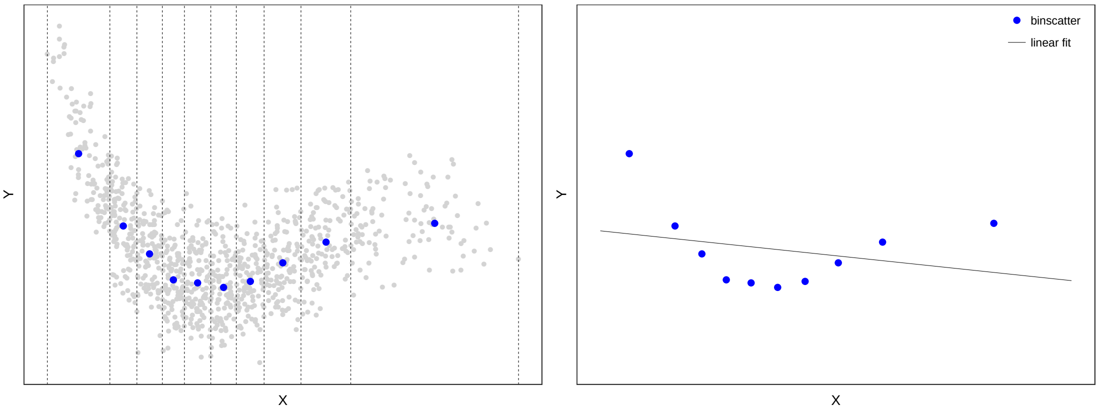
The *scatter plot* itself is a classical tool for data visualization. It shows all the data on two variables y and x, and allows a researcher to visualize not only the relationship between y and x but also the variability, areas of high or low mass, and all other features of their joint distribution. However, in the era of large sample sizes, scatter plots are less useful. A scatter plot of a typical administrative data set with millions of observations yields a solid black cloud, and in particular, obscures the underlying relationship between y and x. This is where binning enters.

To construct a binned scatter plot one first divides the support of x into some number of bins, denoted herein by J. In most cases, ad-hoc choices such as $J = 10$ or 20 are most predominant in practice, with bins themselves divided at the empirical quantiles of observations of x (ignoring y), an approach we call quantile-spaced binning (or partitioning) herein. Then a single dot is placed in the plot representing the mean of y for the observations falling in each bin. The final plot consists of only these J dots, usually depicted at the center of each quantile-spaced bin. Often added is the regression line from a OLS fit to the underlying data. See Figure III.1. It is typical in applications to “control” for covariates in both the regression line and the plot itself, which as discussed below, requires additional care.

The question is: what aspect of the data is being visualized with a binned scatter plot? This turns out to be not the data itself, but only the conditional expectation of y given x; the regression function. A binned scatter plot is nothing more than the fitted values of a particular nonparametric regression of y on x. This is not a disadvantage, indeed, we view it as the reverse: starting from this insight we can deliver a host of tools and results, both formal and visual, for binned scatter plots.

But it is nonetheless important to point out the limitations of what can be learned from a binned scatter plot. The plot *is not* a visualization of the whole data set in any meaningful way. That is, it is not at all analogous to a traditional scatter plot. Many different data sets can give rise to identical binned scatter plots, as in Figure III.2. In particular, the variance (of y given x) is entirely suppressed. Figure III.2 shows four different data sets, with different amounts of variance and heteroskedasticity, which nonetheless yield identical plots. This is not a new revelation, but it does seem to be the source of some confusion in practice. Indeed, Chetty, Friedman, and Rockoff (2014, p. 2650) note “that this binned scatter plot provides a nonparametric representation of the conditional expectation function but does not show the underlying variance

Figure III.1 The Basic Construction of a Binned Scatter Plot.



(a) Scatter and Binscatter Plots **(b)** Binscatter and Linear Fit

Notes. The data is divided into $J = 10$ bins according to the observed x. Within each bin a single dot is plotted at the mean of y for observations falling in the bin. The final plot (b) consists of only these J dots, and the fit from a least squares linear regression of y on x. Constructed using simulated data described in Appendix B.

in the individual-level data." To show the underlying variance from a large data set, one can plot a small random sample of the data. For a large data set, this is perhaps most akin to a traditional scatter plot. However, the sample may need to be small enough as to render the conclusions unreliable, and further, given our results, this is not necessary in most cases.

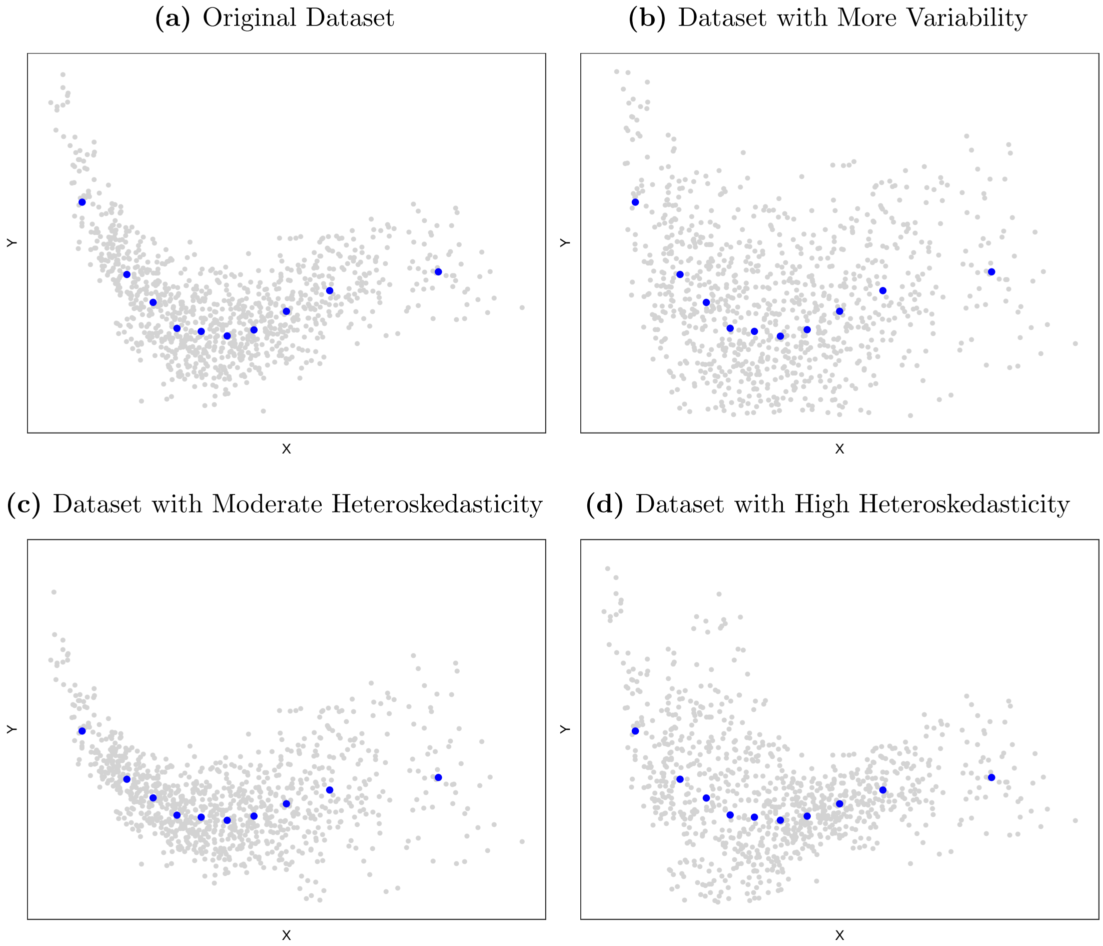
Turning back to how binscatter plots are constructed, in this paper we analyze these plots from an econometric point of view. This allows us not only to formalize its properties, but also develop new tools for empirical practice. These include a valid and optimal choice for the number of bins, valid confidence intervals and bands reflecting the true uncertainty in the data, and formal (and visual) assessment of substantive hypotheses of interest, such as whether the relationship between y and x is monotonic, or of a particular parametric form such as linear, or different between two groups. Here we give only an introduction to these ideas; formal details are spelled out below.

The canonical nonparametric regression model is

$$y_i = \vartheta(x_i) + \varepsilon_i, \qquad \mathbb{E}[\varepsilon_i|x_i] = 0, \tag{3.1}$$

where (y_i, x_i), $i = 1, 2, \ldots, n$, is random sample from (y, x). Notice that we introduce another function $\vartheta(\cdot)$, and reserve the notation $\mu(\cdot)$ for the partial mean relation between y and x with additional variables $\mathbf{w}$ controlled for. Here we are interested in the function $\vartheta(x)$ and its properties. For example, we might like to know if it

Figure III.2 Scatter and Binscatter Plots with Different Variability.



(a) Original Dataset

(b) Dataset with More Variability

(c) Dataset with Moderate Heteroskedasticity

(d) Dataset with High Heteroskedasticity

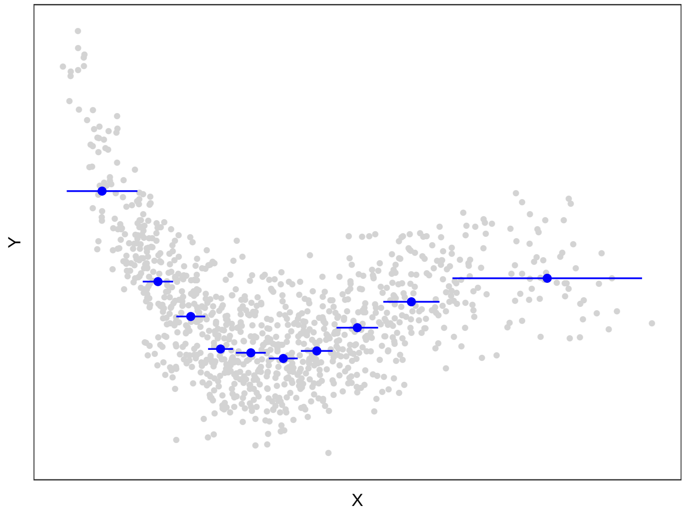
Notes. Four simulated different data sets, each with different variance of y given x, but identical binned scatter plots. Constructed using simulated data described in Appendix B.

is (well-approximated by) a linear function, i.e. $\vartheta(x) = \theta_0 + x\theta_1$, or quadratic, i.e. $\vartheta(x) = \theta_0 + x\theta_1 + x^2\theta_2$. This is implicitly behind plots of the form of Figure III.1: we want to assume that the linear approximation is sound enough that conclusions from an OLS regression of y on x are useful for policy.

Binned scatter plots estimate the unknown regression function $\vartheta(x)$ by exploiting the fact that $\vartheta(x_1) \approx \vartheta(x_2)$ if $x_1 \approx x_2$. Broadly speaking, all nonparametric regression exploits this same idea. For binscatter regressions, "$x_1 \approx x_2$" is translated as being in the same bin, and then further, the estimator sets $\hat{\vartheta}(x) = \bar{y}_j$ for all x in the j-th bin, $j = 1, 2, \ldots, J$, where $\bar{y}_j$ denotes the sample average of the y_i's with x_i's in that j-th bin. This results in a piecewise constant estimate, as shown in Figure III.3. A typical binned scatter plot shows only one point within each bin, but it is important to observe that a binned scatter plot is *equivalent* to this piecewise constant fit,

Figure III.3 The actual nonparametric estimator corresponding to a binned scatter plot.

(a) Binned Scatter Plot with Piecewise Constant Fit



Notes. Constructed using simulated data described in Appendix B.

however unfamiliar it may look. As a way of contrast, a traditional kernel regression is distinct almost everywhere from the canonical binscatter, coinciding for a very special implementation and then only at J points: at the center of each bin and employing the uniform kernel with bandwidth equal to half the block length both procedures will yield the same fitted values, but only at these J points. To "fill in" the rest of the regression curve, traditional kernel regression rolls out the window, implying new bandwidths and associated new "bins", distinct almost everywhere from those used to form canonical binscatter.

Despite its appearance, piecewise constant fits over pre-determined quantile-spaced bins is not a "bad" nonparametric estimation method in any sense, when implemented properly it shares many favorable theoretical properties with more familiar methods such as traditional kernel smoothing and, in fact, they are the building block for popular spline approximations. Applying binning to regression problems dates back at least to the regressogram of Tukey (1961b), and in nonparametric regression more broadly it is known as partitioning regression (Györfi, Kohler, Krzyżak, and Walk, 2002; Cattaneo and Farrell, 2013; Cattaneo, Farrell, and Feng, 2018a). The use of binned scatter plots in applied economics is most closely related to this strand of literature, and our theory below can be thought of as a generalization and complete formal treatment of the regressogram. Binning has been applied in many other areas due to its intuitive appeal and ease of implementation: in density estimation as the classical histogram; in program evaluation for subclassification (Cochran, 1968; Cattaneo and Farrell, 2011b), and for visualization in regression discontinuity designs (Calonico, Cattaneo, and Titiunik, 2015) and bunching designs (Kleven, 2016); in empirical finance it is related to portfolio sorting (Fama, 1976; Cattaneo, Crump, Farrell, and Schaumburg, 2019a); and in machine learning it is at the heart of regression trees and similar methods (Friedman, 1977; Hastie, Tibshirani, and Friedman, 2009). We do not address these other applied contexts directly here, as each is different enough to require a separate analysis, but our results and tools can be adapted and exported to those other settings.

This chapter offers three main methodological contributions to the understanding and correct use of binscatter methods in applied microeconomics, which we summarize and illustrate in the remaining of this overview section. In closing this section, we also mention briefly some other contributions related to software and theory.

Contribution 1: Framework and Construction

Understanding binscatter requires formalizing it in a principled way. Thus, our first contribution is to outline a framework that not only correctly incorporates additional covariates, and gives the baseline for further extensions to clustered data, but also permits us to introduce more flexible polynomial regression approximations within bins as well as to incorporate smoothness restrictions across bins. These extensions are particularly useful in applications because it is common for researchers both to control for additional factors in their regression specifications and to prefer more "smooth" global approximations and confidence bands, in combination with canonical binscatter.

In Section 3.3, we first recast canonical binscatter as a very basic nonparametric least squares regression procedure, and then extend it to incorporate additional covariate adjustments and several other features. Adjusting for additional covariates is standard in applications, and to formalize it we extend model (3.1) to include a vector of controls, $\mathbf{w}$, as follows:

$$y_i = \mu(x_i) + \mathbf{w}_i'\boldsymbol{\gamma} + \epsilon_i, \qquad \mathbb{E}[\epsilon_i|x_i, \mathbf{w}_i] = 0, \tag{3.2}$$

where $(y_i, x_i, \mathbf{w}_i')$, $i = 1, 2, \ldots, n$, is random sample from $(y, x, \mathbf{w})$. In this case the object of interest, both visually and numerically, is the function $\mu(x)$. This regression model, variously referred to as partially linear, semi-linear, or semiparametric, retains the interpretation familiar from linear models of "the effect of x on y, controlling for

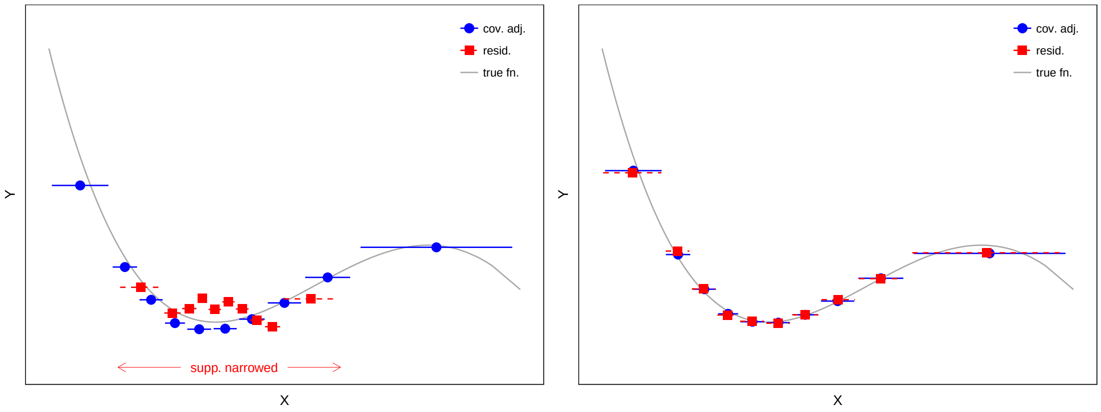
$\mathbf{w}$".

The regression model (3.2) justifies a particular way of covariate adjustment, which is not the way encountered in practice: see Section 3.3.3 for a detail comparison and discussion. In particular, if $\mu(x)$ is not linear, then standard Frisch-Waugh logic does not apply: one cannot estimate (or binned scatter plot) the function $\mu(x)$ using the residuals from least squares regressions of y on $\mathbf{w}$ and x on $\mathbf{w}$. This highlights an important methodological issue with most current applications of covariate-adjusted binscatter, since it is often the case that practitioners first regress out the additional covariates $\mathbf{w}$ and only after construct the binscatter based on the resulting residuals. The latter approach, which differs from our proposed method for covariate-adjustment, can lead to very different empirical findings. In this paper, we refer to the latter approach as (canonical, covariate-adjusted) residualized binscatter.

We illustrate this issue of covariate adjustment numerically in Figure III.4. The true regression function, $\mu(x)$, is depicted in solid grey, while the two approaches to covariate-adjusted binscatter are presented in solid blue circles (ours) and solid red squares (residualized binscatter). Our recommended method implements binscatter via model (3.2), while residualized binscatter implements binscatter via model (3.1) after replacing y_i and x_i by the residuals obtained from regressing y_i on $\mathbf{w}_i$ and regressing x_i on $\mathbf{w}_i$, respectively. As Figure III.4 clearly indicates, the two approaches for covariate adjustment lead to quite different results if x and $\mathbf{w}$ are correlated. The reason is simple: our approach is valid for model (3.2), while residualized binscatter is invalid in general. Figure III.4(a) shows that residualized binscatter is unable to correctly approximate the true function of interest $\mu(x)$, while our semi-linear covariate-adjustment approach works well.

This numerical illustration relies on data generated as in model (3.2), but even when the true regression function of y_i given $(x_i, \mathbf{w}_i')$ does not satisfy the semi-linear structure, our approach to covariate adjustment retains a natural interpretation as a "best" semi-linear approximation in mean square, just as it occurs with simple least squares methods (e.g., Angrist and Pischke, 2008, for more discussion), while residualized binscatter would be fundamentally misspecified and uninterpretable in such case. To put this another way, in the case when the true $\mu(x)$ is nonlinear, the conclusions reached from the currently dominant binscatter approach are incompatible with the often-presented table of results from a regression of y_i on x_i and $\mathbf{w}_i$. While such dominant approach is not completely "wrong" in all cases, it does not match how the results are usually interpreted. See Section 3.3.3 for more technical details and discussion on our recommended approach to covariate adjustment vis-á-vis residualized

Figure III.4 Comparison of Covariate Adjustment Approaches.



(a) x correlated with $\mathbf{w}$

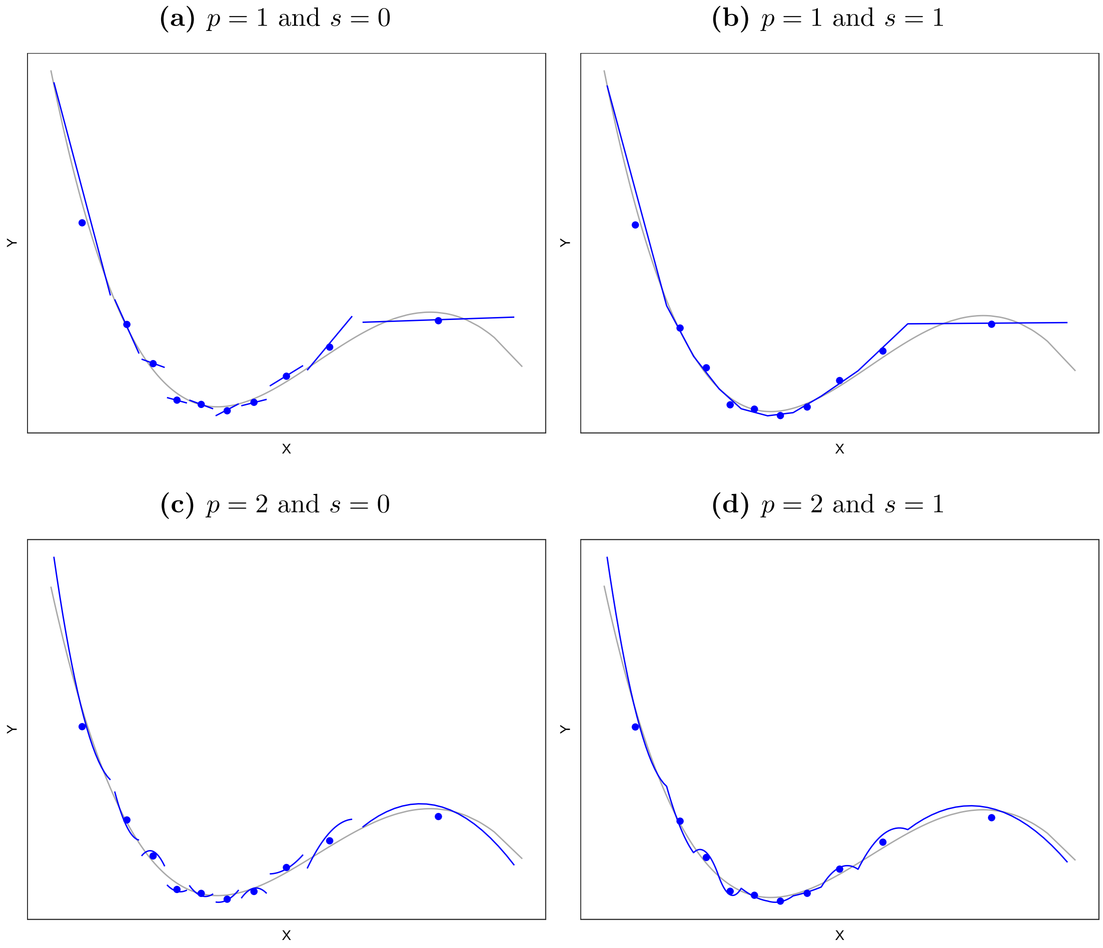
(b) x independent of $\mathbf{w}$

Notes. Two plots comparing semi-linear covariate-adjustment and residualized covariate adjustment for binscatter. Plot (a) illustrates the biases introduced by residualization when there is non-zero correlation between x and the other covariates $\mathbf{w}$ controlled for. Plot (b) shows that the residualization biases are not present when x and $\mathbf{w}$ are independent, and the location of binscatter is adjusted: see Section 3.3.3 for more details. Constructed using simulated data described in Appendix B.

binscatter.

In addition to incorporating covariate adjustments in an appropriate and interpretable way, our proposed framework allows us to introduce new, related binscatter procedures. In particular, we consider two useful extensions for empirical work: fitting a p-th order polynomial regression within each bin and imposing s-th order smoothness restrictions across bins, both with and without covariate adjustments. These generalizations of binscatter are exhibited in Figure III.5. Increasing the polynomial order p used within bins allows for a more “flexible” local approximation within each bin, while increasing the smoothness order s forces the approximation to be smoother across bins. Thus, the user-choices p and s control flexibility and smoothness from a local and global perspectives, respectively. For example, if $p = 1$, then $s = 1$ corresponds to piecewise linear fits that are forced to be connected at the bin’s boundaries: a continuous but not differentiable global fit based on binscatter. This is illustrated in Figure III.5(b). Of course, removing the smoothness constraint ($s = 0$) leads to piecewise linear fits within bins ($p = 1$) that need not to agree at the bins’ boundaries: Figure III.5(a). An example of within-bin quadratic fit ($p = 2$) without smoothness constraints ($s = 0$) is given in Figure III.5(c), while imposing only continuity at the bins’ edges ($s = 1$) for the quadratic case is depicted in Figure III.5(d). A within-bin quadratic fit ($p = 2$) with continuously differentiable restrictions at bins’ boundaries ($s = 2$) is not depicted to conserve space, but follows the same logic.

Figure III.5 Binscatter Generalizations.



(a) $p = 1$ and $s = 0$

(b) $p = 1$ and $s = 1$

(c) $p = 2$ and $s = 0$

(d) $p = 2$ and $s = 1$

Notes. Constructed using simulated data described in Appendix B.

This generalization of binscatter can be implemented with or without covariate adjustment, as discussed in Section 3.3. It should be clear that canonical binscatter corresponds to the specific choice $p = s = 0$ (Figure III.3), but one can consider more or less smooth versions thereof by appropriate choice of $s \leq p$. Another advantage of considering $p > 0$ polynomial fits, with or without covariate adjustments and/or smoothness restrictions, is that approximating the derivatives $\mu^{(v)}(x) = \frac{d^v}{dx^v}\mu(x)$ is enabled: estimating derivatives of the regression function $\mu(x)$ is crucial for testing shape features such as monotonicity or concavity, as we discuss further below.

Employing our econometrics framework, we obtain an array of methodological results for canonical binscatter and its generalizations, which we summarize and illustrate next.

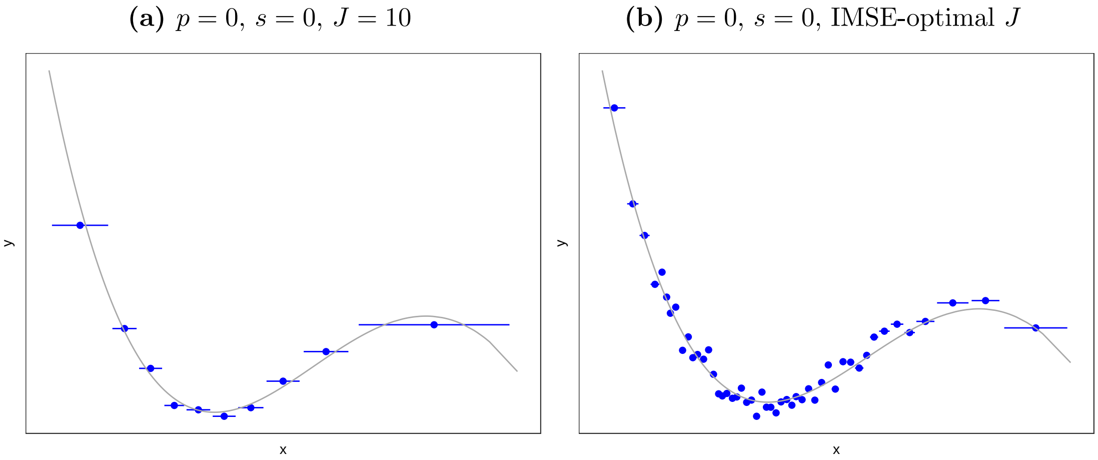
Contribution 2: Valid and Optimal Number of Bins Selection

Implementing our standard binned scatter plot requires one choice: the number of bins to be used, J. Given a choice of J, the position of the bins is set by the empirical quantiles of x, via the quantile-spaced binning used in all applications. Because the bin positions are determined by estimated quantiles, the random binning structure underlying binscatter introduces some additional technical issues. Nevertheless, in Section 3.4, we employ our formalization of binscatter to view the choice of J as that of a tuning parameter in nonparametric estimation, just as a bandwidth is the tuning parameter in kernel regression. As such, it reflects a bias/variance trade-off: as J increases the bias decreases but the variability of the estimator increases. This is depicted in Figure III.6.

Our second main contribution is to give a precise choice of the number of quantiles J that trades off bias and variance in a valid and optimal way. Specifically, we study the asymptotic properties of the integrated mean square error (IMSE) of binscatter and its generalizations, and show that an IMSE-optimal choice of J is always proportional to $n^{\frac{1}{2p+3}}$, up to rounding to the closest integer, where recall p denotes the order of polynomial fit within each bin. For example, if a constant fit is used (i.e., the canonical binscatter), as in Figure III.3, then the optimal choice of number of bins is $J \propto n^{1/3}$. The role of covariate adjustment, smoothness restrictions across bins, and other related features of binscatter, show up only through the constant of proportionality in the optimal rule for J. For implementation, we make the optimal choice of J, including its constant, fully data-driven and automatic, and readily available for empirical work in our companion software.

Most of the current binscatter applications employ an ad-hoc number of bins, usually $J = 10$ or $J = 20$. There is no a priori reason why these choices would be valid: these ad-hoc choices can be “too” small to account for a non-linear relationship $\mu(x)$ (i.e., too much misspecification bias), leading to incorrect empirical conclusions. Furthermore, even when “too” large, there is no a priori reason why they would be optimal in terms of the usual bias-variance trade-off. Depending on the underlying unknown features of the data, such an arbitrary choice of J could be “too small” or “too large”, leading to graphical and formal procedures with invalid or at least unreliable statistical properties. Section 3.4 presents our formal approach to this problem, where we rely on an objective measure (IMSE) to select in a data-driven way the number of bins J to use in each application of binscatter.

Figure III.6 Number of Bins (J).



Notes. The plots illustrate the potential effects on binscatter of choosing the number of bins J too small vis-á-vis in an IMSE-optimal way. Constructed using simulated data described in Appendix B.

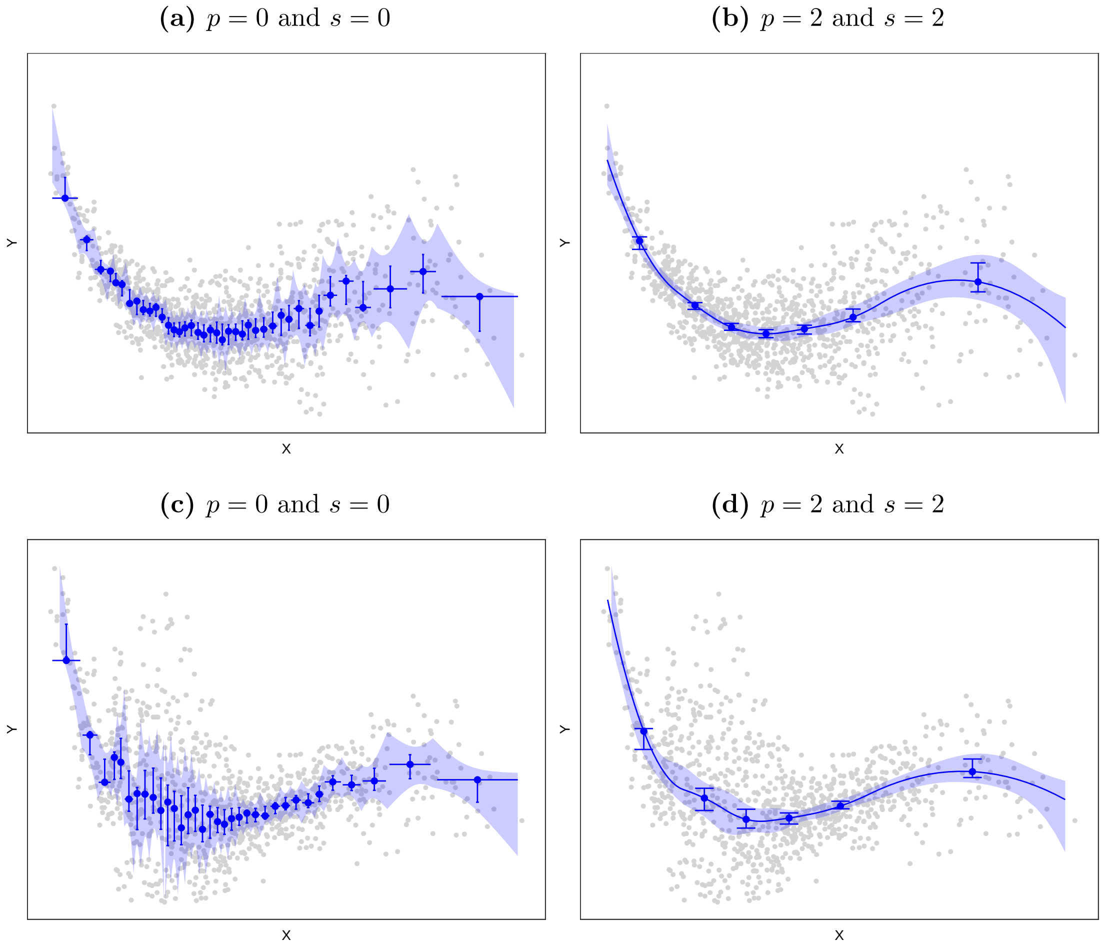
Contribution 3: Confidence Bands and Valid Inference

Armed with an IMSE-optimal estimator of the regression function we now turn to inference. Binned scatter plots are often used in applications to guide subsequent regression analyses, essentially as a visual specification check. A second common usage is to visually assess economically meaningful properties such as monotonicity or concavity. Our results allow for a valid assessment, both visually and formally, of these questions, as well as faithful display of the variability in the outcome y in the underlying data set. None of these are possible with a traditional scatter plot nor are currently available in the literature for binscatter and its generalizations.

The first, most intuitive display of these results is a confidence *band*. One may, for each bin $j = 1, 2, \ldots, J$, place a standard confidence interval around the sample mean $\hat{\mu}(x) = \bar{y}_j$. However, this is not a correct visualization of the uncertainty about $\mu(x)$ in the data set, and as such, can not be used to assess hypotheses of interest. For example, just because one cannot fit a line through all these intervals does not allow a researcher to conclude that $\mu(x)$ is nonlinear. A confidence band is the tool required here, which naturally generalizes the idea of confidence interval.

Loosely speaking, a band is simply a confidence "interval" for a function, and like a traditional confidence interval, it is given by the area between two "endpoint" functions, say $\hat{\mu}_{\mathtt{U}}(x)$ and $\hat{\mu}_{\mathtt{L}}(x)$. We may then make statements analogous to those for usual confidence intervals. For example, if this band does not contain a line (or quadratic function), then we say that at level α we reject the null hypothesis that

Figure III.7 Confidence Intervals and Confidence Bands.



(a) $p = 0$ and $s = 0$ **(b)** $p = 2$ and $s = 2$

(c) $p = 0$ and $s = 0$ **(d)** $p = 2$ and $s = 2$

Notes. Constructed using simulated data described in Appendix B.

$\mu(x)$ is linear (or quadratic). Visually, the size of the band reflects the uncertainty in the data, both in terms of overall uncertainty and any heteroskedasticity patterns. Figure III.7 shows a confidence bands for the same four data sets as in Figure III.2, and we see that the size and shape of the band reflects the underlying data.

We can use confidence bands, and associated statistical procedures, to test for a variety of substantive hypotheses, both for guiding further analysis and for economic meaning directly. Figure III.8 shows two examples: the left plot shows a rejection of linearity while the right plot indicates statistically significant group differences. Given the left result, a researcher may consider nonlinear regression modeling in the empirical analysis. Given the right plot, we conclude that the different relationship between y and x is different between the two groups shown, which may be of substantive interest in its own right. This is a nonparametric analogue of testing the significance of the interaction between x and a group dummy in a linear model. In this paper, we formalize this kind of test, which we refer to as parametric specification testing because a particular parametric specification for $\mu(x)$, namely the linear-in-parameters model $\mu(x) = \theta_0 + \theta_1 x$ is contrasted against the binscatter approximation of $\mu(x)$. Of course, we can test for any given parametric functional form for $\mu(x)$, including examples such as the Probit model $\mu(x) = \Phi(\theta_0 + \theta_1 x)$ or the log-linear model $\mu(x) = e^{\theta_0+\theta_1 x}$. Visually, we cannot reject a certain functional form if the confidence band contains a function of that type. Numerically, we can give a precise p-value for such a test.

Heuristically, our formal parametric specification testing approach is based on comparing the maximal empirical deviation between binscatter and the desired parametric specification for $\mu(x)$. If the parametric specification is correct, then there should no deviation beyond what is explained by random sampling for all evaluation points x; hence the connection with the confidence band for $\mu(x)$. The first three rows of Table III.1 illustrate our approach numerically.

In addition to parametric specification testing, we also develop graphical and formal testing procedures for nonparametric shape restrictions of $\mu(x)$. Prime examples of such tests include negativity, monotonicity or concavity of $\mu(x)$, and their reciprocals positivity and convexity, or course. Graphically, this can be tested as before: using Figure III.8 again we can assess whether $\mu(x)$ is "likely" to be monotonic, concave or positive, say. Formally, we can test all these features as a one-sided hypothesis test on $\mu(x)$ or its derivatives. To be more precise, negativity means $\mu(x) \leq 0$, monotonicity means $\mu^{(1)}(x) \leq 0$, and concavity means $\mu^{(2)}(x) \leq 0$. The second three rows of Table III.1 illustrate our approach to shape restriction testing numerically.

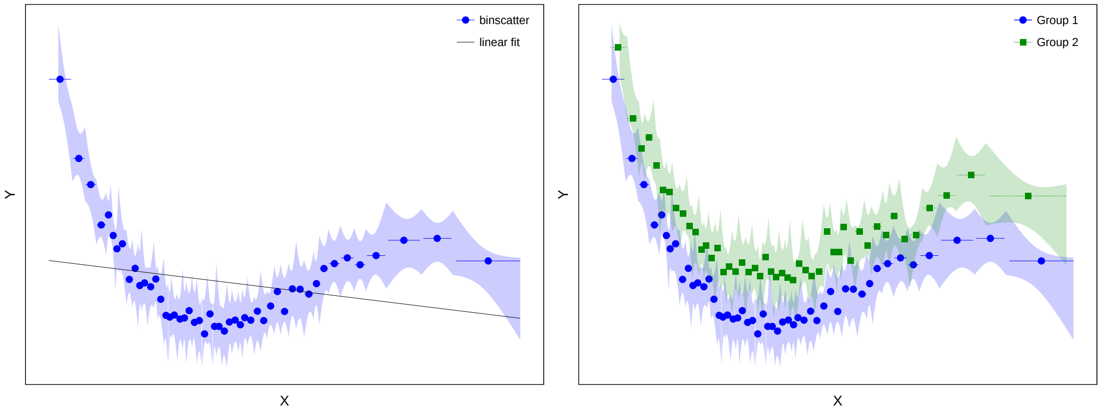
Table III.1 Formal Testing of Substantive Hypothesis.

	Half Support ($n = 482$)			**Full Support** ($n = 1000$)		
	Test Statistic	P-value	J	Test Statistic	P-value	J
Parametric Specification						
Constant	21.761	0.000	37	22.680	0	50
Linear	8.968	0.000	37	20.433	0	50
Quadratic	4.478	0.000	37	44.650	0	50
Shape Restrictions						
Negativity	187.185	0.000	37	188.414	0	50
Decreasingness	0.339	0.996	6	6.149	0	11
Concavity	6.009	0.000	3	8.976	0	5

Notes. Constructed using simulated data described in Appendix B.

Figure III.8 Graphical Testing of Substantive Hypotheses.

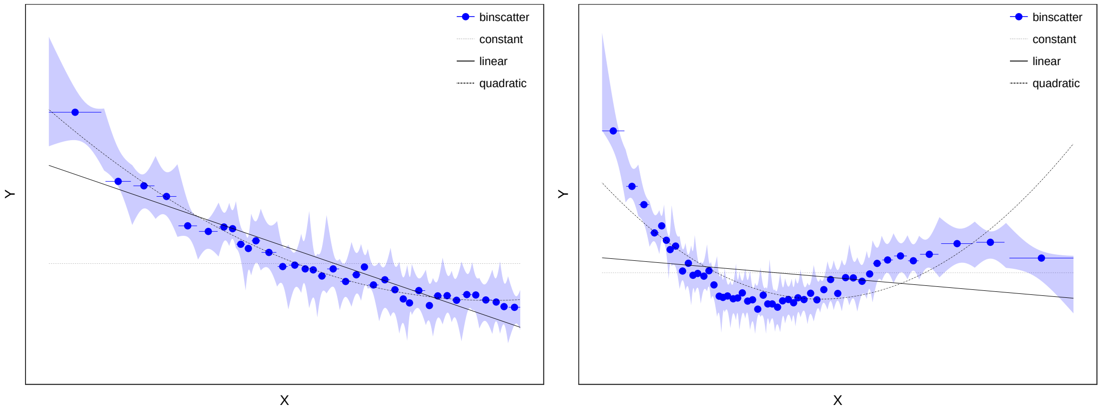
(a) Linear Regression Fit vs. Binscatter **(b)** Two-sample Comparison



Notes. Constructed using simulated data described in Appendix B.

Figure III.9 Graphical Representation of Parametric Specification Testing

(a) Half Support of x **(b)** Full Support of x



Notes. Constructed using simulated data described in Appendix B.

Precise results for estimation and inference, including regularity conditions and other technicalities, are summarized in Section 3.5.

Other Contributions: Software and Technicalities

Stepner (2014) gives an introduction to a very popular `Stata` software package implementing binscatter. This package implements canonical binscatter and residualized (covariate-adjusted) binscatter. Accompanying this chapter, we provide new software packages in `Stata` and `R`, which improve on current software implementations in several directions. First, we implement polynomial fits within bins and smoothness restrictions across bins for binscatter, and hence consider estimation and inference for both the regression function and its derivatives. Second, we implement fully data-driven selections of J, the number of bins, reflecting the features of the underlying data. Third, we implement covariate adjustments as discussed above, avoiding residualization, which leads to valid and interpretable methods for practice. Fourth, we implement valid distributional approximations leading to confidence intervals, confidence bands, and a wide range of parametric specification and nonparametric shape restriction hypothesis tests. Cattaneo, Crump, Farrell, and Feng (2019a) discusses all the details concerning our accompanying software and further illustrates it.

Finally, while not the focus on our paper, it is fair to underscore that studying in full generality standard empirical practice using binscatter forced us to develop new technical results that may be of independent interest. Our theoretical work is connected to the literature on nonparametric series estimation because binscatter is a partitioning-based nonparametric least squares estimator (e.g., Belloni, Chernozhukov, Chetverikov, and Kato, 2015; Belloni, Chernozhukov, Chetverikov, and Fernandez-Val, 2019; Cattaneo and Farrell, 2013; Cattaneo, Farrell, and Feng, 2018a, and references therein), and to the literature on partially linear semiparametric regression because of the way we incorporate covariate adjustments (e.g., Cattaneo, Jansson, and Newey, 2018a,b, and references therein). However, available technical results can not be used to analyze binscatter because it is implemented with a quantile-spaced binning, an example of random partitioning, generated by estimated quantiles.

As a consequence, our theoretical work necessarily relies on new results concerning non-/semi-parametric partitioning-based estimation on quantile-spaced (data-driven) partitions, which may be of independent interest. To be specific, we establish three main set of new theoretical results. First, we formally handle quantile-spaced (random) partitions underlying binscatter by resorting to appropriate empirical process techniques, substantially extending the results in Nobel (1996). Second, we obtain a general characterization of a linear map between piecewise polynomials and B-splines and give several technical results for it, properly accounting for quantile-spaced binning. Third, we develop a new strong approximation approach for the supremum of the t-statistic process building on ideas related to uniform distributional approximations of the supremum of stochastic process in Chernozhukov, Chetverikov, and Kato (2014a,b) and on the conditional coupling lemma used in Chapter II (Lemma A.3). Since this chapter is purposely practical, we relegate most discussions on our underlying technical work to Appendix B, unless it is strictly necessary for practical implementation or methodological interpretation of binscatter.

3.3 Formalizing Binscatter

We now begin our formal econometric-theoretical treatment of binscatter. Canonical binscatter builds on the standard regression model (3.1), and is constructed employing a quantile-spaced, disjoint partitioning of the support of x_i based on the observed data. To be precise, J disjoint intervals are constructed employing the empirical quantiles of x_i, leading to the partitioning scheme $\widehat{\Delta} = \{\widehat{\mathcal{B}}_1, \dots, \widehat{\mathcal{B}}_J\}$, where

$$\widehat{\mathcal{B}}_j = \begin{cases} \left[x_{(1)}, x_{(\lfloor n/J \rfloor)}\right) & \text{if } j = 1 \\ \left[x_{(\lfloor n(j-1)/J \rfloor)}, x_{(\lfloor nj/J \rfloor)}\right) & \text{if } j = 2, 3, \dots, J-1 \ , \\ \left[x_{(\lfloor n(J-1)/J \rfloor)}, x_{(n)}\right] & \text{if } j = J \end{cases}$$

$x_{(i)}$ denotes the i-th order statistic of the sample $\{x_1, x_2, \dots, x_n\}$, $\lfloor \cdot \rfloor$ is the floor operator, and $J < n$. Each estimated bin $\widehat{\mathcal{B}}_j$ contains roughly the same number of observations $N_j = \sum_{i=1}^{n} \mathbb{1}_{\widehat{\mathcal{B}}_j}(x_i)$, where $\mathbb{1}_{\mathcal{A}}(x) = \mathbb{1}(x \in \mathcal{A})$ with $\mathbb{1}(\cdot)$ denoting the indicator function. It follows that units are binned according to their rank in the x_i dimension.

Given the quantile-spaced partitioning scheme $\widehat{\Delta}$ for a choice of total number of bins J, the *canonical* binscatter estimator is

$$\widehat{\vartheta}(x) = \widehat{\mathbf{b}}(x)'\widehat{\boldsymbol{\beta}}, \qquad \widehat{\boldsymbol{\beta}} = \operatorname*{arg\,min}_{\boldsymbol{\beta} \in \mathbb{R}^J} \sum_{i=1}^{n} (y_i - \widehat{\mathbf{b}}(x_i)'\boldsymbol{\beta})^2, \tag{3.3}$$

where

$$\widehat{\mathbf{b}}(x) = \left[\ \mathbb{1}_{\widehat{\mathcal{B}}_1}(x) \quad \mathbb{1}_{\widehat{\mathcal{B}}_2}(x) \quad \cdots \quad \mathbb{1}_{\widehat{\mathcal{B}}_J}(x)\ \right]',$$

is the binscatter basis given by a J-dimensional vector of orthogonal dummy variables, that is, the j-th component of $\widehat{\mathbf{b}}(x)$ records whether the evaluation point x belongs to the j-th bin in the partition $\widehat{\Delta}$. Therefore, canonical binscatter can be expressed as the collection of J sample averages of the response variable y_i, one for each bin $\widehat{\mathcal{B}}_j$: $\bar{y}_j = \frac{1}{N_j}\sum_{i=1}^{n} \mathbb{1}_{\widehat{\mathcal{B}}_j}(x_i)y_i$ for $j = 1, 2, \ldots, J$. As illustrated in Section 3.2, empirical work employing canonical binscatter typically plots these binned sample averages along with some other estimate of the regression function $\vartheta(x)$.

3.3.1 Polynomial and Covariate Adjustments

We investigate the properties of binscatter in more generality. First, we allow for a more flexible polynomial regression approximation within each bin $\widehat{\mathcal{B}}_j$ forming the partitioning scheme $\widehat{\Delta}$, and thus expand the binscatter basis to allow for p-th order polynomial fits within each bin. For a choice of $p = 0, 1, 2, \ldots,$ we redefine

$$\widehat{\mathbf{b}}(x) = \begin{bmatrix} \mathbb{1}_{\widehat{\mathcal{B}}_1}(x) & \mathbb{1}_{\widehat{\mathcal{B}}_2}(x) & \cdots & \mathbb{1}_{\widehat{\mathcal{B}}_J}(x) \end{bmatrix}' \otimes \begin{bmatrix} 1 & x & \cdots & x^p \end{bmatrix}',$$

where now the binscatter basis is of dimension $(p + 1)J$. Setting $p = 0$ restores canonical binscatter. This generalization allows us to consider two important extensions of binscatter: (i) estimating derivatives of $\vartheta(\cdot)$, and (ii) incorporating smoothness restrictions across bins. Both will be very useful in Section 3.5 when we develop novel smooth confidence band estimators and formal hypothesis tests for shape restrictions.

Our second generalization of binscatter concerns covariate adjustment. As discussed in Section 3.2, we allow for additive separable covariate regression-based adjustment. Given the quantile-spaced partitioning scheme already introduced and a choice of p-th order polynomial approximation within bin, our proposed covariate-adjusted binscatter estimator is

$$\widehat{\mu}^{(v)}(x) = \widehat{\mathbf{b}}^{(v)}(x)'\widehat{\boldsymbol{\beta}}, \qquad \begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{\gamma}} \end{bmatrix} = \operatorname*{arg\,min}_{\boldsymbol{\beta},\boldsymbol{\gamma}} \sum_{i=1}^{n} (y_i - \widehat{\mathbf{b}}(x_i)'\boldsymbol{\beta} - \mathbf{w}_i'\boldsymbol{\gamma})^2, \qquad v \leq p, \quad (3.4)$$

using the standard notation $\mathbf{g}^{(v)}(x) = d^v\mathbf{g}(x)/dx^v$ for a function $\mathbf{g}(x)$ and $\mathbf{g}(x) = \mathbf{g}^{(0)}(x)$. The partially linear structure of model (3.2) naturally justifies our way of covariate adjustment, and sharply contrasts with the most common approach based on least squares residualization. See Section 3.3.3 below for more details. Note that with additional control variables $\mathbf{w}$, the estimand now is $\mu(\cdot)$ rather than $\vartheta(\cdot)$. We focus our analysis in the following mostly on $\mu(\cdot)$.

The generalized binscatter (3.4) reduces to the canonical binscatter (3.3) when

$p = 0 = v$ and $\boldsymbol{\gamma} = \mathbf{0}_d$, in which case $\widehat{\mu}(x) = \widehat{\mu}^{(0)}(x)$ becomes an step function (Figure III.3) reporting the sample averages $\bar{y}_j$ according to whether $x \in \widehat{\mathcal{B}}_j$, $j = 1, 2, \ldots, J$. The generalized binscatter $\widehat{\mu}^{(v)}(x)$ is useful to formalize commonly used empirical procedures based on binscatter, and to develop new binscatter-based estimation and inference procedures with better theoretical and practical properties.

3.3.2 Smoothness Restrictions

The binscatter estimator $\widehat{\mu}^{(v)}(x)$ retains the main features of canonical binscatter: estimation is conducted using only information within each (estimated) bin forming the quantile-spaced partition of the support of x_i. It follows that $\widehat{\mu}(x)$ is discontinuous at the boundaries of the J bins forming the partition $\widehat{\Delta}$; see Figure III.5. For some empirical analyses, both graphical and formal, researchers prefer a smoother binscatter of $\mu(\cdot)$, where the fits within each bin are constrained so that the final estimator exhibits some overall smoothness over the support of x_i. In this section we further generalize binscatter to provide such an alternative.

Given the quantile-spaced partitioning scheme, a smooth binscatter is the p-th order polynomial, s-times continuously differentiable, covariate-adjusted estimator given by

$$\widehat{\mu}^{(v)}(x) = \widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\boldsymbol{\beta}}, \qquad \begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{\gamma}} \end{bmatrix} = \operatorname*{arg\,min}_{\boldsymbol{\beta},\boldsymbol{\gamma}} \sum_{i=1}^{n} (y_i - \widehat{\mathbf{b}}_s(x_i)'\boldsymbol{\beta} - \mathbf{w}_i'\boldsymbol{\gamma})^2, \qquad s \leq p, \tag{3.5}$$

where $\widehat{\mathbf{b}}_s(x) = \widehat{\mathbf{T}}_s \widehat{\mathbf{b}}(x)$ with $\widehat{\mathbf{T}}_s$ being a $[(p+1)J - (J-1)s] \times (p+1)J$ matrix of linear restrictions ensuring that the $(s-1)$-th derivative of $\widehat{\mu}(x) = \widehat{\mu}^{(0)}(x)$ is continuous. When $s = 0$, $\widehat{\mathbf{T}}_0 = \mathbf{I}_{(p+1)J}$, the identity matrix of dimension $(p+1)J$, and therefore no restrictions are imposed: $\widehat{\mathbf{b}}(x) = \widehat{\mathbf{b}}_0(x)$, as given in the previous subsection. Consequently, if $s = 0$, we obtain the binscatter $\widehat{\mu}(x)$, which is not a continuous function estimate. On the other hand, $p \geq s$ implies that a least squares p-th order polynomial fit is constructed within each bin $\widehat{\mathcal{B}}_j$, in which case setting $s = 1$ forces these fits to be connected at the boundaries of adjacent bins, $s = 2$ forces these fits to be connected and continuously differentiable at the boundaries of adjacent bins, and so on, for $s = 3, 4, \ldots, p$. This is the formalization leading to Figure III.5.

Enforcing smoothness for binscatter boils down to incorporating restrictions on the binscatter basis. The resulting constrained basis, $\widehat{\mathbf{b}}_s(x)$, corresponds to a choice of spline basis for approximation of $\mu(\cdot)$, with estimated quantile-spaced knots according to the partition $\widehat{\Delta}$. In this paper, we employ $\widehat{\mathbf{T}}_s$ leading to B-splines, which tend to have very good finite sample properties, but other choices are of course possible. Smooth binscatter (3.5) reduces to binscatter (3.4) when $s = 0$, and therefore the former is a strict generalization of latter and hence, in particular, of the canonical binscatter (3.3).

3.3.3 Comparison to the Canonical Residualized Binscatter

Current widespread empirical practice for covariate adjustment of binscatter proceeds by first regressing out the covariates $\mathbf{w}_i$, and then applying canonical binscatter on the residualized variables of interest. To be precise, standard practice applies (3.3) upon replacing y_i by $y_i - \widetilde{\mathbf{w}}_i'\widehat{\boldsymbol{\delta}}_{y.\widetilde{w}}$ and x_i by $x_i - \widetilde{\mathbf{w}}_i'\widehat{\boldsymbol{\delta}}_{x.\widetilde{w}}$, where $\widetilde{\mathbf{w}}_i = (1, \mathbf{w}_i')'$, and $\widehat{\boldsymbol{\delta}}_{y.\widetilde{w}}$ and $\widehat{\boldsymbol{\delta}}_{x.\widetilde{w}}$ denote the OLS coefficients obtained from regressing y on $\mathbf{w}$ and x on $\mathbf{w}$, respectively, with each regression including a constant term. This is the default (and only) implementation of covariate adjustment in standard binscatter software widely used in practice (Stepner, 2014).

Under mild assumptions, the estimators $\widehat{\boldsymbol{\delta}}_{y.\widetilde{w}}$ and $\widehat{\boldsymbol{\delta}}_{x.\widetilde{w}}$ are consistent for $\boldsymbol{\delta}_{y.\widetilde{w}} = \mathbb{E}[\widetilde{\mathbf{w}}_i\widetilde{\mathbf{w}}_i']^{-1}\mathbb{E}[\widetilde{\mathbf{w}}_i y_i]$ and $\boldsymbol{\delta}_{x.\widetilde{w}} = \mathbb{E}[\widetilde{\mathbf{w}}_i\widetilde{\mathbf{w}}_i']^{-1}\mathbb{E}[\widetilde{\mathbf{w}}_i x_i]$, respectively. As it is customary in applied work, $\widetilde{\mathbf{w}}'\boldsymbol{\delta}_{y.\widetilde{w}}$ and $\widetilde{\mathbf{w}}'\boldsymbol{\delta}_{x.\widetilde{w}}$ can be interpreted as a "best" linear approximation to $\mathbb{E}[y|\mathbf{w}]$ and $\mathbb{E}[x|\mathbf{w}]$, respectively. It can be argued that, under non-trivial assumptions, the residualized binscatter approximates the conditional expectation $\mathbb{E}[y - \widetilde{\mathbf{w}}'\boldsymbol{\delta}_{y.\widetilde{w}}|x - \widetilde{\mathbf{w}}'\boldsymbol{\delta}_{x.\widetilde{w}}]$, a parameter that is quite difficult to interpret. Consequently, as illustrated in Figure III.4, residualized binscatter does not consistently estimate $\mu(x)$ in model (3.2), nor $\mathbb{E}[y_i|x_i]$ in general. Under additional restrictive assumptions, the probability limit of residualized binscatter does have some interpretation when model (3.2) holds: if x and $\mathbf{w}$ are uncorrelated, then $\boldsymbol{\delta}_{x.\widetilde{w}} = (\mathbb{E}[x], \mathbf{0}')'$, and the residualized binscatter procedure consistently estimates

$$\mathbb{E}[y - \widetilde{\mathbf{w}}'\boldsymbol{\delta}_{y.\widetilde{w}}|x - \mathbb{E}[x]] = \mu(x) - \mathbb{E}[y] + \mathbb{E}\big[\mathbf{w}|x - \mathbb{E}[x]\big]'(\boldsymbol{\gamma} - \check{\boldsymbol{\delta}}_{y.\widetilde{w}}),$$

where $\check{\boldsymbol{\delta}}_{y.\widetilde{w}} = \mathbb{E}[(\mathbf{w}_i - \mathbb{E}[\mathbf{w}_i])\mathbf{w}_i']^{-1}\mathbb{E}[(\mathbf{w}_i - \mathbb{E}[\mathbf{w}_i])y_i]$. This estimand is clearly not equal to $\mu(x)$ unless additional assumptions hold.

When model (3.2) is misspecified for $\mathbb{E}[y|x, \mathbf{w}]$, the probability limit of both residualized binscatter and our recommended covariate-adjusted binscatter changes. In the case of residualized binscatter, the probability limit becomes quite difficult to interpret and relate to any meaningful "partial effect" of x on y. On the other hand, our approach to covariate adjustment retains the usual interpretation of standard semiparametric semi-linear models, where the true unknown "long" conditional expectation $\mathbb{E}[y|x, \mathbf{w}]$ is approximated by the closest model $\mu(x) + \mathbf{w}'\boldsymbol{\gamma}$ in a mean square error sense. See Angrist and Pischke (2008) for further discussion on the interpretation of (semi-)linear least squares approximations, and its uses in applied work.

For the reasons above, we recommend to covariate-adjust binscatter by incorporating covariates in an additively separable way, as in (3.5), and not via residualization as currently done in most empirical applications.

3.4 Implementing Binscatter

Binscatter is readily implementable once the number of bins J is chosen, for any polynomial order p, level of smoothness constrain $s \leq p$, and derivative of interest $v \leq p$. Therefore, for implementation purposes, we discuss first a valid and optimal choice of J based on minimizing the IMSE of binscatter as a point estimator of $\mu^{(v)}(x)$ in model (3.2), given the researchers' choice of p, s, and v. This IMSE-optimal selection procedure can be viewed as a special case of applying the general results given in Chapter II, Section 2.4, with the randomness of quantile-based partitions appropriately treated. We defer the more detailed discussion of IMSE expansion and the corresponding implementation procedures to Chapter IV.

The following basic assumption is the only one used throughout this chapter.

Assumption III.1. *The sample $(y_i, x_i, \mathbf{w}_i')$, $i = 1, 2, \ldots, n$, is i.i.d and satisfies model* (3.2). *Further, the covariate x_i has a continuous density function $f(x)$ bounded away from zero on the support $\mathcal{X}$, $\mathbb{E}[\mathbb{V}[\mathbf{w}_i|x_i]] > 0$, $\sigma^2(x) = \mathbb{E}[\epsilon_i^2|x_i = x]$ is continuous and bounded away from zero, and $\mathbb{E}[\|\mathbf{w}_i\|^4|x_i = x]$, $\mathbb{E}[|\epsilon_i|^4|x_i = x]$ and $\mathbb{E}[|\epsilon_i|^2|x_i = x, \mathbf{w}_i = \mathbf{w}]$ are uniformly bounded. Finally, $\mu(x)$ and $\mathbb{E}[\mathbf{w}_i|x_i = x]$ are $(p + q + 1)$-times continuously differentiable form some $q \geq 1$.*

This assumption is not minimal, but is nonetheless mild because it imposes standard conditions in classical regression settings. When the covariates $\mathbf{w}_i$ are not adjusted for in the binscatter, all statements involving these covariates in Assumption III.1 can be ignored.

To select the number of bins J forming the quantile-spaced partition $\widehat{\Delta}$ used by binscatter, we proposed to minimize an approximation to the density-weighted integrated mean square error of the estimator $\widehat{\mu}^{(v)}(x)$. Letting $\approx_{\mathbb{P}}$ denote an approximation in probability, we show that

$$\int \left(\widehat{\mu}^{(v)}(x) - \mu^{(v)}(x)\right)^2 f(x)dx \approx_{\mathbb{P}} \frac{J^{1+2v}}{n}\mathscr{V}_n(p, s, v) + J^{-2(p+1-v)}\mathscr{B}_n(p, s, v)$$

where these two terms capture the asymptotic variance and (squared) bias of binscatter, respectively, as a function of the polynomial order used (p), the desired derivative to be approximated (v), and the level of smoothness imposed across bins (s). Both quantities are fully characterized in Chapter IV, where they are shown to be non-random functions of the sample size n, at this level of generality. The variance $\mathscr{V}_n(p,s,v)$, depending on $\sigma^2(x)$ and $f(x)$, is bounded and bounded away from zero under minimal assumptions, while the (squared) bias $\mathscr{B}_n(p,s,v)$, depending on $\mu^{(p+1)}(x)$ and $f(x)$, is generally bounded and bounded away from zero. Our precise characterization of $\mathscr{V}_n(p,s,v)$ and $\mathscr{B}_n(p,s,v)$ is useful to approximate them in practice for implementation purposes. Furthermore, we show that $\mathscr{V}_n(p,0,v) \to \mathscr{V}(p,0,v)$ and $\mathscr{B}_n(p,0,v) \to \mathscr{B}(p,0,v)$, where $\mathscr{V}(p,0,v)$ and $\mathscr{B}(p,0,v)$ are cumbersome quantities in general. However, for special leading cases, the variance and (squared) bias take very simple forms: $\mathscr{V}(0,0,0) = \mathbb{E}[\sigma^2(x_i)]$ and $\mathscr{B}(0,0,0) = \frac{1}{12}\mathbb{E}\big[\big(\frac{\mu^{(1)}(x_i)}{f(x_i)}\big)^2\big]$, which corresponds to canonical binscatter ($p = v = s = 0$).

The main takeaway is that the IMSE of binscatter naturally depends on the squared bias and variance of the estimator, and these factors can be balanced out in order to select the IMSE-optimal number of bins to use in applications. The following theorem summarizes this result.

Theorem III.1 (IMSE-Optimal Binscatter)**.** *Let Assumption III.1 hold,* $0 \leq v, s \leq p$, *and* $J\log(J)/n \to 0$ *and* $nJ^{-4p-5} \to 0$. *Then, the IMSE-optimal number of bins for implementing binscatter is*

$$J_{IMSE} = \left\lceil \left(\frac{2(p-v+1)\mathscr{B}_n(p,s,v)}{(1+2v)\mathscr{V}_n(p,s,v)} \right)^{\frac{1}{2p+3}} \; n^{\frac{1}{2p+3}} \right\rceil,$$

where $\lceil\cdot\rceil$ *denotes the ceiling operator.*

This theorem gives the optimal choice of J for the general class of binscatter considered in this paper, that is, allowing for higher-order polynomial fits within bins and imposing smooth restrictions on the fits across bins, with or without covariate adjustment, when the main object of interest is possibly a derivative of the unknown function $\mu(\cdot)$. This additional versatility will be useful in upcoming sections when constructing formal statistical testing procedures based on binscatter derivative estimates. In particular, the optimal number of bins for the canonical binscatter is obtained when $p = v = s = 0$.

As discussed in Section 3.2, most common practice set $s = 0$ first, in which the size of the partition is chosen without smoothness restrictions, even if later those restrictions are imposed and used for constructing smoother regression estimates and confidence bands. An important result emerging from Theorem III.1 is that this approach is justified in large samples because the optimal number of bins for any $0 \leq s \leq p$ is proportional to $n^{\frac{1}{2p+3}}$, and therefore choosing J with or without imposing smoothness restrictions leads to an IMSE rate optimal binscatter —only the constant of proportionality changes slightly depending on the s chosen.

3.5 Using Binscatter

Our generalized binscatter estimator $\widehat{\mu}^{(v)}(x)$, with $0 \leq p$ and $0 \leq v, s \leq p$, is constructed to approximate the function $\mu^{(v)}(x)$ in model (3.2), which captures the v-th derivative partial effect of x on y, after controlling for $\mathbf{w}$. Viewed as a semi-/nonparametric estimator, binscatter can be implemented in a valid and IMSE-optimal way by setting $J = J_{\texttt{IMSE}}$ (Theorem III.1) when forming the bins partitioning the support of x.

In this section we employ binscatter for three main purposes. First, we discuss valid and optimal graphical presentation of the regression function and its derivatives. Second, we offer valid confidence intervals and bands for $\mu^{(v)}(x)$. Finally, we develop valid hypothesis tests for parametric specification and nonparametric shape restrictions of $\mu^{(v)}(x)$. All the results discussed in this section are formalizations of the procedures already illustrated in Section 3.2.

3.5.1 Graphical Presentation

We proved that the binscatter estimator $\widehat{\mu}^{(v)}(x)$, implemented with $J = J_{\texttt{IMSE}}$ as in Theorem III.1, is an IMSE-optimal point estimator of $\mu^{(v)}(x)$. Furthermore, we also show there that binscatter can achieve the fastest uniform rate of convergence. These results highlight some of the good statistical properties of binscatter, and justify its use for depicting an approximation to the unknown function $\mu(x)$.

In Section 3.2, we illustrated several of the resulting binned scatter plots, all constructed using $\widehat{\mu}^{(v)}(x)$ for appropriate choice of polynomial order within bin (p), smoothness level (s), and derivative of interest (v). To describe how these plots are constructed, let $\bar{b}_j$ denote the center of the j-th quantile-spaced bin $\widehat{\mathcal{B}}_j$, where $j = 1, 2, \ldots, J$. Then, the dots in the binned scatter plot correspond to $\widehat{\mu}(\bar{b}_j)$ for any choice of $0 \leq v, s \leq p$. In Figure III.5 we illustrated the effects of varying s and p by plotting $\widehat{\mu}(x)$ as a function of x. When $s = 0$, the resulting estimator $\widehat{\mu}(x)$ is discontinuous at the bins' edges, while when $s > 0$ it is at least continuous.

In constructing a binned scatter plot, it may be convenient to report more than one estimate of $\mu(x)$ over the same quantile-spaced bins. For example, researchers can report a collection of "dots" using $\widehat{\mu}(\bar{b}_j)$, $j = 1, 2, \ldots, J$, with $p = s = 0$ (canonical binscatter), and a "line" representing a smoother estimate such as $\widehat{\mu}(x)$, $x \in \mathcal{X}$, with $p = s = 3$ (cubic B-spline).

Finally, while not graphically illustrated in Section 3.2, derivative estimates can also lead to powerful and useful binned scatter plots. Specifically, in some applications researchers may be interested in an "average marginal effect" of x on y, possibly controlling for other factors $\mathbf{w}$, which is naturally captured by $\mu^{(1)}(x)$. Such a quantity is of interest in many different setups, ranging from reduced form latent variable models to structural non-separable models. Furthermore, derivatives of $\mu(x)$ are of interest in testing for substantive hypotheses such as marginal diminishing returns. We formalize these latter ideas further below.

3.5.2 Pointwise Inference and Confidence Intervals

We turn now to confidence interval and confidence band estimators based on binscatter. The Studentized t-statistic is

$$\widehat{T}_p(x) = \frac{\widehat{\mu}^{(v)}(x) - \mu^{(v)}(x)}{\sqrt{\widehat{\Omega}(x)/n}}, \qquad 0 \leq v, s \leq p,$$

where the binscatter variance estimator is

$$\widehat{\Omega}(x) = \widehat{\mathbf{b}}_s^{(v)}(x)' \widehat{\mathbf{Q}}^{-1} \widehat{\mathbf{\Sigma}} \widehat{\mathbf{Q}}^{-1} \widehat{\mathbf{b}}_s^{(v)}(x),$$

$$\widehat{\mathbf{Q}} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)', \qquad \widehat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'(y_i - \widehat{\mathbf{b}}_s(x_i)'\widehat{\boldsymbol{\beta}} - \mathbf{w}_i'\widehat{\boldsymbol{\gamma}})^2.$$

Lemma III.1 (Distributional Approximation: Pointwise)**.** *Let Assumption III.1 hold,* $0 \leq v, s \leq p$*, and* $J^2 \log^2(J)/n \to 0$ *and* $nJ^{-2p-3} \to 0$*. Then,*

$$\sup_{u \in \mathbb{R}} \left| \mathbb{P}\big[\widehat{T}_p(x) \leq u\big] - \Phi(u) \right| \to 0, \qquad \textit{for each } x \in \mathcal{X},$$

where $\Phi(u)$ *denotes the distribution function of a standard normal random variable.*

Lemma III.1 can be used to form asymptotically valid confidence intervals for $\mu^{(v)}(x)$, pointwise in $x \in \mathcal{X}$, provided the misspecification error introduced by binscatter is removed from the distributional approximation. Specifically, for a choice p, the confidence intervals take the form:

$$\widehat{I}_p(x) = \Big[\, \widehat{\mu}^{(v)}(x) \pm \mathfrak{c} \cdot \sqrt{\widehat{\Omega}(x)/n}\, \Big], \qquad 0 \le v, s \le p,$$

where $\mathfrak{c}$ denotes a choice of quantile (e.g., $\mathfrak{c} \approx 1.96$ for a 95% Gaussian confidence intervals). However, employing an IMSE-optimal binscatter (e.g., Theorem III.1) introduces a first-order misspecification error leading to invalidity of these confidence intervals. To address this problem, we rely on a simple application of robust bias correction (Calonico, Cattaneo, and Titiunik, 2014; Calonico, Cattaneo, and Farrell, 2018b; Cattaneo, Farrell, and Feng, 2018a) to form valid confidence intervals based on IMSE-optimal binscatter, that is, without altering the partitioning scheme $\widehat{\Delta}$ used.

Our proposed robust bias-corrected binscatter confidence intervals are constructed as follows. First, for a given choice of p, we select the number of bins in $\widehat{\Delta}$ according to Theorem III.1, and construct the binscatter accordingly. Then, we employ the confidence interval $\widehat{I}_{p+q}(x)$ with $\mathfrak{c} = \Phi^{-1}(1-\alpha/2)$. The following theorem summarizes this approach.

Theorem III.2 (Confidence Intervals). *For given p, suppose the conditions in Lemma III.1 hold and $J = J_{\mathtt{IMSE}}$. If $\mathfrak{c} = \Phi^{-1}(1-\alpha/2)$, then*

$$\mathbb{P}\Big[\mu^{(v)}(x) \in \widehat{I}_{p+q}(x)\Big] \to 1-\alpha, \qquad \textit{for all } x \in \mathcal{X}.$$

The confidence intervals constructed in the above theorem are based on an IMSE-optimal implementation of binscatter and robust bias correction. They were illustrated in Figure III.7 as individual vertical segments inside the shaded bands, which are discussed in the next subsection.

3.5.3 Uniform Inference and Confidence Bands

In many empirical applications of binscatter, the goal is to conduct inference uniformly over $x \in \mathcal{X}$ as opposed to pointwise as in the previous section. Examples include reporting confidence bands for $\mu(x)$ and its derivatives, as well as testing for linearity, monotonicity, concavity, or other shape features of $\mu^{(v)}(x)$. This section applies a formal approach for uniform inference employing binscatter and its generalizations, and constructs valid confidence bands based on binscatter and its generalizations. In the following subsections, we employ these uniform inference results to develop asymptotically valid testing procedures for parametric model specification and nonparametric shape restrictions.

Our approach to uniform inference extends the work on strong approximations in Chapter II to allow for estimated quantile-spaced partitioning $\widehat{\Delta}$, as commonly used in binscatter settings, which requires non-trivial additional technical work. In fact, it is not possible to obtain a valid strong approximation for the entire stochastic process $\{\widehat{T}_p(x) : x \in \mathcal{X}\}$, as done in Chapter II, because uniformity fundamentally fails when the partitioning scheme is random: see Appendix B for details. Inspired by the work in Chernozhukov, Chetverikov, and Kato (2014a,b), our approach circumvents this technical hurdle by retaining the randomness introduced by $\widehat{\Delta}$, and focusing instead on the specific functional of interest (i.e., suprema).

In this section we apply these results to construct valid robust bias-corrected confidence bands for $\mu^{(v)}(x)$, while in the next two upcoming sections we employ them to develop valid testing procedures. For a choice of p, $0 \leq v, s \leq p$, and quantile-spaced partition size J, we define

$$\left\{\widehat{I}_{p+q}(x) : x \in \mathcal{X}\right\} \quad \text{with} \quad \mathfrak{c} = \inf\left\{c \in \mathbb{R}_+ : \mathbb{P}\left[\sup_{x\in\mathcal{X}} \left|\widehat{T}_{p+q}(x)\right| \leq c\right] \geq 1-\alpha\right\}.$$

By construction, this band covers the entire function $\mu^{(v)}(x)$ with probability $1-\alpha$.

The main drawback in the construction above, however, is that the quantiles $\mathfrak{c}$ are unknown because the finite sample distribution of $\sup_{x\in\mathcal{X}} \left|\widehat{T}_{p+q}(x)\right|$ is unknown. Our strong approximation results allow us to approximate this distribution by resampling from a Gaussian vector of length $(p+q+1)J-(J-1)s$. To be more precise, let $\mathbf{N}_K$ be a K-dimensional standard normal random vector, and define the following (conditional) Gaussian process:

$$\widehat{Z}_p(x) = \frac{\widehat{\mathbf{b}}^{(v)}(x)'\widehat{\mathbf{Q}}^{-1}\widehat{\mathbf{\Sigma}}^{-1/2}}{\sqrt{\widehat{\Omega}(x)/n}}\mathbf{N}_{(p+1)J-(J-1)s}, \qquad x \in \mathcal{X}, \quad 0 \leq v, s \leq p. \tag{3.6}$$

We show that the distribution of $\sup_{x\in\mathcal{X}} \left|\widehat{T}_p(x)\right|$ is well approximated by that of $\sup_{x\in\mathcal{X}} \left|\widehat{Z}_p(x)\right|$, conditional on the data $\mathbf{D} = \{(y_i, x_i, \mathbf{w}_i') : i = 1, 2, \ldots, n\}$. This result implies that the quantiles used to construct confidence bands can be approximated by resampling from the normal random vectors $\mathbf{N}_{(p+1)J-(J-1)s}$, keeping the data $\mathbf{D}$ fixed. We make this approach precise in the following theorem.

Lemma III.2 (Distributional Approximation: Supremum). *Let Assumption III.1 hold,* $0 \leq v, s \leq p$, *and* $J^2\log^6(J)/n \to 0$ *and* $nJ^{-2p-3}\log J \to 0$. *Then,*

$$\sup_{u\in\mathbb{R}} \left|\mathbb{P}\left[\sup_{x\in\mathcal{X}} |\widehat{T}_p(x)| \leq u\right] - \mathbb{P}\left[\sup_{x\in\mathcal{X}} |\widehat{Z}_p(x)| \leq u \middle| \mathbf{D}\right]\right| \to_{\mathbb{P}} 0.$$

where $\to_{\mathbb{P}}$ *denotes convergence in probability.*

Putting the above together, we have the following main result for robust bias-corrected confidence bands.

Theorem III.3 (Confidence Bands)**.** *For given p, suppose the conditions in Lemma III.2 hold and* $J = J_{\texttt{IMSE}}$*. If* $\mathfrak{c} = \inf\big\{c \in \mathbb{R}_+ : \mathbb{P}\big[\sup_{x\in\mathcal{X}} \big|\widehat{Z}_{p+q}(x)\big| \le c \,\big|\, \mathbf{D}\big] \ge 1-\alpha\big\}$, *then*

$$\mathbb{P}\Big[\mu^{(v)}(x) \in \widehat{I}_{p+q}(x),\ \textit{for all } x \in \mathcal{X}\Big] \to 1-\alpha.$$

This theorem offers a valid confidence bands construction for $\mu^{(v)}(\cdot)$, which relies on resampling from a particular random variable: $\sup_{x\in\mathcal{X}} |\widehat{Z}_{p+q}(x)|$, conditional on the original data. In practice, this supremum is replaced by a maximum over a fine grid of evaluation points, and each realization of $\widehat{Z}_{p+q}(x)$ is obtained by resampling from the standard normal random vector $\mathbf{N}_{(p+q+1)J-(J-1)s}$ and then computing $\widehat{Z}_{p+q}(x)$ as in (3.6), where all other quantities are fixed and known given the original data. As a consequence, the quantile $\mathfrak{c}$ is actually estimated conditional on $\mathbf{D}$. Further details on implementation are given in our companion software package (Cattaneo, Crump, Farrell, and Feng, 2019a).

3.5.4 Testing Parametric Specifications

Binscatter is often used to heuristically assess different types of shape features of the unknown regression function and its derivatives. In this section, we provide a rigorous formalization of one such kind of hypothesis tests: parametric specifications of $\mu^{(v)}(x)$. In the next section, we discuss another type of shape-related hypothesis test: testing for nonparametric features such as monotonicity or concavity of $\mu^{(v)}(x)$.

One type of informal analysis commonly encountered in empirical work concerns comparing the binscatter $\widehat{\mu}^{(v)}(x)$ relative to some parametric fit. For example, $\widehat{\mu}(x)$ can be compared to $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ to assess whether there is a relationship between y_i and x_i or, more formally, whether $\mu(x)$ is a constant function. Similarly, it is common to see binscatter used to assess whether there is a linear or perhaps quadratic relationship, that is, whether $\mu(x) = \theta_0 + x\theta_1$ or perhaps $\mu(x) = \theta_0 + x\theta_1 + x^2\theta_2$ for some unknown coefficients $\boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2)'$. More generally, researchers are often interested in assessing formally whether $\mu(x) = m(x, \boldsymbol{\theta})$ for some $m(\cdot)$ known up to a finite parameter $\boldsymbol{\theta}$, which can be estimated using the available data. We formalize this class of hypothesis tests as follows: for a choice of v and function $m^{(v)}(x, \boldsymbol{\theta})$ with

$\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^{d_\theta}$, the null and alternative hypotheses are

$$\begin{aligned} \ddot{\mathsf{H}}_0: &\quad \sup_{x\in\mathcal{X}} \left|\mu^{(v)}(x) - m^{(v)}(x,\boldsymbol{\theta})\right| = 0, \qquad \text{for some } \boldsymbol{\theta} \in \boldsymbol{\Theta}, \qquad \mathit{vs.} \\ \ddot{\mathsf{H}}_\mathrm{A}: &\quad \sup_{x\in\mathcal{X}} \left|\mu^{(v)}(x) - m^{(v)}(x,\boldsymbol{\theta})\right| > 0, \qquad \text{for all } \boldsymbol{\theta} \in \boldsymbol{\Theta}. \end{aligned}$$

As is evident from its formulation, this testing problem can be implemented using test statistics involving the supremum of (derivatives of) binscatter, with or without employing higher-order polynomials, imposing smoothness restrictions, or adjusting for additional covariates. Crucially, in all cases it is required to approximate the quantiles of the finite sample distribution of such statistics, which can be done in a similar fashion as discussed above for constructing confidence bands.

Since $\boldsymbol{\theta}$ is unknown and not set by the null hypothesis, we construct a feasible testing procedure by assuming that there exists an estimator $\widehat{\boldsymbol{\theta}}$ that consistently estimates $\boldsymbol{\theta}$ under the null hypothesis (correct parametric specification), and that is "well behaved" under the alternative hypothesis (parametric misspecification). See Theorem III.4 below for precise restrictions. Then, we consider the following test statistic

$$\ddot{T}_p(x) = \frac{\widehat{\mu}^{(v)}(x) - m^{(v)}(x,\widehat{\boldsymbol{\theta}})}{\sqrt{\widehat{\Omega}(x)/n}}, \qquad 0 \le v, s \le p,$$

leading to the hypothesis test:

$$\text{Reject } \ddot{\mathsf{H}}_0 \qquad \text{if and only if} \qquad \sup_{x\in\mathcal{X}} |\ddot{T}_p(x)| \ge \mathfrak{c}, \tag{3.7}$$

for an appropriate choice of critical value $\mathfrak{c}$ to control false rejections (Type I error).

The following theorem gives the remaining details, and makes the hypothesis testing procedure (3.7) feasible.

Theorem III.4 (Hypothesis Testing: Parametric Specification)**.** *Let Assumption III.1 hold. For given p, and $0 \le v, s \le p$, set $J = J_{\texttt{IMSE}}$ and $\mathfrak{c} = \inf\big\{c \in \mathbb{R}_+ : \mathbb{P}\big[\sup_{x\in\mathcal{X}} |\widehat{Z}_{p+q}(x)| \le c \,\big|\, \mathbf{D}\big] \ge 1-\alpha\big\}$.*

Under $\ddot{\mathsf{H}}_0$, if $\sup_{x\in\mathcal{X}} |m^{(v)}(x,\widehat{\boldsymbol{\theta}}) - \mu^{(v)}(x)| = O_\mathbb{P}(n^{-1/2})$, then

$$\lim_{n\to\infty} \mathbb{P}\Big[\sup_{x\in\mathcal{X}} \big|\ddot{T}_{p+q}(x)\big| > \mathfrak{c}\Big] = \alpha.$$

Under $\ddot{\mathsf{H}}_A$, if there exists some $\bar{\boldsymbol{\theta}}$ such that $\sup_{x\in\mathcal{X}} |m^{(v)}(x,\widehat{\boldsymbol{\theta}}) - m^{(v)}(x,\bar{\boldsymbol{\theta}})| =$

$O_{\mathbb{P}}(n^{-1/2})$, *then*

$$\lim_{n\to\infty} \mathbb{P}\Big[\sup_{x\in\mathcal{X}} \big|\ddot{T}_{p+q}(x)\big| > \mathfrak{c}\Big] = 1.$$

This theorem formalizes a very intuitive idea: if the confidence band for $\mu^{(v)}(x)$ does not contain entirely the parametric fit considered, then such parametric fit is inconsistent with the data, i.e., should be rejected. Formally, this leads to the hypothesis testing procedure (3.7), which relies on a proper (simulated) critical value. The condition $\sup_{x\in\mathcal{X}} |m^{(v)}(x, \widehat{\boldsymbol{\theta}}) - \mu^{(v)}(x)| = O_{\mathbb{P}}(n^{-1/2})$ under the null hypothesis is very mild: it states that the unknown parameters entering the parametric specification $\mu(x) = m(x, \boldsymbol{\theta})$ is $\sqrt{n}$-estimable, provided some mild regularity holds for the known regression function $m(x, \boldsymbol{\theta})$. For example, a simple sufficient condition is $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = O_{\mathbb{P}}(1)$ and $m(x, \boldsymbol{\theta})$ continuous in x and continuously differentiable in $\boldsymbol{\theta}$. Most standard parametric models in applied microeconomics satisfy such conditions, including linear and non-linear regression models, discrete choice models (Probit or Logit), censored and truncation models, and many more.

In practice, it is natural to combine the formal hypothesis test emerging from Theorem III.4 with a binned scatter plot that includes a binscatter confidence band and a line representing the parametric fit. Section 2 illustrated this with Table III.1 and Figure III.8.

Remark III.1 (Other Metrics)*.* The parametric specification test in (3.7) is based on the maximum discrepancy between the fit of the hypothesized parametric model for $\mu(x)$ and the nonparametric binscatter approximation. Some practitioners, however, may prefer to assess the discrepancy by means of an alternative metric, such as the mean square difference between the parametric and nonparametric fits. Our theoretical results given in the appendix are general enough to accommodate such alternative comparisons, but we do not discuss them here only to conserve space.

3.5.5 Testing Shape Restrictions

The hypothesis test (3.7) concerns parametric specification testing for a choice of $m(x, \boldsymbol{\theta})$, but it can also be used to conduct certain nonparametric shape restriction testing. For example, if the function $\mu(x)$ is constant, then $\mu^{(1)}(x) = 0$ for all $x \in \mathcal{X}$, which can be implemented using Theorem III.4 upon setting $m(\cdot) = 0$ and $v = 1$, for any $p \geq 1$ and $0 \leq s \leq p$. Similarly, linearity or other related nonparametric shape restrictions can be tested for via the results in Theorem III.4, for appropriate choice of v. The common feature in all cases is that the null hypothesis of interest is two-sided.

There are, however, other important nonparametric shape restriction hypotheses about $\mu(x)$ that correspond to one-sided null hypothesis, and thus cannot be implemented using Theorem III.4. For example, negativity, monotonicity and concavity of $\mu(x)$ all correspond to formal statements of the form $\mu(x) \leq 0$, $\mu^{(1)}(x) \leq 0$ and $\mu^{(2)}(x) \leq 0$, respectively. Thus, in this section we also study the following class of hypothesis tests: for a choice of v, the null and alternative hypotheses are

$$\dot{\mathsf{H}}_0 : \quad \sup_{x\in\mathcal{X}} \mu^{(v)}(x) \leq 0, \qquad vs. \qquad \dot{\mathsf{H}}_\text{A} : \quad \sup_{x\in\mathcal{X}} \mu^{(v)}(x) > 0.$$

These hypotheses highlight the importance of extending binscatter to derivative estimation, which necessarily requires considering $p \geq v > 0$, with or without smoothness restrictions or covariate adjustments. In other words, considering higher-order polynomial fits within bins is not a spurious generalization of binscatter, but rather a fundamental input for implementing the above nonparametric shape-related hypothesis tests.

To make our hypothesis testing procedures precise, we employ the following feasible, Studentized statistic:

$$\dot{T}_p(x) = \frac{\widehat{\mu}^{(v)}(x)}{\sqrt{\widehat{\Omega}(x)/n}}, \qquad x \in \mathcal{X}, \quad 0 \leq v, s \leq p$$

leading to the hypothesis test:

$$\text{Reject } \dot{\mathsf{H}}_0 \qquad \text{if and only if} \qquad \sup_{x\in\mathcal{X}} \dot{T}_p(x) \geq \mathfrak{c}, \tag{3.8}$$

for an appropriate choice of critical value $\mathfrak{c}$ to control false rejections (Type I error). Of course, the other one-sided hypothesis tests are constructed in the obvious symmetric way.

Theorem III.5 (Hypothesis Testing: Nonparametric Shape Restriction)**.** *Let Assumption III.1 hold. For given p, and $0 \leq v, s \leq p$, set $J = J_{\texttt{IMSE}}$ and $\mathfrak{c} = \inf\big\{c \in \mathbb{R}_+ : \mathbb{P}\big[\sup_{x\in\mathcal{X}} \widehat{Z}_{p+q}(x) \leq c \,\big|\, \mathbf{D}\big] \geq 1-\alpha\big\}$.*

Under $\dot{\mathsf{H}}_0$, then

$$\lim_{n\to\infty} \mathbb{P}\Big[\sup_{x\in\mathcal{X}} \dot{T}_{p+q}(x) > \mathfrak{c}\Big] \leq \alpha.$$

Under $\dot{\mathsf{H}}_A$, then

$$\lim_{n\to\infty} \mathbb{P}\Big[\sup_{x\in\mathcal{X}} \dot{T}_{p+q}(x) > \mathfrak{c}\Big] = 1.$$

This theorem shows that the hypothesis testing procedure (3.8) is valid. Because of its one-sided nature, the test is conservative in general. Further, because it relies on a supremum-type statistic, this nonparametric shape restriction test also employs a simulated critical value, just like those used in the previous sections to construct confidence bands or to conduct parametric specification testing. Theorem III.5 corresponds to the one-sided "left" hypothesis test, but of course the analogous theorem "to the right" also holds. Our software implementation allows for all three possibilities: one-sided (left or right) and two-sided hypothesis testing. See Cattaneo, Crump, Farrell, and Feng (2019a) for more details.

Remark III.2 (Two-Sample Nonparametric Testing). Our results can also be extended to handle nonparametric testing about features of $\mu(x)$ for two or more groups. For example, assuming that two (sub-)samples are available, our methods can be used to test the null hypothesis: $\mathsf{H}_0 : \mu_1(x) = \mu_2(x)$ for all $x \in \mathcal{X}$, where $\mu_1(x)$ and $\mu_2(x)$ denote the $\mu(x)$ function in our framework for two distinct (sub-)samples. Such a hypothesis test can be formally implemented using a uniform measure of discrepancy, as we used above, or some other metric (see Remark III.1). Our theoretical results given in Appendix B are general enough to accomodate this extension, which we plan to undertake in upcoming work.

3.6 Conclusion

We introduced a general econometrics framework to understand binscatter, a very popular methodology for approximating the conditional expectation function in applied microeconomics. Our framework leads to a variety of new methodological (and theoretical) results for the canonical binscatter, including novel smooth and/or polynomial approximation approaches, principled covariate adjustment implementation, and valid inference and hypothesis testing methods. In particular, we highlight important problems with the way covariate adjustment is currently done in practice via current popular software implementations (Stepner, 2014).

In addition to providing the first foundational methodological study of binscatter and extensions thereof, we also offer new accompanying `Stata` and `R` software implementing all our results (Cattaneo, Crump, Farrell, and Feng, 2019a).

Chapter IV

Implementation Methodology and Numerical Evidence

4.1 Introduction

Chapter II and III provide an array of theoretical results for partitioning-based estimators in nonparametric and semiparametric models. In practice it is crucial for researchers to select tuning parameters and conduct corresponding estimation and inference in a principled way. As illustrated by our theory, one could select a tuning parameter that minimizes integrated mean squared error (IMSE), and then proceed to valid inference relying on various bias correction strategies proposed in Chapter II. This chapter aims at offering more detailed implementation methodology for this practice.

First, we discuss several popular basis choices, including splines, piecewise polynomials and wavelets, in Section 4.2. In particular, the expressions of their leading asymptotic errors are presented, which form the basis of our tuning parameter selection and bias correction methods. To show the usefulness of our theory, other high level conditions specified in Chapter II are verified as well.

Second, we derive a more explicit IMSE approximation for tensor-product partitions in Section 4.3. The result given in Theorem II.1 is stated at a very high level of generality, allowing for both tensor-product and non-tensor-product partitioning schemes. The *whole* partition plays the role of a tuning parameter, rendering the selection procedure still difficult to implement. In this chapter we will specialize it to a more detailed result for partitions formed via tensor products of intervals, where the tuning parameter reduces to a vector of partitioning knots. Under additional regularity conditions, we give explicit limiting constants in IMSE approximation which can be estimated consistently in practice. Similar results exist in literature only for some particular basis choices, e.g. splines (Agarwal and Studden, 1980) and piecewise polynomials (Cattaneo and Farrell, 2013). Our results are stated under a set of high-level conditions, covering these examples as special cases, and for compactly supported wavelets, our results appear to be new to literature.

Given the IMSE approximation and bias characterization, we propose in Section 4.4 two data-driven procedures (rule-of-thumb and direct plug-in) for tuning parameter selection, which are implemented in companion software packages `lspartition` in `R` and `binsreg` in `R` and `STATA`.

In Section 4.5 we illustrate our methods using an empirical example, offering a list of practical recommendations. Finally, Section 4.6 provides Monte Carlo evidence.

Notation

Throughout this chapter, we focus on the setup of Chapter II, and all notation used in the following is understood accordingly, if no special explanation. Moreover, we will assume the support of the regressors is of tensor product form, and then each dimension of $\mathcal{X}$ is partitioned marginally into intervals and Δ is the tensor product of these intervals. Let $\mathcal{X}_\ell = [\underline{x}_\ell, \overline{x}_\ell]$ be the support of covariate $\ell = 1, 2, \dots, d$ and partition this into κ_ℓ disjoint subintervals defined by $\{\underline{x}_\ell = t_{\ell,0} < t_{\ell,1} < \cdots < t_{\ell,\kappa_\ell - 1} < t_{\ell,\kappa_\ell} = \overline{x}_\ell\}$. If this partition of $\mathcal{X}_\ell$ is Δ_ℓ, then a complete partition of $\mathcal{X}$ can be formed by tensor products of the one-dimensional partitions: $\Delta = \otimes_{\ell=1}^{d} \Delta_\ell$, with $\boldsymbol{\kappa} = (\kappa_1, \kappa_2, \dots, \kappa_d)'$ subintervals in each dimension of $\mathbf{x}_i$ and $\bar{\kappa} = \kappa_1 \kappa_1 \cdots \kappa_d$. A generic cell of this partition is a rectangle

$$\delta_{l_1 \dots l_d} = \{\mathbf{x} : t_{\ell,l_\ell} < x_\ell < t_{\ell,l_\ell+1}, \quad 0 \le l_\ell \le \kappa_\ell - 1 \quad \text{and} \quad 1 \le \ell \le d\}. \tag{4.1}$$

Given this setup, Assumption II.2 can be verified by choosing the knot positions/configuration appropriately, often dividing $\mathcal{X}_\ell$ uniformly or by empirical quantiles.

Moreover, let $\delta_{\mathbf{x}}$ be a subrectangle in Δ containing $\mathbf{x}$, and $\mathbf{b}_{\mathbf{x}}$ be the vector collecting the interval lengths of $\delta_{\mathbf{x}}$ (see Assumption IV.1 below). $\mathbf{t}_{\mathbf{x}}^{L}$ denotes the start point of $\delta_{\mathbf{x}}$. For $\ell = 1, \dots, d$, for a generic cell $\delta_{l_1 \dots l_d}$ as in (4.1), we write $b_{\ell,l_\ell} = t_{\ell,l_\ell+1} - t_{\ell,l_\ell}$, $l_\ell = 0, \dots, \kappa_\ell - 1$, $b_\ell = \max_{0 \le l_\ell \le \kappa_\ell - 1} b_{\ell,l_\ell}$ and $b = \max_{1 \le \ell \le d} b_\ell$ ($b \asymp h$ by Assumption II.2). Finally, $\oslash$ denotes the entrywise division operator (Hadamard division), and $\mathbf{q}! = q_1! \cdots q_d!$

4.2 Several Basis Choices

This section discusses several examples of basis expansions that are commonly used in practice, including B-splines, wavelets and piecewise polynomials (i.e., generalized regressogram). We illustrate how Assumptions II.3, II.4, II.5 and the orthogonality condition (2.5) are verified for these basis choices. In particular, their leading approximation errors are presented.

4.2.1 B-Splines

A univariate spline is a piecewise polynomial satisfying certain smoothness constraints. For some integer $m_\ell \geq 2$, let $\mathcal{S}_{\Delta_\ell, m_\ell}$ be the set of splines of order m_ℓ with univariate partition Δ_ℓ. Then

$$\mathcal{S}_{\Delta_\ell,m_\ell} = \Big\{ s \in \mathcal{C}^{m_\ell - 2}(\mathcal{X}_\ell) \colon s(x) \text{ is a polynomial of degree } (m_\ell - 1) \\ \text{on each subinterval } [t_{\ell,l_\ell}, t_{\ell,l_\ell+1}] \Big\},$$

and hence $\mathcal{S}_{\Delta_\ell,m_\ell}$ is a vector space and can be spanned by many equivalent representing bases. B-splines as a local basis are well studied in literature and enjoy many nice properties. The detailed construction of B-splines can be found in many textbooks, e.g., Schumaker (2007), and is omitted here to conserve space.

We focus on tensor-product polynomial splines of order $\mathbf{m} = (m_1, \ldots, m_d)$ with partition Δ, which are formed by a tensor product of univariate B-splines:

$$\mathcal{S}_{\Delta,\mathbf{m}} = \otimes_{\ell=1}^{d} \mathcal{S}_{\Delta_\ell,m_\ell} = \operatorname{span}\{p_{l_1}(x_1)p_{l_2}(x_2)\cdots p_{l_d}(x_d)\}_{l_1=1,\ldots,l_d=1}^{K_1,\ldots,K_d}.$$

Each collection of basis functions $\{p_{l_\ell}(x_\ell)\}_{l_\ell=1}^{K_\ell}$ forms the univariate B-spline basis of order m_ℓ for dimension ℓ, $\ell = 1, \ldots d$. We have a total of $K = \prod_{\ell=1}^{d} K_\ell$ basis functions. The order of univariate basis could vary across dimensions, but for simplicity we assume that $m_1 = \cdots = m_d = m$.

The following lemma shows that Assumptions II.3 and II.4 hold for B-splines.

Lemma IV.1 (B-Splines Estimators)**.** *Let* $\mathbf{p}(\mathbf{x})$ *be a tensor-product* B*-Spline basis of order* m*, and suppose Assumptions II.1 and II.2 hold with* $m \leq S$*. Then:*

1. $\mathbf{p}(\mathbf{x})$ *satisfies Assumption II.3.*

2. *If, in addition,*

$$\max_{0 \leq l_\ell \leq \kappa_\ell - 2} |b_{\ell,l_\ell+1} - b_{\ell,l_\ell}| = O(b^{1+\varrho}), \qquad \ell = 1, \ldots, d, \tag{4.2}$$

then Assumption II.4 holds with $\varsigma = m - 1$ *and*

$$\mathscr{B}_{m,\boldsymbol{\varsigma}}(\mathbf{x}) = -\sum_{\boldsymbol{u}\in\Lambda_m} \frac{\partial^{\boldsymbol{u}}\theta(\mathbf{x})h_{\mathbf{x}}^{m-[\boldsymbol{\varsigma}]}}{(\boldsymbol{u}-\boldsymbol{\varsigma})!} \frac{\mathbf{b}_{\mathbf{x}}^{\boldsymbol{u}-\boldsymbol{\varsigma}}}{h_{\mathbf{x}}^{m-[\boldsymbol{\varsigma}]}} B_{\boldsymbol{u}-\boldsymbol{\varsigma}}^{\mathsf{S}}\Big((\mathbf{x}-\mathbf{t}_{\mathbf{x}}^{L}) \oslash \mathbf{b}_{\mathbf{x}}\Big)$$

where $\Lambda_m = \{\boldsymbol{u} \in \mathbb{Z}_+^d : [\boldsymbol{u}] = m,$ *and* $u_\ell = m$ *for some* $1 \leq \ell \leq d\}$ *and* $B_{\boldsymbol{u}}^{\mathsf{S}}(\mathbf{x})$ *is the product of univariate Bernoulli polynomials; that is,* $B_{\boldsymbol{u}}^{\mathsf{S}}(\mathbf{x}) := \prod_{\ell=1}^{d} B_{u_\ell}(x_\ell)$ *with* $B_{u_\ell}(\cdot)$ *being the* u_ℓ*-th Bernoulli polynomial and* $B_{\boldsymbol{u}}^{\mathsf{S}}(\cdot) = 0$ *if* $\boldsymbol{u}$ *contains negative elements. Furthermore, Equation* (2.5) *holds.*

3. *Let* $\tilde{\mathbf{p}}(\mathbf{x})$ *be a tensor-product B-Spline basis of order* $\tilde{m} > m$ *on the same partition* Δ*, and assume* $\tilde{m} \leq S$*. Then Assumption II.5 is satisfied.*

Equation (4.2) gives a precise definition of the strong quasi-uniform condition on the partition scheme. Assumption II.2 requires that the volumes of all cells vanish at the same rate but allows for any constant proportionality between neighboring cells. Presently, cells are further restricted to be of the same volume asymptotically, and further, a specific rate is required that is related to the smoothness of $\theta(\cdot)$. Note that, for example, equally spaced knots satisfy this conditional trivially. For other schemes, additional work may be needed. Under (4.2), Barrow and Smith (1978) obtained an expression for the leading asymptotic error of univariate splines, which was later used by Zhou, Shen, and Wolfe (1998) and Zhou and Wolfe (2000), among others. Lemma IV.1 extends previous results to the multi-dimensional case, in addition to showing that the high-level conditions in Assumption II.3 and II.4 hold for B-Splines.

4.2.2 Wavelets

Our results apply to compactly supported wavelets, such as Cohen-Daubechies-Vial wavelets (Cohen, Daubechies, and Vial, 1993). For more background details see Meyer (1995); Härdle, Kerkyacharian, Picard, and Tsybakov (2012); Chui (2016), and references therein.

Specifically, let $\{\phi_{s_n,l}, l \in \mathcal{L}_{s_n}\}$ be a collection of (compactly supported) father wavelet functions at resolution level s_n, where $\mathcal{L}_{s_n}$ denotes some properly defined index set. We employ a tensor-product (father) wavelet basis

$$\mathbf{p}(\mathbf{x}) := \otimes_{\ell=1}^{d} 2^{-s_n/2} \boldsymbol{\phi}_{s_n}(x_\ell) \tag{4.3}$$

where $\boldsymbol{\phi}_{s_n}$ is a vector containing all functions in $\{\phi_{s_n,l}, l \in \mathcal{L}_{s_n}\}$. The resolution level s_n plays the role of the tuning parameter for wavelets. As $n \to \infty$, $s_n \to \infty$, and it is linked with the mesh width by $b = 2^{-s_n}$.

As the next lemma shows, this large class of orthogonal wavelet bases satisfies our assumptions.

Lemma IV.2 (Wavelets Estimators)**.** *Let ϕ and ψ be a scaling function and a wavelet function of degree $m-1$ with $q+1$ continuous derivatives, $\mathbf{p}(\mathbf{x})$ be the tensor product orthogonal (father) wavelet basis of degree $m-1$ generated by ϕ, and suppose Assumption II.1 holds with $m \leq S$.*

1. *$\mathbf{p}(\mathbf{x})$ satisfies Assumption II.3.*

2. *Assumption II.4 holds with $\varsigma = q$ and*

$$\mathscr{B}_{m,\boldsymbol{\varsigma}}(\mathbf{x}) = -\sum_{\boldsymbol{u}\in\Lambda_m} \frac{\partial^{\boldsymbol{u}}\theta(\mathbf{x})h^{m-[\varsigma]}}{\boldsymbol{u}!}\frac{b^{m-[\varsigma]}}{h^{m-[\varsigma]}} B^{\mathtt{W}}_{\boldsymbol{u},\boldsymbol{\varsigma}}(\mathbf{x}/b)$$

where $\Lambda_m = \{\boldsymbol{u} \in \mathbb{Z}^d_+ : [\boldsymbol{u}] = m,$ and $u_\ell = m$ for some $1 \leq \ell \leq d\}$ and $B^{\mathtt{W}}_{\boldsymbol{u},\boldsymbol{\varsigma}}(\mathbf{x}) = \sum_{s\geq 0} \partial^{\varsigma}\xi_{\boldsymbol{u},s}(\mathbf{x})$ converges uniformly, with $\xi_{\boldsymbol{u},s}(\cdot)$ being a linear combination of products of univariate father wavelet ϕ and the mother wavelet ψ; its exact form is notationally cumbersome and is given in Equation (C.3)*. Furthermore, Equation* (2.5) *holds.*

3. *Let $\tilde{\phi}$ be a scaling function of degree $\tilde{m}-1$ with $m+1$ continuous derivatives for some $\tilde{m} > m$, $\tilde{\mathbf{p}}(\mathbf{x})$ be the tensor-product orthogonal wavelet basis generated by $\tilde{\phi}$ having the same resolution level as $\mathbf{p}(\mathbf{x})$, and assume $\tilde{m} \leq S$. Then Assumption II.5 is satisfied.*

In addition to verifying our high-level conditions for wavelets, this result gives a novel asymptotic error expansion for multidimensional compactly supported wavelets. Our derivation employs the ideas in Sweldens and Piessens (1994) and exploits the tensor-product structure of these bases.

4.2.3 Generalized Regressograms

The generalized regressogram is distinguished from splines in that (i) each polynomial is supported on exactly one cell, and relatedly (ii) no continuity is assumed between cells. Specifically, for some fixed integer $m \in \mathbb{Z}_+$, let $\mathbf{r}(x_\ell) = (1, x_\ell, \ldots, x_\ell^{m-1})'$ denote a vector of powers up to degree $m-1$. To extend it to a multidimensional basis, we take the tensor product of $\mathbf{r}(x_\ell)$, denoted by a column vector $\mathbf{R}(\mathbf{x})$. The total order of such a basis is not fixed, and its behavior is more similar to tensor-product B-splines. Following Cattaneo and Farrell (2013, and references therein), we exclude all terms with degree greater than $m-1$ in $\mathbf{R}(\mathbf{x})$. Hence the remaining elements in $\mathbf{R}(\mathbf{x})$ are given by $\mathbf{x}^{\boldsymbol{\alpha}} = x_1^{\alpha_1}\cdots x_d^{\alpha_d}$ for a unique d-tuple $\boldsymbol{\alpha}$ such that $[\boldsymbol{\alpha}] \leq m-1$. Then we "localize" this basis by restricting it to a particular subrectangle $\delta_{l_1\dots l_d}$. Specifically, we write $\mathbf{p}_{l_1\dots l_d}(\mathbf{x}) = \mathbb{1}_{\delta_{l_1\dots l_d}}(\mathbf{x})\mathbf{R}(\mathbf{x})$, where $\mathbb{1}_{\delta_{l_1\dots l_d}}(\mathbf{x})$ is equal to 1 if $\mathbf{x} \in \delta_{l_1\dots l_d}$ and 0 otherwise. Finally, we rotate the basis by centering each basis function at the start point of the corresponding cell and scale it by interval lengths.

The following lemma shows that Assumptions II.3 and II.4 hold for generalized regressograms.

Lemma IV.3 (Generalized Regressograms)**.** *Let* $\mathbf{p}(\mathbf{x})$ *be the rotated piecewise polynomial basis of degree* $m-1$ *based on Legendre polynomials, and suppose that Assumptions II.1 and II.2 hold with* $m \leq S$. *Then,*

1. $\mathbf{p}(\mathbf{x})$ *satisfies Assumption II.3.*

2. *Assumption II.4 holds with* $\varsigma = m-1$ *and*

$$\mathscr{B}_{m,\varsigma}(\mathbf{x}) = -\sum_{\boldsymbol{u}\in\Lambda_m} \frac{\partial^{\boldsymbol{u}}\theta(\mathbf{x})h_{\mathbf{x}}^{m-[\boldsymbol{\varsigma}]}}{(\boldsymbol{u}-\boldsymbol{\varsigma})!}\frac{\mathbf{b}_{\mathbf{x}}^{\boldsymbol{u}-\boldsymbol{\varsigma}}}{h_{\mathbf{x}}^{m-[\boldsymbol{\varsigma}]}} B^{\mathsf{P}}_{\boldsymbol{u}-\boldsymbol{\varsigma}}\Big((\mathbf{x}-\mathbf{t}^{L}_{\mathbf{x}})\oslash\mathbf{b}_{\mathbf{x}}\Big),$$

where $\Lambda_m = \{\boldsymbol{u} : [\boldsymbol{u}] = m\}$ *and*

$$B^{\mathsf{P}}_{\boldsymbol{u}}(\mathbf{x}) := \prod_{\ell=1}^{d} \binom{2u_\ell}{u_\ell}^{-1} P_{u_\ell}(x_\ell),$$

with $P_{u_\ell}(\cdot)$ *being the* u_ℓ*-th shifted Legendre polynomial orthogonal on* $[0,1]$*, and* $B^{\mathsf{P}}_{\boldsymbol{u}}(\cdot) = 0$ *if* $\boldsymbol{u}$ *contains negative elements. Furthermore, Equation* (2.5) *holds.*

3. *Let* $\tilde{\mathbf{p}}(\mathbf{x})$ *be a piecewise polynomial basis of degree* $\tilde{m}-1$ *on the same partition* Δ *for some* $\tilde{m} > m$*, and assume* $\tilde{m} \leq S$*. Then Assumption II.5 is satisfied.*

The leading asymptotic error obtained in Lemma IV.3 differs from the one in Cattaneo and Farrell (2013) because it is expressed in terms of orthogonal polynomials. Here we employ Legendre polynomials, $\bar{P}_m(x)$, that are orthogonal with respect to the Lebesgue measure on $[-1,1]$, and then shift them to $P_m(x) = \bar{P}_m(2x-1)$. Thus the shifted Legendre polynomials are orthogonal on $[0,1]$.

4.3 IMSE: Tensor-Product Partitions

Theorem II.1 gives a general IMSE approximation where leading constants rely on the whole partition Δ. To illustrate the usefulness of this result in applications, we study the special case of a tensor-product partition where the tuning parameter Δ reduces to a vector of partitioning knots $\boldsymbol{\kappa} = (\kappa_1, \dots, \kappa_d)'$, where κ_ℓ is the number of subintervals used for the ℓ-th covariate. We further assume that Δ and $\mathbf{p}(\cdot)$ obey the following regularity conditions, so that the limiting constants in the IMSE approximation can be characterized.

Assumption IV.1 (Regularity for Asymptotic IMSE). *Suppose that $\mathcal{X} = \otimes_{\ell=1}^{d} \mathcal{X}_\ell \subset \mathbb{R}^d$, which is normalized to $[0,1]^d$ without loss of generality, and Δ is a tensor-product partition. For $\mathbf{x} \in [0,1]^d$, denote $\delta_{\mathbf{x}} = \{t_{\ell,l_{\mathbf{x}}} \leq x_\ell \leq t_{\ell,l_{\mathbf{x}}+1}, 1 \leq \ell \leq d\}$, where $l_{\mathbf{x}} < \kappa_\ell$. Let $\mathbf{b}_{\mathbf{x}} = (b_{\mathbf{x},1}, \dots, b_{\mathbf{x},d})$ collect the interval lengths $b_{\mathbf{x},\ell} = |t_{\ell,l_{\mathbf{x}}+1} - t_{\ell,l_{\mathbf{x}}}|$. In addition:*

1. *For $\ell = 1, \dots, d$, $\sup_{\mathbf{x}\in[0,1]^d} |b_{\mathbf{x},\ell} - \kappa_\ell^{-1} g_\ell(\mathbf{x})^{-1}| = o(\kappa_\ell^{-1})$, where $g_\ell(\cdot)$ is bounded away from zero and continuous.*

2. *For all $\delta \in \Delta$ and $\boldsymbol{u}_1, \boldsymbol{u}_2 \in \Lambda_m$, there exist constants $\eta_{\boldsymbol{u}_1,\boldsymbol{u}_2,\mathbf{q}}$ such that*

$$\int_\delta \frac{h_{\mathbf{x}}^{2m-2[\mathbf{q}]}}{\mathbf{b}_{\mathbf{x}}^{\boldsymbol{u}_1+\boldsymbol{u}_2-2\mathbf{q}}} B_{\boldsymbol{u}_1,\mathbf{q}}(\mathbf{x}) B_{\boldsymbol{u}_2,\mathbf{q}}(\mathbf{x})\, d\mathbf{x} = \eta_{\boldsymbol{u}_1,\boldsymbol{u}_2,\mathbf{q}} \operatorname{vol}(\delta)$$

   where $\operatorname{vol}(\delta)$ denotes the volume of δ.

3. *There exists a set of points $\{\boldsymbol{\tau}_k\}_{k=1}^K$ such that $\boldsymbol{\tau}_k \in \operatorname{supp}(p_k(\cdot))$ for each $k = 1, \dots, K$, and $\{\boldsymbol{\tau}_k\}_{k=1}^K$ can be assigned into $J + \breve{J} < \infty$ groups such that $\{\boldsymbol{\tau}_{s,k_s}\}_{k_s=1}^{K_s}$, $s = 1, \dots, J + \breve{J}$, $\sum_{s=1}^{J+\breve{J}} K_s = K$, and the following conditions hold: (i) For all $1 \leq s \leq J$, $\{\delta_{\boldsymbol{\tau}_{s,k_s}}\}_{k_s=1}^{K_s}$ are pairwise disjoint and $\operatorname{vol}\left([0,1]^d \setminus \bigcup_{k_s=1}^{K_s} \delta_{\boldsymbol{\tau}_{s,k_s}}\right) = o(1)$; and (ii) for all $J + 1 \leq s \leq J + \breve{J}$, $\operatorname{vol}\left(\bigcup_{k_s=1}^{K_s} \delta_{\boldsymbol{\tau}_{s,k_s}}\right) = o(1)$.*

Part (1) slightly strengthens the quasi-uniform condition imposed in Assumption II.2, but allows for quite general transformations of the knot location. Part (2) ensures that the *local* integral of the product between any two $B_{\boldsymbol{u},\mathbf{q}}(\cdot)$ for $\boldsymbol{u} \in \Lambda_m$, which depend on the basis but not $\theta(\mathbf{x})$, is proportional to the volume of the cell. The scaling factor is due to the use of the lengths of intervals on each axis (denoted by $\mathbf{b}_{\mathbf{x}}$) to characterize the approximation error for a tensor-product partition, instead of the more general diameter used in Section 2.2. Finally, part (3) describes how the supports of the basis functions cover the whole support of data. Specifically, it requires that the approximating basis $\mathbf{p}$ can be divided into $J + \breve{J}$ groups. The supports of functions in each of the first J groups constitute "almost" complete covers of $\mathcal{X}$. In contrast, the supports of functions in other groups are negligible in terms of volume. In such a case, we refer to J as the number of complete covers generated by the supports of basis functions. For tensor product B-splines (with simple knots) and wavelets, each subrectangle in Δ can be associated with one basis function in $\mathbf{p}$ and the supports of the remaining functions are asymptotically negligible in terms of volume. Thus, $J = 1$ in these two examples. For piecewise polynomials of total order m, within each subrectangle the unknown function is approximated by a multivariate polynomial of degree $m-1$, and thus $J = \binom{d+m-1}{m-1}$. This condition is used to ensure that the summation over the number of basis functions converges to a well-defined integral as $K \asymp h^{-d} \to \infty$.

We then have the following result for $\widehat{\theta}_0(\mathbf{x})$.

Theorem IV.1 (Asymptotic IMSE)**.** *Suppose that the conditions in Theorem II.1 and Assumption IV.1 hold. Then, for* $[\mathbf{q}] = 0$,

$$\mathscr{V}_{\boldsymbol{\kappa},\mathbf{0}} = \Big(\prod_{\ell=1}^{d} \kappa_\ell\Big)\mathscr{V}_{\mathbf{0}} \ + o(h^{-d}), \qquad \mathscr{V}_{\mathbf{0}} = J\int_{[0,1]^d} \frac{\sigma^2(\mathbf{x})}{f(\mathbf{x})}\Big(\prod_{\ell=1}^{d} g_\ell(\mathbf{x})\Big) w(\mathbf{x})\, d\mathbf{x},$$

and, provided that (2.5) *holds,*

$$\mathscr{B}_{\boldsymbol{\kappa},\mathbf{0}} = \sum_{\boldsymbol{u}_1,\boldsymbol{u}_2\in\Lambda_m} \boldsymbol{\kappa}^{-(\boldsymbol{u}_1+\boldsymbol{u}_2)}\mathscr{B}_{\boldsymbol{u}_1,\boldsymbol{u}_2,\mathbf{0}} + o(h^{2m}),$$

$$\mathscr{B}_{\boldsymbol{u}_1,\boldsymbol{u}_2,\mathbf{0}} = \eta_{\boldsymbol{u}_1,\boldsymbol{u}_2,\mathbf{0}} \int_{[0,1]^d} \frac{\partial^{\boldsymbol{u}_1}\theta(\mathbf{x})\partial^{\boldsymbol{u}_2}\theta(\mathbf{x})}{\mathbf{g}(\mathbf{x})^{\boldsymbol{u}_1+\boldsymbol{u}_2}} w(\mathbf{x})d\mathbf{x}.$$

The bias approximation requires the approximate orthogonality condition (2.5) which is satisfied by B-splines, wavelets, and piecewise polynomials. It appears to be an open question whether $\mathscr{V}_{\boldsymbol{\kappa},\mathbf{q}}$ and $\mathscr{B}_{\boldsymbol{\kappa},\mathbf{q}}$ converge to a well-defined limit when general basis functions are considered. Cattaneo and Farrell (2013) showed convergence to well defined limits for piecewise polynomials, but their result is not easy to extend to cover other basis functions without imposing $\mathbf{q} = \mathbf{0}$ and the approximate orthogonality condition (2.5). This is why Theorem IV.1 only considers $\mathbf{q} = \mathbf{0}$ (i.e., the IMSE of $\widehat{\theta}_0(\mathbf{x})$) and imposes condition (2.5).

Theorem IV.1 justifies the IMSE-optimal choice of number of knots:

$$\boldsymbol{\kappa}_{\mathtt{IMSE},\mathbf{0}} = \operatorname*{arg\,min}_{\boldsymbol{\kappa}\in\mathbb{Z}^d_{++}} \left\{ \frac{1}{n}\Big(\prod_{\ell=1}^{d}\kappa_\ell\Big)\mathscr{V}_{\mathbf{0}} + \sum_{\boldsymbol{u}_1,\boldsymbol{u}_2\in\Lambda_m} \boldsymbol{\kappa}^{-(\boldsymbol{u}_1+\boldsymbol{u}_2)}\mathscr{B}_{\boldsymbol{u}_1,\boldsymbol{u}_2,\mathbf{0}} \right\},$$

and, in particular, when the same number of knots is used in all margins,

$$\kappa_{\mathtt{IMSE},\mathbf{0}} = \left\lceil \left(\frac{2m\sum_{\boldsymbol{u}_1,\boldsymbol{u}_2\in\Lambda_m}\mathscr{B}_{\boldsymbol{u}_1,\boldsymbol{u}_2,\mathbf{0}}}{d\mathscr{V}_{\mathbf{0}}} \right)^{\frac{1}{2m+d}} n^{\frac{1}{2m+d}} \right\rceil$$

Data-driven versions of this IMSE-optimal choice, and extensions to derivative estimation, are discussed in Section 4.4 and fully implemented in our companion general-purpose R package `lspartition` (Cattaneo, Farrell, and Feng, 2018b).

4.4 Tuning Parameter Selection

In this section, we discuss implementation details about choosing the IMSE-optimal tuning parameters. We restrict our attention to tensor-product partitions with the same number of knots used in every dimension. Thus the tuning parameter reduces to a scalar κ which denotes the number of subintervals used in every dimension. We offer two approaches: rule-of-thumb (ROT) and direct plug-in (DPI).

4.4.1 Rule-of-Thumb Choice

The rule-of-thumb choice is based on the special case considered in Theorem IV.1. Specifically, we assume $\mathbf{q} = \mathbf{0}$ and knots are evenly spaced. The implementation steps are summarized as follows.

- *Preliminary regression.* Implement a preliminary regression using a global polynomial of degree $(m+2)$, and denote this estimate of $\theta(\cdot)$ by $\hat{\theta}_{\texttt{pre}}(\cdot)$.

- *Bias constant.* Let the weighting function $w(\mathbf{x})$ be the density function of $\mathbf{x}_i$. Use the preliminary regression coefficients to obtain an estimate of the mth derivatives of $\theta(\cdot)$, i.e., $\widehat{\partial^{\boldsymbol{u}}\theta}(\cdot) = \partial^{\boldsymbol{u}}\widehat{\theta}_{\texttt{pre}}(\cdot)$, for each $\boldsymbol{u} \in \Lambda_m$. Then an estimate of the bias constant is

$$\widehat{\mathscr{B}}_{\boldsymbol{u}_1,\boldsymbol{u}_2,\mathbf{0}} = \eta_{\boldsymbol{u}_1,\boldsymbol{u}_2,\mathbf{0}} \times \frac{1}{n}\sum_{i=1}^{n} \widehat{\partial^{\boldsymbol{u}_1}\theta}(\mathbf{x}_i)\widehat{\partial^{\boldsymbol{u}_2}\theta}(\mathbf{x}_i).$$

- *Variance constant.* Implement another regression of y_i^2 on $\mathbf{x}_i$ using global polynomials of degree $(m+2)$, and thus an estimate of $\mathbb{E}[y_i^2|\mathbf{x}_i = \mathbf{x}]$ is obtained. Combining it with $\hat{\theta}_{\texttt{pre}}(\cdot)$, we obtain an estimate of conditional variance function, denoted by $\hat{\sigma}^2(\cdot)$, since $\sigma^2(\mathbf{x}) = \mathbb{E}[y_i^2|\mathbf{x}_i = \mathbf{x}] - (\mathbb{E}[y_i|\mathbf{x}_i = \mathbf{x}])^2$. Then an estimate of the variance constant is

$$\widehat{\mathscr{V}}_{\mathbf{0}} = \begin{cases} \frac{1}{n}\sum_{i=1}^{n}\hat{\sigma}^2(\mathbf{x}_i) & \text{for splines and wavelets,} \\ \binom{d+m-1}{m-1} \times \frac{1}{n}\sum_{i=1}^{n}\hat{\sigma}^2(\mathbf{x}_i) & \text{for piecewise polynomial.} \end{cases}$$

- *Rule-of-thumb* $\hat{\kappa}_{\texttt{ROT}}$. Using the above results, a simple rule-of-thumb choice of κ is

$$\hat{\kappa}_{\texttt{ROT}} = \left\lceil \left(\frac{2(m-[\mathbf{q}]) \sum_{\boldsymbol{u}_1,\boldsymbol{u}_2 \in \Lambda_m} \widehat{\mathscr{B}}_{\boldsymbol{u}_1,\boldsymbol{u}_2,\mathbf{0}}}{(d+2[\mathbf{q}])\widehat{\mathscr{V}}_{\mathbf{0}}} \right)^{\frac{1}{2m+d}} n^{\frac{1}{2m+d}} \right\rceil.$$

  Clearly, this choice of κ is derived based on many strong assumptions, but it still has the correct rate ($\asymp n^{\frac{1}{2m+d}}$) in other cases.

Remark IV.1. The above procedure assumes $\mathbf{q} = \mathbf{0}$, since a general limiting variance constant similar to that given in Theorem IV.1 is still unavailable for other cases. However, for piecewise polynomials, limiting variance and bias constants are available for any m and $\mathbf{q}$ given a partition scheme studied in Theorem IV.1. See Cattaneo and Farrell (2013) and Cattaneo, Crump, Farrell, and Feng (2019b) for more discussion.

4.4.2 Direct Plug-in Choice

Assume that the weighting function $w(\mathbf{x})$ is equal to the density function of $\mathbf{x}_i$. We propose a direct-plug-in (DPI) procedure summarized in the following steps.

- *Preliminary choice of* κ: Implement the rule-of-thumb procedure to obtain $\hat{\kappa}_{\texttt{ROT}}$.

- *Preliminary regression.* Given the user-specified basis (splines, wavelets, or piecewise polynomials), knot placement scheme ("uniform" or "quantile") and rule-of-thumb choice $\hat{\kappa}_{\texttt{ROT}}$, implement a series regression of order $(m+1)$ to obtain derivative estimates for every $\boldsymbol{u} \in \Lambda_m$. Denote this preliminary estimate by $\widehat{\partial^{\boldsymbol{u}}\theta}_{\texttt{pre}}(\cdot)$.

- *Bias constant.* Construct an estimate $\widehat{\mathscr{B}}_{m,\mathbf{q}}(\cdot)$ of the leading error $\mathscr{B}_{m,\mathbf{q}}(\cdot)$ simply by replacing $\partial^{\boldsymbol{u}}\theta(\cdot)$ by $\widehat{\partial^{\boldsymbol{u}}\theta}_{\texttt{pre}}(\cdot)$. $\widehat{\mathscr{B}}_{m,\mathbf{0}}(\cdot)$ can be obtained similarly. Then we use the pre-asymptotic version of conditional bias to estimate the bias constant:

$$\widehat{\mathscr{B}}_{\kappa,\mathbf{q}} = \frac{1}{n}\sum_{i=1}^{n} \left(\widehat{\mathscr{B}}_{m,\mathbf{q}}(\mathbf{x}_i) - \widehat{\boldsymbol{\gamma}}_{\mathbf{q},0}(\mathbf{x})'\mathbb{E}_n[\boldsymbol{\Pi}_0(\mathbf{x}_i)\widehat{\mathscr{B}}_{m,\mathbf{0}}(\mathbf{x}_i)] \right)^2.$$

- *Variance constant.* Implement a series regression of order m using $\hat{\kappa}_{\texttt{rot}}$, and then use the pre-asymptotic version of conditional variance to obtain an estimate of variance constant. Specifically, we have

$$\widehat{\mathscr{V}}_{\kappa,\mathbf{q}} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\boldsymbol{\gamma}}_{\mathbf{q},0}(\mathbf{x})'\widehat{\boldsymbol{\Sigma}}_0\widehat{\boldsymbol{\gamma}}_{\mathbf{q},0}(\mathbf{x}), \quad \widehat{\boldsymbol{\Sigma}}_0 = \mathbb{E}_n[\boldsymbol{\Pi}_0(\mathbf{x}_i)\boldsymbol{\Pi}_0(\mathbf{x}_i)'\widehat{\varepsilon}^2_{i,0}]$$

where $\widehat{\varepsilon}_{i,0}$'s are regression residuals. Different weighting schemes for residuals may be used, leading to various "heteroskedasticity-consistent" variance estimates.

- *Direct plug-in $\hat{\kappa}$.* Collecting all these results, a direct plug-in choice of κ is

$$\hat{\kappa}_{\texttt{DPI}} = \left\lceil \left(\frac{2(m-[\mathbf{q}])\kappa_{\texttt{ROT}}^{2(m-[\mathbf{q}])}\widehat{\mathscr{B}}_{\kappa,\mathbf{q}}}{(d+2[\mathbf{q}])\kappa_{\texttt{ROT}}^{-(d+2[\mathbf{q}])}\widehat{\mathscr{V}}_{\kappa,\mathbf{q}}} \right)^{\frac{1}{2m+d}} n^{\frac{1}{2m+d}} \right\rceil.$$

Remark IV.2. The above DPI procedure relies on an explicit expression of leading asymptotic bias. Note that in Chapter III we discuss binscatter estimation with smoothness restrictions, which can be viewed as least squares estimators based on B-splines with knots of certain multiplicities (see Schumaker, 2007, Definition 4.1). The bias representation for B-splines with simple knots given in Lemma IV.1 does not immediately apply to other cases. Nevertheless, a spline basis with simple knots leads to the smoothest estimate of the regression function, given the order of basis. Removing some smoothness restrictions only enlarges the underlying approximation space. Hence the particular spline function resulting in the leading error given in Lemma IV.1 still forms a valid L_∞ approximation of the regression function, though it is not the best one. From the theoretical perspective, the main cost of employing this sub-optimal L_∞-approximation is that the orthogonality condition (2.5) fails. However, as emphasized previously, our IMSE approximation given in Theorem II.1 *does not* rely on (2.5). Thus, the suggested DPI procedure is still feasible for piecewise polynomials with any smoothness restrictions, provided that they do not degenerate to global polynomials. This approach has been applied to binning selection for binscatter regressions in the companion software package `Binsreg`.

4.5 Empirical Example

In this section we illustrate our methods using an empirical example, with a list of recommendations offered for practitioners. We will focus on generalized binscatter methods discussed in Chapter III, which relies on piecewise polynomial and B-spline bases. Recall that in this context J denotes the number of bins and $\mu(\cdot)$ is the function to be estimated (mean relationship between y and x with $\mathbf{w}$ controlled for). p, s and v denote the degree of polynomial, the number of smoothness restrictions and the derivative order respectively.

The real dataset is obtained from the American Community Survey (ACS) using the 5-year survey estimates (2013-2017). All analyses are performed at the zip code tabulation area level for the United States (excluding Puerto Rico), with the data downloaded from the Census Bureau website: `https://factfinder.census.gov/faces/nav/jsf/pages/programs.xhtml?program=acs`.

The outcome variable is the Gini index, and the independent variable of interest is median household income (dollars in thousands). Such relationship will reflect how inequality varies across regions of different income levels.

For graphical presentation we recommend the following:

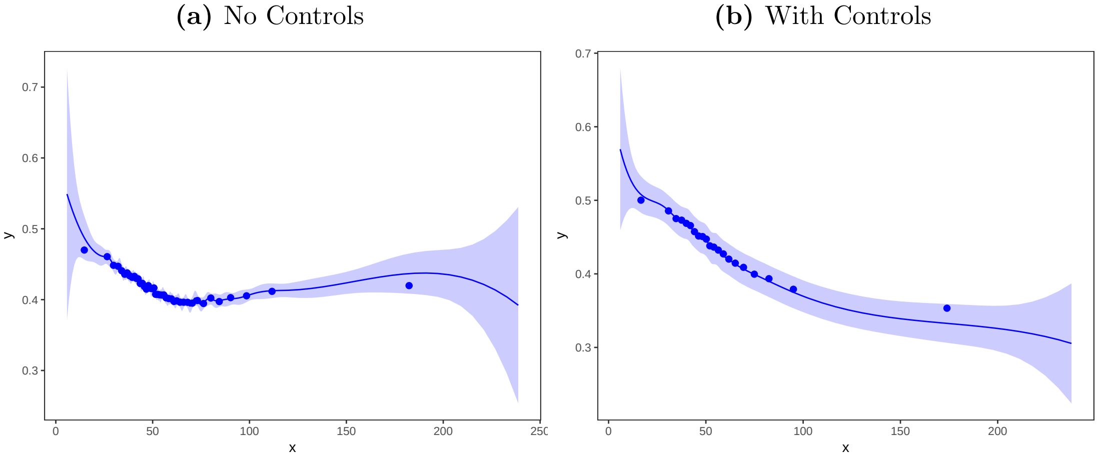
Step 1. Use IMSE-optimal canonical binscatter to depict data, with covariate adjustment and accounting for clustered data as appropriate. Formally, set $p = s = v = 0$ and $J = J_{\texttt{IMSE}}$ (Theorem II.1), and then plot $\widehat{\mu}(\bar{b}_j)$ as "dots", where $\bar{b}_j$ denotes the center of the j-th quantile-spaced bin (Section 3.5), $j = 1, 2, \dots, J$.

Step 2. On the same quantile-spaced partition determined by $J_{\texttt{IMSE}}$ in Step 1, construct a cubic B-spline for flexible approximation of $\mu(x)$, with covariate adjustment and accounting for clustered data as appropriate. Formally, set $p = 3$, $s = 3$ and $v = 0$, and then plot $\widehat{\mu}(x)$ as a solid line.

Step 3. Under the baseline configuration in Steps 1 and 2, confidence bands can be constructed on the same quantile-spaced partitioning and for the same cubic B-spline choices using the simulated quantiles (Lemma III.2). These bands can be plotted directly on top of the "dots" from Step 1 and the solid line from Step 2.

This approach is illustrated in Figure IV.1. When we apply controls, the control variables are (i) percentage of residents with a high school degree, (ii) percentage of residents with a bachelor's degree, (iii) median age of residents, (iv) percentage of residents without health insurance, and (v) the local unemployment rate. All control variables are also observed at the zip code tabulation area level. Plot (a) shows that their relationship is nonlinear and unstable, while it becomes monotonically decreasing when the control variables are added, as shown in Plot (b).

Figure IV.1 Gini Index versus Household Income.


(a) No Controls

(b) With Controls

Notes. Data are obtained from the American Community Survey (ACS) using the 5-year survey estimates (2013-2017).

For formal testing of substantive features of $\mu(x)$ we recommend the following:

Step 1. Use IMSE-optimal cubic B-spline to approximate the function $\mu(x)$, with covariate adjustment and accounting for clustered data as appropriate. Specifically, set $p = 3$ and $s = 3$, and $J = J_{\mathtt{IMSE}}$ (Theorem II.1), for $v = 0, 1, 2$ (see Section 3.5 for details). For $v > 2$, use $p = 3 + v$ and $s = 3 + v$, and $J = J_{\mathtt{IMSE}}$ (Theorem II.1).

Step 2. Conduct formal hypothesis testing procedures using Theorem III.4 and Theorem III.5, as appropriate, with $q = 1$.

This approach is illustrated in Table IV.1. Specifically, the results indicate that the relation between the Gini index and household income, controlling for covariates, is nonlinear, but can be modelled well by a quadratic polynomial. Moreover, the hypotheses of monotonicity and convexity are also supported. In this empirical example, positivity holds by construction, but the test is nonetheless included for completeness.

4.6 Simulations

Finally, we conducted a Monte Carlo investigation of the finite sample performance of our methods. We considered three univariate ($d = 1$), two bivariate ($d = 2$) and two trivariate ($d = 3$) models, but only summarize one univariate design here for brevity.

Table IV.1 Testing of Substantive Hypothesis.

	Test Statistic	P-value	J
Parametric Specification			
Constant	17.665	0.000	21
Linear	6.152	0.000	21
Quadratic	1.947	0.268	21
Shape Restrictions			
Positivity	9.666	1.000	21
Decreasingness	−0.634	1.000	9
Convexity	−1.678	0.420	5

Notes. A set of control variables are added. The number of bins is IMSE-optimal, selected based on a fully data-driven procedure.

Complete results and details are available in the online supplemental appendix to Cattaneo, Farrell, and Feng (2018a). All numerical results were obtained using our companion R package `lspartition` (Cattaneo, Farrell, and Feng, 2018b).

We set $\theta(x) = \sin(\pi x - \pi/2)/(1+2(2x-1)^2(\text{sign}(2x-1)+1))$, with $\text{sign}(\cdot)$ denoting the sign function. We generate samples $\{(y_i, x_i) : i = 1, \ldots, n\}$ from $y_i = \theta(x_i) + \varepsilon_i$, where $x_i \sim \mathsf{U}[0,1]$ and $\varepsilon_i \sim \mathsf{N}(0,1)$, independent of each other. We consider $5,000$ simulated datasets with $n = 1,000$ each time. Results based on splines and wavelets are presented. Specifically, we use linear splines or Daubechies (father) wavelets of order 2 ($m = 2$) to form the point estimator $\widehat{\theta}_0(x)$, and quadratic splines or Daubechies wavelets of order 3 ($\tilde{m} = 3$) for bias correction, on the same evenly spaced partitioning scheme for point estimation and bias correction ($\Delta = \tilde{\Delta}$).

The results are presented in Table IV.2. Column "RMSE" reports (simulated) root mean squared error for point estimators, while the columns "CR" and "AL" report coverage rate and average interval length of pointwise 95% nominal confidence intervals at $x = 0.5$. The columns under "Uniform" present uniform inference results, including the uniform coverage rate (UCR) and the average width (AW) of the 95% nominal confidence band. For B-splines, we employ either the infeasible IMSE-optimal size choice ($\kappa_{\texttt{IMSE}}$), a rule-of-thumb estimate ($\hat{\kappa}_{\texttt{ROT}}$), or a direct plug-in estimate ($\hat{\kappa}_{\texttt{DPI}}$). For wavelets, the tuning parameter is instead the resolution level (resp., $s_{\texttt{IMSE}}$, $\hat{s}_{\texttt{ROT}}$, or $\hat{s}_{\texttt{DPI}}$), which is the logarithm of the number of subintervals (to base 2). Finally, the table reports all four (estimation and) inference methods discussed in this dissertation, indexed by $j = 0, 1, 2, 3$. Due to the lack of smoothness of low-order wavelet bases, plug-in bias correction ($j = 3$) is practically cumbersome and hence not implemented.

All the numerical findings are consistent with our theoretical results. To briefly summarize: robust bias-correction seems to perform quite well, always delivering close-to-correct coverage, both pointwise and uniformly. The improvement is less pronounced for wavelets since the number of basis increases rapidly with the resolution. However, if the underlying model is highly nonlinear, bias correction does make a difference. In addition, the numerical performance of our rule-of-thumb (ROT) and direct plug-in (DPI) knot selection procedures for tensor-product partitions worked well in this simulation study.

Table IV.2 Simulation Evidence

(a) B-Splines ($m = 2$, $\tilde{m} = 3$, $\Delta = \tilde{\Delta}$, Evenly Spaced Partition)

			Pointwise		**Uniform**	
	κ	RMSE	CR	AL	UCR	AW
$j = 0$						
$\kappa_{\texttt{IMSE}}$	3.0	0.046	91.5	0.328	79.7	0.384
$\hat{\kappa}_{\texttt{ROT}}$	4.9	0.009	94.6	0.317	92.2	0.469
$\hat{\kappa}_{\texttt{DPI}}$	5.1	0.007	94.4	0.318	91.4	0.478
$j = 1$						
$\kappa_{\texttt{IMSE}}$	3.0	0.003	94.8	0.226	93.9	0.426
$\hat{\kappa}_{\texttt{ROT}}$	4.9	0.006	95.0	0.298	93.7	0.506
$\hat{\kappa}_{\texttt{DPI}}$	5.1	0.006	95.1	0.306	93.4	0.514
$j = 2$						
$\kappa_{\texttt{IMSE}}$	3.0	0.004	94.7	0.268	94.1	0.443
$\hat{\kappa}_{\texttt{ROT}}$	4.9	0.003	95.0	0.336	93.8	0.536
$\hat{\kappa}_{\texttt{DPI}}$	5.1	0.003	94.9	0.342	93.3	0.546
$j = 3$						
$\kappa_{\texttt{IMSE}}$	3.0	0.034	92.7	0.321	89.0	0.413
$\hat{\kappa}_{\texttt{ROT}}$	4.9	0.006	94.8	0.328	93.6	0.499
$\hat{\kappa}_{\texttt{DPI}}$	5.1	0.005	94.3	0.331	93.0	0.509

(b) Wavelets ($m = 2$, $\tilde{m} = 3$, $\Delta = \tilde{\Delta}$, Evenly Spaced Partition)

			Pointwise		**Uniform**	
	s	RMSE	CR	AL	UCR	AW
$j = 0$						
$s_{\texttt{IMSE}}$	3.0	0.001	94.1	0.497	90.9	0.509
$\hat{s}_{\texttt{ROT}}$	2.4	0.001	94.1	0.497	90.9	0.509
$\hat{s}_{\texttt{DPI}}$	2.9	0.001	94.0	0.501	90.8	0.514
$j = 1$						
$s_{\texttt{IMSE}}$	3.0	0.037	93.6	0.450	89.9	0.504
$\hat{s}_{\texttt{ROT}}$	2.4	0.037	93.6	0.450	89.9	0.504
$\hat{s}_{\texttt{DPI}}$	2.9	0.035	93.8	0.455	89.7	0.510
$j = 2$						
$s_{\texttt{IMSE}}$	3.0	0.007	94.1	0.533	91.4	0.576
$\hat{s}_{\texttt{ROT}}$	2.4	0.007	94.1	0.533	91.4	0.576
$\hat{s}_{\texttt{DPI}}$	2.9	0.007	94.1	0.538	91.3	0.581

Notes:
(i) Pointwise = pointwise inference at $x = 0.5$, Uniform = uniform inference.
(ii) RMSE = root MSE of point estimator, CR = coverage rate of 95% nominal confidence intervals, AL = average interval length of 95% nominal confidence intervals.
(iii) UCR = uniform coverage rate of 95% nominal confidence band, AW = average width of 95% nominal confidence band.
(iv) $\kappa_{\texttt{IMSE}}$ ($s_{\texttt{IMSE}}$) = infeasible IMSE-optimal number of intervals (resolution level), $\hat{\kappa}_{\texttt{ROT}}$ ($\hat{s}_{\texttt{ROT}}$) = feasible rule-of-thumb (ROT) implementation of $\kappa_{\texttt{IMSE}}$ ($s_{\texttt{IMSE}}$), $\hat{\kappa}_{\texttt{DPI}}$ ($\hat{s}_{\texttt{DPI}}$) = feasible direct plug-in (DPI) implementation of $\kappa_{\texttt{IMSE}}$ ($s_{\texttt{IMSE}}$).

Chapter V
Conclusion

This dissertation studies estimation and inference using partitioning-based least squares estimators in nonparametric and semiparametric models. An array of theoretical and practical results is presented, including bias approximations, integrated mean squared error (IMSE) expansions, and pointwise and uniform distributional approximations. In particular, these results can be used to construct valid pointwise confidence intervals and uniform confidence bands, allowing for mean squared error minimizing tuning parameter choices. A special focus is put on binscatter, a particular class of partitioning-based estimators in semiparametric models. We provide novel smooth and/or polynomial approximation approaches, principled covariate adjustment and number of bins selection, and valid inference and hypothesis testing methods.

A number of important extensions are left for future work. This dissertation focuses on cross-sectional data only. Extending the theory presented here to time series and panel data models seems like a promising avenue. Furthermore, another important question of both theoretical and practical interest is whether this work can be extended to scenarios in which a general data-dependent partition is employed. Chapter III considers a very special case: the partitioning knots are empirical quantiles of an independent variable. Such partitions are determined by covariates only, and under weak regularity conditions, the quasi-uniformity condition specified in Chapter II is satisfied. However, many modern machine learning techniques, like regression trees, construct data-dependent partitions in more convoluted ways, leading to further technical complications and probable violation of the quasi-uniformity assumption. A systematic approach to regularizing partitions, combined with sample splitting, may be needed in such cases.

Appendices

Appendix A

Proof for Chapter II

The proofs for Chapter II rely on a collection of technical results, which are summarized in the following lemma. Let $\widehat{\mathbf{Q}}_m = \mathbb{E}_n[\mathbf{p}(\mathbf{x}_i)\mathbf{p}(\mathbf{x}_i)']$, $\widehat{\mathbf{Q}}_{\tilde{m}} = \mathbb{E}_n[\tilde{\mathbf{p}}(\mathbf{x}_i)\tilde{\mathbf{p}}(\mathbf{x}_i)']$, $\mathbf{Q}_m = \mathbb{E}[\mathbf{p}(\mathbf{x}_i)\mathbf{p}(\mathbf{x}_i)']$, and $\mathbf{Q}_{\tilde{m}} = \mathbb{E}[\tilde{\mathbf{p}}(\mathbf{x}_i)\tilde{\mathbf{p}}(\mathbf{x}_i)']$. $C, C_1, C_2, \ldots$ are universal constants.

Lemma A.1. *Let Assumptions II.1, II.2, II.3, and II.5 hold. If* $\frac{\log n}{nh^d} = o(1)$, *then,*

(i) $\|\widehat{\mathbf{Q}}_m - \mathbf{Q}_m\| \lesssim_{\mathbb{P}} h^d\sqrt{\frac{\log n}{nh^d}}$, $\quad \|\widehat{\mathbf{Q}}_m - \mathbf{Q}_m\|_\infty \lesssim_{\mathbb{P}} h^d\sqrt{\frac{\log n}{nh^d}}$;

(ii) $\|\widehat{\mathbf{Q}}_m\| \lesssim_{\mathbb{P}} h^d$, $\quad \|\widehat{\mathbf{Q}}_m^{-1}\|_\infty \lesssim_{\mathbb{P}} h^{-d}$;

(iii) $\sup_{\mathbf{x}\in\mathcal{X}} \|\boldsymbol{\gamma}_{\mathbf{q},j}(\mathbf{x})'\|_\infty \lesssim h^{-d-[\mathbf{q}]}$, $\quad \sup_{\mathbf{x}\in\mathcal{X}} \|\widehat{\boldsymbol{\gamma}}_{\mathbf{q},j}(\mathbf{x})' - \boldsymbol{\gamma}_{\mathbf{q},j}(\mathbf{x})'\|_\infty \lesssim h^{-d-[\mathbf{q}]}\sqrt{\frac{\log n}{nh^d}}$,

$\inf_{\mathbf{x}\in\mathcal{X}} \|\boldsymbol{\gamma}_{\mathbf{q},j}(\mathbf{x})'\| \gtrsim h^{-d-[\mathbf{q}]}$ *for each* $j = 0, 1, 2, 3$;

(iv) $\sup_{\mathbf{x}\in\mathcal{X}} \Omega_j(\mathbf{x}) \lesssim h^{-d-2[\mathbf{q}]}$, $\quad \inf_{\mathbf{x}\in\mathcal{X}} \Omega_j(\mathbf{x}) \gtrsim h^{-d-2[\mathbf{q}]}$ *for each* $j = 0, 1, 2, 3$.

Proof. See Section SA-10 of the supplemental appendix to Cattaneo, Farrell, and Feng (2018a). □

Notice that the results for $\widehat{\mathbf{Q}}_m$ and $\mathbf{Q}_m$ also hold for $\widehat{\mathbf{Q}}_{\tilde{m}}$ and $\mathbf{Q}_{\tilde{m}}$ under Assumption II.5.

Proof of Lemma II.1. For s^* in Assumption II.4,

$$\begin{aligned}
&\mathbb{E}[\widehat{\partial^{\mathbf{q}}\theta}_0(\mathbf{x})|\mathbf{X}] - \partial^{\mathbf{q}}\theta(\mathbf{x})\\
&= \mathscr{B}_{m,\mathbf{q}}(\mathbf{x}) + \widehat{\boldsymbol{\gamma}}_{\mathbf{q},0}(\mathbf{x})'\mathbb{E}_n[\mathbf{p}(\mathbf{x}_i)(\theta(\mathbf{x}_i) - s^*(\mathbf{x}_i))] + O(h^{m+\varrho-[\mathbf{q}]})\\
&= \mathscr{B}_{m,\mathbf{q}}(\mathbf{x}) - \widehat{\boldsymbol{\gamma}}_{\mathbf{q},0}(\mathbf{x})'\mathbb{E}_n\left[\mathbf{p}(\mathbf{x}_i)\mathscr{B}_{m,\mathbf{0}}(\mathbf{x}_i)\right] + O(h^{m+\varrho-[\mathbf{q}]})\\
&\quad + \widehat{\boldsymbol{\gamma}}_{\mathbf{q},0}(\mathbf{x})'\mathbb{E}_n\Big[\mathbf{p}(\mathbf{x}_i)(\theta(\mathbf{x}_i) - s^*(\mathbf{x}_i) + \mathscr{B}_{m,\mathbf{0}}(\mathbf{x}_i))\Big].
\end{aligned}$$

By Assumption II.3 and II.4, $\|\mathbb{E}[\mathbf{p}(\mathbf{x}_i)(\theta(\mathbf{x}_i) - s^*(\mathbf{x}_i) + \mathscr{B}_{m,\mathbf{0}}(\mathbf{x}_i))\|_\infty \lesssim_{\mathbb{P}} h^{m+\varrho+d}$. Also, $\|\mathbb{G}_n[\mathbf{p}(\mathbf{x}_i)(\theta(\mathbf{x}_i) - s^*(\mathbf{x}_i) + \mathscr{B}_{m,\mathbf{0}}(\mathbf{x}_i))\|_\infty \lesssim_{\mathbb{P}} h^{m+\varrho+\frac{d}{2}}\sqrt{\log n}$ by Bernstein's inequality. Then, by Lemma A.1, the last term in the above expansion is $O_{\mathbb{P}}(h^{m+\varrho-[\mathbf{q}]})$. □

Proof of Theorem II.1. Regarding the integrated conditional variance,

$$\begin{aligned}
&\int_{\mathcal{X}} \mathbb{V}[\widehat{\partial^{\mathbf{q}}\theta}_0(\mathbf{x})|\mathbf{X}]w(\mathbf{x})\,d\mathbf{x} \\
&= \frac{1}{n}\operatorname{trace}\left[\boldsymbol{\Sigma}_0 \int_{\mathcal{X}} \boldsymbol{\gamma}_{\mathbf{q},0}(\mathbf{x})\boldsymbol{\gamma}_{\mathbf{q},0}(\mathbf{x})' w(\mathbf{x})\,d\mathbf{x}\right] + o_{\mathbb{P}}\left(\frac{1}{nh^{d+2[\mathbf{q}]}}\right) \\
&\lesssim_{\mathbb{P}} \frac{1}{nh^d}\operatorname{trace}\left[\int_{\mathcal{X}} \partial^{\mathbf{q}}\mathbf{p}(\mathbf{x})\partial^{\mathbf{q}}\mathbf{p}(\mathbf{x})' w(\mathbf{x}) d\mathbf{x}\right] \lesssim \frac{1}{nh^{d+2[\mathbf{q}]}}
\end{aligned}$$

where the second line holds by Lemma A.1, the third by Trace Inequality, the continuity of $w(\cdot)$ and Lemma A.1. Since $\sigma^2(\cdot)$ and $w(\cdot)$ are bounded away from zero, the other side of the bound follows similarly.

Regarding the integrated squared bias, we have

$$\begin{aligned}
&\int_{\mathcal{X}} \left(\mathbb{E}[\widehat{\partial^{\mathbf{q}}\theta}_0(\mathbf{x})|\mathbf{X}] - \partial^{\mathbf{q}}\theta(\mathbf{x})\right)^2 w(\mathbf{x})d\mathbf{x} \\
&= \int_{\mathcal{X}} \mathscr{B}_{m,\mathbf{q}}(\mathbf{x})^2 w(\mathbf{x})d\mathbf{x} + \int_{\mathcal{X}} \left(\boldsymbol{\gamma}_{\mathbf{q},0}(\mathbf{x})'\mathbb{E}[\mathbf{p}(\mathbf{x}_i)\mathscr{B}_{m,\mathbf{0}}(\mathbf{x}_i)]\right)^2 w(\mathbf{x})d\mathbf{x} \\
&\quad - 2\int_{\mathcal{X}} \mathscr{B}_{m,\mathbf{q}}(\mathbf{x})\boldsymbol{\gamma}_{\mathbf{q},0}(\mathbf{x})'\mathbb{E}[\mathbf{p}(\mathbf{x}_i)\mathscr{B}_{m,\mathbf{0}}(\mathbf{x}_i)]w(\mathbf{x})d\mathbf{x} + o_{\mathbb{P}}(h^{2m-2[\mathbf{q}]}) \\
&=: \mathrm{B}_1 + \mathrm{B}_2 - 2\mathrm{B}_3 + o_{\mathbb{P}}(h^{2m-2[\mathbf{q}]}).
\end{aligned} \tag{A.1}$$

Let h_δ be the diameter of δ and $\mathbf{t}_\delta^*$ be an arbitrary point in δ. Then

$$\begin{aligned}
\mathrm{B}_1 &= \sum_{\boldsymbol{u}_1,\boldsymbol{u}_2\in\Lambda_m} \sum_{\delta\in\Delta} \int_\delta \left[h_\delta^{2m-2[\mathbf{q}]}\partial^{\boldsymbol{u}_1}\theta(\mathbf{t}_\delta^*)\partial^{\boldsymbol{u}_2}\theta(\mathbf{t}_\delta^*)B_{\boldsymbol{u}_1,\mathbf{q}}(\mathbf{x})B_{\boldsymbol{u}_2,\mathbf{q}}(\mathbf{x})\right]w(\mathbf{t}_\delta^*)d\mathbf{x} \\
&\qquad + o(h^{2m-2[\mathbf{q}]}) \\
&= \sum_{\boldsymbol{u}_1,\boldsymbol{u}_2\in\Lambda_m} \sum_{\delta\in\Delta} \left\{h_\delta^{2m-2[\mathbf{q}]}\partial^{\boldsymbol{u}_1}\theta(\mathbf{t}_\delta^*)\partial^{\boldsymbol{u}_2}\theta(\mathbf{t}_\delta^*)w(\mathbf{t}_\delta^*)\int_\delta B_{\boldsymbol{u}_1,\mathbf{q}}(\mathbf{x})B_{\boldsymbol{u}_2,\mathbf{q}}(\mathbf{x})d\mathbf{x}\right\} \\
&\qquad + o(h^{2m-2[\mathbf{q}]}),
\end{aligned}$$

where the second line holds by the continuity of $\partial^{\boldsymbol{u}_1}\theta(\cdot)$, $\partial^{\boldsymbol{u}_2}\theta(\cdot)$ and $w(\cdot)$, and by Assumption II.2 and II.4, $\mathrm{B}_1 \lesssim h^{2m-2[\mathbf{q}]}$. The other two terms can be bounded similarly. □

To prove Theorem II.2, the following lemma is needed:

Lemma A.2 (Linearization)**.** *Let Assumptions II.1–II.5 hold. if* $\frac{\log n}{nh^d} = o(1)$*, then for* *each $x \in \mathcal{X}$, for each $j = 0, 1, 2, 3$,*

$$\begin{aligned} \widehat{\partial^{\mathbf{q}}\theta}_j(\mathbf{x}) - \partial^{\mathbf{q}}\theta(\mathbf{x}) &= \boldsymbol{\gamma}_{\mathbf{q},j}(\mathbf{x})'\mathbb{E}_n[\boldsymbol{\Pi}_j(\mathbf{x}_i)\varepsilon_i] + R_{1n,\mathbf{q}}(\mathbf{x}) + R_{2n,\mathbf{q}}(\mathbf{x}), \quad \textit{where} \\ R_{1n,\mathbf{q}}(\mathbf{x}) &:= (\widehat{\boldsymbol{\gamma}}_{\mathbf{q},j}(\mathbf{x}) - \boldsymbol{\gamma}_{\mathbf{q},j}(\mathbf{x})')\,\mathbb{E}_n[\boldsymbol{\Pi}_j(\mathbf{x}_i)\varepsilon_i] \lesssim_{\mathbb{P}} \frac{\sqrt{\log n}}{nh^{d+[\mathbf{q}]}}, \\ R_{2n,\mathbf{q}}(\mathbf{x}) &:= \mathbb{E}[\widehat{\partial^{\mathbf{q}}\theta}_j(\mathbf{x})|\mathbf{X}] - \partial^{\mathbf{q}}\theta(\mathbf{x}) \lesssim_{\mathbb{P}} h^{m-[\mathbf{q}]}. \end{aligned}$$

For $j = 1, 2, 3$, $R_{2n,\mathbf{q}}(\mathbf{x}) \lesssim_{\mathbb{P}} h^{m+\varrho-[\mathbf{q}]}$.

If, in addition, one of the following holds:

(i) $\mathbb{E}[|\varepsilon_i|^{2+\nu}] < \infty$ *for some* $\nu > 0$, *and* $\frac{n^{\frac{2}{2+\nu}}(\log n)^{\frac{2\nu}{4+2\nu}}}{nh^d} \lesssim 1$; *or*

(ii) $\mathbb{E}[|\varepsilon_i|^3 \exp(|\varepsilon_i|)] < \infty$, *and* $\frac{(\log n)^3}{nh^d} \lesssim 1$.

Then, $\sup_{\mathbf{x}\in\mathcal{X}} |R_{1n,\mathbf{q}}(\mathbf{x})| \lesssim_{\mathbb{P}} \frac{\log n}{nh^{d+[\mathbf{q}]}} =: \bar{R}_{1n,\mathbf{q}}$. *For $j = 0$,* $\sup_{\mathbf{x}\in\mathcal{X}} |R_{2n,\mathbf{q}}(\mathbf{x})| \lesssim_{\mathbb{P}} h^{m-[\mathbf{q}]} =: \bar{R}_{2n,\mathbf{q}}$, *and for $j = 1, 2, 3$,* $\sup_{\mathbf{x}\in\mathcal{X}} |R_{2n,\mathbf{q}}(\mathbf{x})| \lesssim_{\mathbb{P}} h^{m+\varrho-[\mathbf{q}]} =: \bar{R}_{2n,\mathbf{q}}$.

Proof. For $j = 0, 1$, the results directly follow from Assumption II.5, Lemma II.1 and Belloni, Chernozhukov, Chetverikov, and Kato (2015, Lemma 4.1). For $j = 2, 3$, conditional on $\mathbf{X}$, $R_{1n,\mathbf{q}}(\mathbf{x})$ has mean zero, and $\mathbb{V}[R_{1n,\mathbf{q}}(\mathbf{x})|\mathbf{X}] \lesssim \frac{1}{n}\|\widehat{\boldsymbol{\gamma}}_{\mathbf{q},j}(\mathbf{x})' - \boldsymbol{\gamma}_{\mathbf{q},j}(\mathbf{x})'\|^2\|\mathbb{E}[\boldsymbol{\Pi}_j(\mathbf{x}_i)\boldsymbol{\Pi}_j(\mathbf{x}_i)']\| \lesssim_{\mathbb{P}} \log n/(n^2h^{2d+2[\mathbf{q}]})$ by Lemma A.1. Then by Chebyshev's inequality, $R_{1n,\mathbf{q}}(\mathbf{x}) \lesssim_{\mathbb{P}} \sqrt{\log n}/(nh^{d+[\mathbf{q}]})$.

Regarding the conditional bias $R_{2n,\mathbf{q}}(\mathbf{x})$, for $j = 2$, by construction,

$$\begin{aligned} \mathbb{E}[\widehat{\partial^{\mathbf{q}}\theta}_2(\mathbf{x})|\mathbf{X}] - \partial^{\mathbf{q}}\theta(\mathbf{x}) &= \left(\mathbb{E}[\widehat{\partial^{\mathbf{q}}\theta}_1(\mathbf{x})|\mathbf{X}] - \partial^{\mathbf{q}}\theta(\mathbf{x})\right) - \partial^{\mathbf{q}}\mathbf{p}(\mathbf{x})'\widehat{\mathbf{Q}}_m^{-1}\mathbb{E}_n[\mathbf{p}(\mathbf{x}_i)\mathfrak{B}_{\tilde{m},\mathbf{0}}(\mathbf{x}_i)] \\ &= O_{\mathbb{P}}(h^{m+\varrho-[\mathbf{q}]}) - \partial^{\mathbf{q}}\mathbf{p}(\mathbf{x})'\widehat{\mathbf{Q}}_m^{-1}\mathbb{E}_n[\mathbf{p}(\mathbf{x}_i)\mathfrak{B}_{\tilde{m},\mathbf{0}}(\mathbf{x}_i)] \end{aligned}$$

where $\mathfrak{B}_{\tilde{m},\mathbf{0}}(\mathbf{x}_i) = \mathbb{E}[\widehat{\theta}_1(\mathbf{x}_i)|\mathbf{X}] - \theta(\mathbf{x}_i)$ is the conditional bias of $\widehat{\theta}_1(\mathbf{x}_i)$ and the last line follows from Lemma II.1. Then it follows from Lemma A.1 and II.1 that the conditional bias of $\widehat{\partial^{\mathbf{q}}\theta}_2(\mathbf{x})$ is $O_{\mathbb{P}}(h^{m+\varrho-[\mathbf{q}]})$.

Next, for $j = 3$, using Lemma II.1, we have

$$\begin{aligned} &\mathbb{E}[\widehat{\partial^{\mathbf{q}}\theta}_3(\mathbf{x})|\mathbf{X}] - \partial^{\mathbf{q}}\theta(\mathbf{x}) \\ =& \sum_{\boldsymbol{u}\in\Lambda_m} h_{\mathbf{x}}^{m-[\mathbf{q}]} B_{\boldsymbol{u},\mathbf{q}}(\mathbf{x})\mathbb{E}\left[\widehat{\partial^{\boldsymbol{u}}\theta}_1(\mathbf{x}) - \partial^{\boldsymbol{u}}\theta(\mathbf{x})|\mathbf{X}\right] + \\ &\partial^{\mathbf{q}}\mathbf{p}(\mathbf{x})'\widehat{\mathbf{Q}}_m^{-1}\mathbb{E}_n\left[\mathbf{p}(\mathbf{x}_i)\left(\mathbb{E}[\widehat{\mathscr{B}}_{m,\mathbf{0}}(\mathbf{x}_i)|\mathbf{X}] - \mathscr{B}_{m,\mathbf{0}}(\mathbf{x}_i)\right)\right] + O_{\mathbb{P}}(h^{m+\varrho-[\mathbf{q}]}) \\ =& \partial^{\mathbf{q}}\mathbf{p}(\mathbf{x})'\widehat{\mathbf{Q}}_m^{-1}\mathbb{E}_n\left[\mathbf{p}(\mathbf{x}_i)\left(\mathbb{E}[\widehat{\mathscr{B}}_{m,\mathbf{0}}(\mathbf{x}_i)|\mathbf{X}] - \mathscr{B}_{m,\mathbf{0}}(\mathbf{x}_i)\right)\right] + O_{\mathbb{P}}(h^{m+\varrho-[\mathbf{q}]}) \end{aligned}$$

where $\widehat{\mathscr{B}}_{m,\mathbf{0}}(\mathbf{x}) = -\sum_{\boldsymbol{u}\in\Lambda_m}\left(\widehat{\partial^{\mathbf{q}}\theta}_1(\mathbf{x})\right) h_{\mathbf{x}}^m B_{\boldsymbol{u},\mathbf{q}}(\mathbf{x})$, and the last line follows from Assumption II.4, II.5 and Lemma II.1. Also by Lemma II.1 and the fact that $B_{\boldsymbol{u},\mathbf{0}}(\cdot)$ is bounded, $\sup_{\mathbf{x}\in\mathcal{X}}|\mathbb{E}[\widehat{\mathscr{B}}_{m,\mathbf{0}}(\mathbf{x})|\mathbf{X}]-\mathscr{B}_{m,\mathbf{0}}(\mathbf{x})|\lesssim_{\mathbb{P}} h^{m+\varrho}$. The desired result immediately follows by using the similar argument for $j=2$.

Now, suppose that the additional conditions in **(i)** hold. We first bound $\sup_{\mathbf{x}\in\mathcal{X}}|R_{1n,\mathbf{q}}(\mathbf{x})|$ for $j=0,1,2,3$. To simplify notation, we write $\mathbf{\Pi}_j(\mathbf{x}_i)=(\pi_1(\mathbf{x}_i),\ldots,\pi_{K_j}(\mathbf{x}_i))'$ where $K_j=\dim(\mathbf{\Pi}_j(\cdot))$. We will truncate the errors by an increasing sequence of constants $\{\vartheta_n : n\geq 1\}$ such that $\vartheta_n\asymp\sqrt{nh^d/\log n}$. Let $H_{ik}=\pi_k(\mathbf{x}_i)(\varepsilon_i\mathbb{1}\{|\varepsilon_i|\leq\vartheta_n\}-\mathbb{E}[\varepsilon_i\mathbb{1}\{|\varepsilon_i|\leq\vartheta_n|\mathbf{x}_i\}])$ and $T_{ik}=\pi_k(\mathbf{x}_i)(\varepsilon_i\mathbb{1}\{|\varepsilon_i|>\vartheta_n\}-\mathbb{E}[\varepsilon_i\mathbb{1}\{|\varepsilon_i|>\vartheta_n|\mathbf{x}_i\}])$. Regarding the truncated term, it follows from the truncation strategy, Assumption II.3 and II.5 that $|H_{ik}|\leq\vartheta_n$ and $\mathbb{E}[H_{ik}^2]\lesssim h^d$. By Bernstein's inequality, $\max_{1\leq k\leq K_j}|\mathbb{E}_n[H_{ik}]|\lesssim_{\mathbb{P}} h^d\sqrt{\log n/(nh^d)}$. It immediately follows from Lemma A.1 that

$$\sup_{\mathbf{x}\in\mathcal{X}}\Big|(\widehat{\boldsymbol{\gamma}}_{\mathbf{q},j}(\mathbf{x})'-\boldsymbol{\gamma}_{\mathbf{q},j}(\mathbf{x})')\mathbb{E}_n[\mathbf{\Pi}_j(\mathbf{x}_i)(\varepsilon_i\mathbb{1}\{|\varepsilon_i|\leq\vartheta_n\}-\mathbb{E}[\varepsilon_i\mathbb{1}\{|\varepsilon_i|\leq\vartheta_n|\mathbf{x}_i\}])]\Big|$$
$$\lesssim_{\mathbb{P}} h^{-[\mathbf{q}]-d}\sqrt{\log n/(nh^d)}\,h^d\sqrt{\log n/(nh^d)}=h^{-[\mathbf{q}]}\log n/(nh^d).$$

Regarding the tails, let $\mathscr{K}_{ji}(\mathbf{x}):=(\widehat{\boldsymbol{\gamma}}_{\mathbf{q},j}(\mathbf{x})'-\boldsymbol{\gamma}_{\mathbf{q},j}(\mathbf{x})')\mathbf{\Pi}_j(\mathbf{x}_i)$. By Lemma A.1 and Assumption II.3, we have $\sup_{\mathbf{x}\in\mathcal{X}}|\mathscr{K}_{ji}(\mathbf{x})|\lesssim_{\mathbb{P}} h^{-d-[\mathbf{q}]}\sqrt{\log n/(nh^d)}$. Let $\mathcal{A}_n(M)$ denote the event on which $\sup_{\mathbf{x}\in\mathcal{X}}|\mathscr{K}_{ji}(\mathbf{x})|\leq Mh^{-d-[\mathbf{q}]}\sqrt{\log n/(nh^d)}$ for some $M>0$, and $\mathbb{1}_{\mathcal{A}_n(M)}$ be an indicator function of $\mathcal{A}_n(M)$. Then by Markov's inequality, for $t>0$,

$$\mathbb{P}\Big(\sup_{\mathbf{x}\in\mathcal{X}}\Big|\mathbb{E}_n[\mathbb{1}_{\mathcal{A}_n(M)}\mathscr{K}_{ji}(\varepsilon_i\mathbb{1}\{|\varepsilon_i|>\vartheta_n\}-\mathbb{E}[\varepsilon_i\mathbb{1}\{|\varepsilon_i|>\vartheta_n|\mathbf{x}_i\}])]\Big|>\frac{t\log n}{nh^{d+[\mathbf{q}]}}\Big)$$
$$\lesssim\frac{Mh^{-d-[\mathbf{q}]}\sqrt{\log n/(nh^d)}\,\mathbb{E}[|\varepsilon_i|\mathbb{1}\{|\varepsilon_i|>\vartheta_n\}]}{th^{-[\mathbf{q}]}\log n/(nh^d)}\leq\frac{M\sqrt{n}}{t\sqrt{h^d\log n}}\,\frac{\mathbb{E}[|\varepsilon_i|^{2+\nu}]}{\vartheta_n^{1+\nu}}$$

which is arbitrarily small for t/M large enough by the additional moment condition specified in the lemma and the rate restriction. Since $\mathbb{P}(\mathcal{A}_n(M)^c)=o(1)$ as $M\to\infty$, simply let $t=M^2$ and $M\to\infty$, then the desired conclusion immediately follows.

The bound on $\sup_{\mathbf{x}\in\mathcal{X}}|R_{2n,\mathbf{q}}(\mathbf{x})|$ follows from Lemma II.1 and Assumption II.4.

Finally, the proof under condition **(ii)** is similar except that we let $\vartheta_n=\log n$ in the proof for $R_{1n,\mathbf{q}}(\mathbf{x})$. □

Proof of Theorem II.2. Regarding the L_2 convergence, using Lemma II.1,

$$\int_{\mathcal{X}}\Big(\widehat{\partial^{\mathbf{q}}\theta}_0(\mathbf{x})-\partial^{\mathbf{q}}\theta(\mathbf{x})\Big)^2\Big)w(\mathbf{x})d\mathbf{x}$$
$$=\Big(\mathbb{E}_n[\mathbf{\Pi}_0(\mathbf{x}_i)\varepsilon_i]'\Big)\Big(\int_{\mathcal{X}}\widehat{\boldsymbol{\gamma}}_{\mathbf{q},0}(\mathbf{x})\widehat{\boldsymbol{\gamma}}_{\mathbf{q},0}(\mathbf{x})'w(\mathbf{x})d\mathbf{x}\Big)\Big(\mathbb{E}_n[\mathbf{\Pi}_0(\mathbf{x}_i)\varepsilon_i]\Big)+O_{\mathbb{P}}(h^{2(m-[\mathbf{q}])}).$$

The uniform bound on the conditional bias does not require explicit expression of leading approximation error. Then by Lemma A.1, we have $\int_{\mathcal{X}} \widehat{\boldsymbol{\gamma}}_{\mathbf{q},0}(\mathbf{x})\widehat{\boldsymbol{\gamma}}_{\mathbf{q},0}(\mathbf{x})'w(\mathbf{x})d\mathbf{x} \lesssim_{\mathbb{P}} h^{-d-2[\mathbf{q}]}$. Also, $\mathbb{E}[\|\mathbb{E}_n[\mathbf{\Pi}_0(\mathbf{x}_i)\varepsilon_i]\|^2] \lesssim \mathbb{E}[\mathbf{\Pi}_0(\mathbf{x}_i)'\mathbf{\Pi}_0(\mathbf{x}_i)/n] \lesssim 1/n$. The desired L_2 convergence rate follows.

Regarding the uniform convergence, consider the case when the conditions of Lemma A.2 hold. We use the same truncation strategy as used in the proof of Lemma A.2. Specifically, separate ε_i into

$$\varepsilon_i \mathbb{1}\{|\varepsilon_i| \leq \vartheta_n\} - \mathbb{E}[\varepsilon_i \mathbb{1}\{|\varepsilon_i| \leq \vartheta_n\}|\mathbf{x}_i] \quad \text{and} \quad \varepsilon_i \mathbb{1}\{|\varepsilon_i| > \vartheta_n\} - \mathbb{E}[\varepsilon_i \mathbb{1}\{|\varepsilon_i| > \vartheta_n\}|\mathbf{x}_i]$$

where $\vartheta_n \asymp \sqrt{nh^d/\log n}$. The remaining argument is similar. □

Proof of Theorem II.3. By Lemma A.2, for each $j = 0, 1, 2, 3$,

$$\widehat{\partial^{\mathbf{q}}\theta}_j(\mathbf{x}) - \partial^{\mathbf{q}}\theta(\mathbf{x}) = \boldsymbol{\gamma}_{\mathbf{q},j}(\mathbf{x})'\mathbb{E}_n[\mathbf{\Pi}_j(\mathbf{x}_i)\varepsilon_i] + O_{\mathbb{P}}\Big(\frac{\sqrt{\log n}}{nh^{d+[\mathbf{q}]}}\Big) + O_{\mathbb{P}}(h^{m-[\mathbf{q}]}).$$

For $j = 1, 2, 3$, the last term is $O_{\mathbb{P}}(h^{m+\varrho-[\mathbf{q}]})$. Under the rate restriction given in the theorem, it suffices to show that the first term satisfies Lindeberg's condition. Clearly, $\mathbb{V}\Big[\frac{\boldsymbol{\gamma}_{\mathbf{q},j}(\mathbf{x})'}{\sqrt{\Omega_j(\mathbf{x})}}\mathbb{G}_n[\mathbf{\Pi}_j(\mathbf{x}_i)\varepsilon_i]\Big] = 1$. Let $a_{ni} = \frac{\boldsymbol{\gamma}_{\mathbf{q},j}(\mathbf{x})'\mathbf{\Pi}_j(\mathbf{x}_i)}{\sqrt{\Omega_j(\mathbf{x})}}$. For all $t > 0$,

$$\begin{aligned}\mathbb{E}_n[\mathbb{E}[a_{ni}^2\varepsilon_i^2\mathbb{1}\{|a_{ni}\varepsilon_i/\sqrt{n}| > t\}]] &\leq \mathbb{E}[a_{ni}^2]\sup_{\mathbf{x}\in\mathcal{X}}\mathbb{E}\left[\varepsilon_i^2\mathbb{1}\{|\varepsilon_i| > t\sqrt{n}/|a_{ni}|\}\middle|\mathbf{x}_i = \mathbf{x}\right] \\ &\lesssim \sup_{\mathbf{x}\in\mathcal{X}}\mathbb{E}\left[\varepsilon_i^2\mathbb{1}\{|\varepsilon_i| > t\sqrt{n}/|a_{ni}|\}\middle|\mathbf{x}_i = \mathbf{x}\right]\end{aligned}$$

where the last line follows from Lemma A.1. Since $|a_{ni}| \lesssim h^{-\frac{d}{2}}$ and $\frac{\log n}{nh^d} = o(1)$, $\sqrt{n}/|a_{ni}| \to \infty$ as $n \to \infty$, and the last line goes to 0 as $n \to \infty$. □

Proof of Theorem II.4. Suppose that the conditions in **(i)** holds. In light of Lemma A.1, it suffices to show $\|\widehat{\mathbf{\Sigma}}_j - \mathbf{\Sigma}_j\| = o_{\mathbb{P}}(h^d)$. Notice that $\widehat{\mathbf{\Sigma}}_j - \mathbf{\Sigma}_j = \mathbb{E}_n[(\widehat{\varepsilon}_{i,j}^2 - \varepsilon_i^2)\mathbf{\Pi}_j(\mathbf{x}_i)\mathbf{\Pi}_j(\mathbf{x}_i)'] + \mathbb{E}_n[\varepsilon_i^2\mathbf{\Pi}_j(\mathbf{x}_i)\mathbf{\Pi}_j(\mathbf{x}_i)'] - \mathbf{\Sigma}_j$.

To control the second term, let $\mathbf{L}_j(\mathbf{x}_i) := \mathbf{W}_j^{-1/2}\mathbf{\Pi}_j(\mathbf{x}_i)$ be the normalized basis where $\mathbf{W}_j = \mathbf{Q}_m$ for $j = 0$, $\mathbf{W}_j = \mathbf{Q}_{\tilde{m}}$ for $j = 1$ and $\mathbf{W}_j = \mathrm{diag}\{\mathbf{Q}_m, \mathbf{Q}_{\tilde{m}}\}$ for $j = 2, 3$. Introduce a sequence of positive numbers: $M_n^2 \asymp \frac{K^{1+1/\nu}n^{1/(2+\nu)}}{(\log n)^{1/(2+\nu)}}$, and write $\mathbf{H}_j(\mathbf{x}_i) = \varepsilon_i^2\mathbf{L}_j(\mathbf{x}_i)\mathbf{L}_j(\mathbf{x}_i)'\mathbb{1}\{\|\varepsilon_i^2\mathbf{L}_j(\mathbf{x}_i)\mathbf{L}_j(\mathbf{x}_i)'\| \leq M_n^2\}$, and $\mathbf{T}_j(\mathbf{x}_i) = \varepsilon_i^2\mathbf{L}_j(\mathbf{x}_i)\mathbf{L}_j(\mathbf{x}_i)'\mathbb{1}\{\|\varepsilon_i^2\mathbf{L}_j(\mathbf{x}_i)\mathbf{L}_j(\mathbf{x}_i)'\| > M_n^2\}$. Regarding the truncated term, by construction, $\|\mathbf{H}_j(\mathbf{x}_i)\| \leq M_n^2$. By Triangle Inequality and Jensen's inequality,

$\|\mathbf{H}_j(\mathbf{x}_i) - \mathbb{E}[\mathbf{H}_j(\mathbf{x}_i)]\| \leq 2M_n^2$. In addition, by Assumption II.1,

$$\begin{aligned}\mathbb{E}[(\mathbf{H}_j(\mathbf{x}_i) - \mathbb{E}[\mathbf{H}_j(\mathbf{x}_i)])^2] &\leq M_n^2 \mathbb{E}[\varepsilon_i^2 \mathbf{L}_j(\mathbf{x}_i)\mathbf{L}_j(\mathbf{x}_i)' \mathbb{1}\{\|\varepsilon_i^2 \mathbf{L}_j(\mathbf{x}_i)\mathbf{L}_j(\mathbf{x}_i)'\| \leq M_n^2\}] \\ &\lesssim M_n^2 \mathbb{E}[\mathbf{L}_j(\mathbf{x}_i)\mathbf{L}_j(\mathbf{x}_i)']\end{aligned}$$

where the inequalities are in the sense of semi-definite matrices. Hence, $\|\mathbb{E}[(\mathbf{H}_j(\mathbf{x}_i) - \mathbb{E}[\mathbf{H}_j(\mathbf{x}_i)])^2]\| \lesssim M_n^2$. Let $\vartheta_n^2 = (\log n)^{\frac{\nu}{2+\nu}}/(n^{\frac{\nu}{2+\nu}} h^d)$. By an inequality of Tropp (2012) for independent matrices, we have for all $t > 0$,

$$\mathbb{P}[\|\mathbb{E}_n(\mathbf{H}_j(\mathbf{x}_i) - \mathbb{E}[\mathbf{H}_j(\mathbf{x}_i)])\| > \vartheta_n t] \leq \exp\Big\{ \log n \Big(1 - \frac{\vartheta_n^2 n t^2/2}{M_n^2 \log n (1 + \vartheta_n t/3)} \Big) \Big\}$$

where $M_n^2 \log n \, \vartheta_n^{-2} n^{-1} \asymp (\log n)^{\frac{1}{2+\nu}}/(n^{\frac{1}{2+\nu}} h^{d/\nu}) = o(1)$ and $\vartheta_n = o(1)$. Hence, we have $\|\mathbb{E}_n(\mathbf{H}_j(\mathbf{x}_i) - \mathbb{E}[\mathbf{H}_j(\mathbf{x}_i)])\| \lesssim_{\mathbb{P}} \vartheta_n = o_{\mathbb{P}}(1)$.

Regarding the tails, by Lemma A.1, $\|\mathbf{T}_j(\mathbf{x}_i)\| \lesssim h^{-d} \varepsilon_i^2 \mathbb{1}\{\varepsilon_i^2 \gtrsim M_n^2 h^d\}$. Then, by Triangle inequality and Jensen's inequality,

$$\mathbb{E}[\|\mathbb{E}_n(\mathbf{T}_j(\mathbf{x}_i) - \mathbb{E}[\mathbf{T}_j(\mathbf{x}_i)])\|] \lesssim \frac{2h^{-d(1+\nu/2)} \mathbb{E}[|\varepsilon_i|^{2+\nu} \mathbb{1}\{|\varepsilon_i| \gtrsim M_n \sqrt{h^d}\}]}{M_n^\nu} \lesssim \vartheta_n.$$

By Markov's inequality, $\|\mathbb{E}_n(\mathbf{T}_j(\mathbf{x}_i) - \mathbb{E}[\mathbf{T}_j(\mathbf{x}_i)])\| \lesssim_{\mathbb{P}} \vartheta_n$. Since $\|\mathbf{W}_j^{1/2}\| \lesssim h^{d/2}$ and $\|\mathbf{W}_j^{-1/2}\| \lesssim h^{-d/2}$, we conclude that $\|\mathbb{E}_n[\mathbf{\Pi}_j(\mathbf{x}_i)\mathbf{\Pi}_j(\mathbf{x}_i)'\varepsilon_i^2] - \mathbf{\Sigma}_j\| \lesssim_{\mathbb{P}} h^d \vartheta_n = o_{\mathbb{P}}(h^d)$.

On the other hand, note that

$$\begin{aligned}&\|\mathbb{E}_n[(\widehat{\varepsilon}_{i,j}^2 - \varepsilon_i^2)\mathbf{\Pi}_j(\mathbf{x}_i)\mathbf{\Pi}_j(\mathbf{x}_i)']\| \\ \leq\; &\max_{1\leq i\leq n} |\theta(\mathbf{x}_i) - \widehat{\theta}_j(\mathbf{x}_i)|^2 \|\mathbb{E}_n[\mathbf{\Pi}_j(\mathbf{x}_i)\mathbf{\Pi}_j(\mathbf{x}_i)']\| \\ &+ \max_{1\leq i\leq n} |\theta(\mathbf{x}_i) - \widehat{\theta}_j(\mathbf{x}_i)| \left(\|\mathbb{E}_n[\mathbf{\Pi}_j(\mathbf{x}_i)\mathbf{\Pi}_j(\mathbf{x}_i)']\| + \|\mathbb{E}_n[\mathbf{\Pi}_j(\mathbf{x}_i)\mathbf{\Pi}_j(\mathbf{x}_i)'\varepsilon_i^2]\| \right)\end{aligned}$$

where the last line follows from the fact that $2|a| \leq 1 + a^2$. By Lemma A.1, Theorem SA-4.1 in the online appendix to Cattaneo, Farrell, and Feng (2018a), $\max_{1\leq i\leq n} |\theta(\mathbf{x}_i) - \widehat{\theta}_j(\mathbf{x}_i)| = o_{\mathbb{P}}(1)$, $\|\mathbb{E}_n[\mathbf{\Pi}_j(\mathbf{x}_i)\mathbf{\Pi}_j(\mathbf{x}_i)']\| \lesssim_{\mathbb{P}} h^d$ and $\|\mathbb{E}_n[\mathbf{\Pi}_j(\mathbf{x}_i)\mathbf{\Pi}_j(\mathbf{x}_i)'\varepsilon_i^2]\| \lesssim_{\mathbb{P}} h^d$. Thus, we conclude that $\|\widehat{\mathbf{\Sigma}}_j - \mathbf{\Sigma}_j\| = o_{\mathbb{P}}(h^d)$. The proof under the conditions in **(ii)** is similar. □

Proof of Lemma II.2. First, suppose that the conditions in (i) hold. In Theorem SA-4.2 of the online appendix to Cattaneo, Farrell, and Feng (2018a), we establish that $\sup_{\mathbf{x}\in\mathcal{X}} \big|\widehat{\Omega}_0(\mathbf{x}) - \Omega_0(\mathbf{x})\big| \lesssim_{\mathbb{P}} n^{-\frac{1}{2}} h^{-\frac{3d}{2} - 2[\mathbf{q}]} \big[(\log n)^{\frac{1}{2}} + n^{\frac{1}{2+\nu}} (\log n)^{\frac{\nu}{4+2\nu}} + \sqrt{n} h^{\frac{d}{2}+m}\big]$ and, for $j = 1, 2, 3$, $\sup_{\mathbf{x}\in\mathcal{X}} \big|\widehat{\Omega}_j(\mathbf{x}) - \Omega_j(\mathbf{x})\big| \lesssim_{\mathbb{P}} n^{-\frac{1}{2}} h^{-\frac{3d}{2} - 2[\mathbf{q}]} \big[(\log n)^{\frac{1}{2}} + n^{\frac{1}{2+\nu}} (\log n)^{\frac{\nu}{4+2\nu}} +$

$\sqrt{n}h^{\frac{d}{2}+m+\varrho}]$. Then, for $j=0,1,2,3$,

$$\begin{aligned}
&\sup_{\mathbf{x}\in\mathcal{X}}\left|\frac{\widehat{\partial^{\mathbf{q}}\theta}_j(\mathbf{x})-\partial^{\mathbf{q}}\theta(\mathbf{x})}{\Omega_j^{1/2}(\mathbf{x})/\sqrt{n}}-\frac{\widehat{\partial^{\mathbf{q}}\theta}_j(\mathbf{x})-\partial^{\mathbf{q}}\theta(\mathbf{x})}{\widehat{\Omega}_j^{1/2}(\mathbf{x})/\sqrt{n}}\right|\\
&\le\sup_{\mathbf{x}\in\mathcal{X}}\frac{\sqrt{n}\big|\widehat{\partial^{\mathbf{q}}\theta}_j(\mathbf{x})-\partial^{\mathbf{q}}\theta(\mathbf{x})\big|\big|\widehat{\Omega}_j(\mathbf{x})-\Omega_j(\mathbf{x})\big|}{\Omega_j^{1/2}(\mathbf{x})\widehat{\Omega}_j(\mathbf{x})+\Omega_j(\mathbf{x})\widehat{\Omega}_j^{1/2}(\mathbf{x})}\\
&\lesssim_{\mathbb{P}}\sqrt{n}h^{3d/2+3[\mathbf{q}]}\sup_{\mathbf{x}\in\mathcal{X}}\big|\widehat{\partial^{\mathbf{q}}\theta}_j(\mathbf{x})-\partial^{\mathbf{q}}\theta(\mathbf{x})\big|\sup_{\mathbf{x}\in\mathcal{X}}\big|\widehat{\Omega}_j(\mathbf{x})-\Omega_j(\mathbf{x})\big|=o_{\mathbb{P}}(r_n^{-1}),
\end{aligned}$$

where the result follows from Lemma A.1, Theorem SA-4.1 in Cattaneo, Farrell, and Feng (2018a), the uniform convergence rate of $\widehat{\Omega}_j(\mathbf{x})$, and the rate conditions imposed.

The result under the conditions in **(ii)** follows similarly. □

Proof of Theorem II.5. We first prove the following general lemma. Let $TV_{\mathcal{X}}(g(\cdot))$ denote the total variation of $g(\cdot)$ on $\mathcal{X}\subseteq\mathbb{R}$.

Lemma A.3 (Kernel-Based KMT Coupling)**.** *Suppose* $\{(x_i,\varepsilon_i):1\le i\le n\}$ *are i.i.d., with* $x_i\in\mathcal{X}\subseteq\mathbb{R}$ *and* $\sigma_i^2:=\sigma^2(x_i)=\mathbb{E}[\varepsilon_i^2|x_i]$*. Let* $\{A(x):=\mathbb{G}_n[\mathscr{K}(x,x_i)\varepsilon_i],x\in\mathcal{X}\}$ *be a stochastic process with* $\mathscr{K}(\cdot,\cdot):\mathbb{R}\times\mathbb{R}\mapsto\mathbb{R}$ *an* n*-varying kernel function possibly depending on* $\mathbf{X}$*. Assume one of the following holds:*

(i) $\sup_{x\in\mathcal{X}}\mathbb{E}[|\varepsilon_i|^{2+\nu}|x_i=x]<\infty$*, for some* $\nu>0$*, and*

$$\sup_{x\in\mathcal{X}}\max_{1\le i\le n}|\mathscr{K}(x,x_i)|=o_{\mathbb{P}}(r_n^{-1}n^{-\frac{1}{2+\nu}+\frac{1}{2}}),$$
$$\sup_{x\in\mathcal{X}}TV_{\mathcal{X}}(\mathscr{K}(x,\cdot))=o(r_n^{-1}n^{-\frac{1}{2+\nu}+\frac{1}{2}});\quad or$$

(ii) $\sup_{x\in\mathcal{X}}\mathbb{E}[|\varepsilon_i|^3\exp(|\varepsilon_i|)|x_i=x]<\infty$ *and*

$$\sup_{x\in\mathcal{X}}\max_{1\le i\le n}|\mathscr{K}(x,x_i)|=o_{\mathbb{P}}(r_n^{-1}(\log n)^{-1}\sqrt{n}),$$
$$\sup_{x\in\mathcal{X}}TV_{\mathcal{X}}(\mathscr{K}(x,\cdot))=o(r_n^{-1}(\log n)^{-1}\sqrt{n}).$$

Then, on a sufficiently rich probability space, there exists a copy $A'(\cdot)$ *of* $A(\cdot)$*, and an i.i.d. sequence* $\{\zeta_i:1\le i\le n\}$ *of standard Normal random variables such that* $A(x)=_d\mathbb{G}_n\big[\mathscr{K}(x,x_i)\sigma_i\zeta_i\big]+o_{\mathbb{P}}(r_n^{-1})$ *in* $\mathcal{L}^\infty(\mathcal{X})$*.*

Proof. Suppose the conditions in **(i)** hold. Let $\{x_{(i)}:1\le i\le n\}$ be the order statistics of $\{x_i:1\le i\le n\}$, such that $x_{(1)}\le x_{(2)}\le\cdots\le x_{(n)}$, which also induces the concomitants $\{\varepsilon_{[i]}:1\le i\le n\}$ and $\{\sigma_{[i]}^2=\sigma^2(x_{(i)}):1\le i\le n\}$. Conditional on $\mathbf{X}$, $\{\varepsilon_{[i]}:1\le i\le n\}$ is still an independent mean zero sequence with $\mathbb{V}[\varepsilon_{[i]}|\mathbf{X}]=\sigma_{[i]}^2$.

By Sakhanenko (1991, Corollary 5), there exists a sequence of i.i.d standard normal random variables $\{\zeta_{[i]} : 1 \leq i \leq n\}$ such that $\max_{1\leq l\leq n}|S_{l,n}| \lesssim_{\mathbb{P}} n^{\frac{1}{2+\nu}}$, where $S_{l,n} := \sum_{i=1}^{l}(\varepsilon_{[i]} - \sigma_{[i]}\zeta_{[i]})$. Then, using summation by parts,

$$\begin{aligned}
&\sup_{x\in\mathcal{X}}\Big|\sum_{i=1}^{n}\mathscr{K}(x,x_{(i)})(\varepsilon_{[i]}-\sigma_{[i]}\zeta_{[i]})\Big| \\
&= \sup_{x\in\mathcal{X}}\Big|\mathscr{K}(x,x_{(n)})S_{n,n} - \sum_{i=1}^{n-1}S_{i,n}\left(\mathscr{K}(x,x_{(i+1)})-\mathscr{K}(x,x_{(i)})\right)\Big| \\
&\leq \left(\sup_{x\in\mathcal{X}}\max_{1\leq i\leq n}|\mathscr{K}(x,x_i)| + \sup_{x\in\mathcal{X}}\sum_{i=1}^{n-1}\Big|\mathscr{K}(x,x_{(i+1)})-\mathscr{K}(x,x_{(i)})\Big|\right)\max_{1\leq l\leq n}|S_{l,n}|.
\end{aligned}$$

Note that $\sum_{i=1}^{n-1}\big|\mathscr{K}(x,x_{(i+1)})-\mathscr{K}(x,x_{(i)})\big| \leq TV_{\mathcal{X}}(\mathscr{K}(x,\cdot))$. Thus, under the conditions given in **(i)**, $A(x) =_d \mathbb{G}_n[\mathscr{K}(x,x_i)\sigma_i\zeta_i] + o_{\mathbb{P}}(r_n^{-1})$.

When condition **(ii)** holds, the proof is the same except that under the stronger moment restriction, $\max_{1\leq l\leq n}|S_{l,n}| \lesssim_{\mathbb{P}} \log n$ by Sakhanenko (1985, Theorem 1). □

To prove Theorem II.5, for each $j = 0,1,2,3$, let $\mathscr{K}(x,u) = \boldsymbol{\gamma}_{\mathbf{q},j}(x)'\mathbf{\Pi}_j(u)/\sqrt{\Omega_j(x)}$ and observe that $\sup_{x\in\mathcal{X}}\sup_{u\in\mathcal{X}}|\mathscr{K}(x,u)| \lesssim h^{-d/2}$, by Lemma A.1, and the uniform bound on the total variation of $\mathscr{K}(x,u)$ can also be verified easily. Alternatively, simply note that

$$\begin{aligned}
&\Big|\sum_{i=1}^{n-1}S_{i,n}\left(\mathscr{K}(x,x_{(i+1)})-\mathscr{K}(x,x_{(i)})\right)\Big| \\
&\leq \Big\|\frac{\boldsymbol{\gamma}_{\mathbf{q},j}(x)'}{\sqrt{\Omega_j(x)}}\Big\|_\infty\Big\|\sum_{i=1}^{n-1}S_{i,n}(\mathbf{\Pi}_j(x_{(i+1)})-\mathbf{\Pi}_j(x_{(i)}))\Big\|_\infty.
\end{aligned}$$

By Assumption II.3 and Lemma A.1, $\sup_{x\in\mathcal{X}}\|\boldsymbol{\gamma}_{\mathbf{q},j}(x)'/\sqrt{\Omega_j(x)}\|_\infty \lesssim h^{-d/2}$. Also, write the lth element of $\mathbf{\Pi}_j(\cdot)$ as $\pi_{j,l}(\cdot)$. Then, $\max_{1\leq l\leq K_j}\big|\sum_{i=1}^{n-1}\big(\pi_{j,l}(x_{(i+1)}) - \pi_{j,l}(x_{(i)})\big)S_{l,n}\big| \leq \max_{1\leq l\leq K_j}\sum_{i=1}^{n-1}\big|\pi_{j,l}(x_{(i+1)})-\pi_{j,l}(x_{(i)})\big|\max_{1\leq\ell\leq n}\big|S_{\ell,n}\big|$. By Assumption II.3 and II.5, $\max_{1\leq l\leq K_j}\sum_{i=1}^{n-1}|\pi_{j,l}(x_{(i+1)})-\pi_{j,l}(x_{(i)})| \lesssim 1$. Thus, using Lemma A.3, under the corresponding moment conditions and rate restrictions, there exists independent standard normal $\{\zeta_i : 1\leq i\leq n\}$ such that $\mathbb{G}_n[\mathscr{K}(x,x_i)\varepsilon_i] =_d z_j(x) + o_{\mathbb{P}}(r_n^{-1})$.

To finish the proof Theorem II.5, note that

$$z_j(\mathbf{x}) =_{d|\mathbf{X}} \frac{\boldsymbol{\gamma}_{\mathbf{q},j}(\mathbf{x})'}{\sqrt{\Omega_j(\mathbf{x})}}\mathbf{\Sigma}_j^{1/2}\mathbf{N}_{K_j} + \frac{\boldsymbol{\gamma}_{\mathbf{q},j}(\mathbf{x})'}{\sqrt{\Omega_j(\mathbf{x})}}(\bar{\mathbf{\Sigma}}_j^{1/2} - \mathbf{\Sigma}_j^{1/2})\mathbf{N}_{K_j}$$

where $\mathbf{N}_{K_j}$ is a K_j-dimensional standard normal vector (independent of $\mathbf{X}$) and "$=_{d|\mathbf{X}}$" denotes that two processes have the same conditional distribution given $\mathbf{X}$.

Regarding the second term, by Gaussian Maximal Inequality (see Chernozhukov, Lee, and Rosen, 2013, Lemma 13), $\mathbb{E}\big[\big\|(\bar{\boldsymbol{\Sigma}}_j^{1/2}-\boldsymbol{\Sigma}_j^{1/2})\mathbf{N}_{K_j}\big\|_\infty\big|\mathbf{X}\big]\lesssim\sqrt{\log n}\,\big\|\bar{\boldsymbol{\Sigma}}_j^{1/2}-\boldsymbol{\Sigma}_j^{1/2}\big\|$. By the same argument used in the proof of Cattaneo, Farrell, and Feng (2018a, Lemma SA-2.1), $\big\|\bar{\boldsymbol{\Sigma}}_j-\boldsymbol{\Sigma}_j\big\|\lesssim_{\mathbb{P}} h^d(\log n/(nh^d))^{1/2}$. Then, by Bhatia (2013, Theorem X.1.1), $\big\|\bar{\boldsymbol{\Sigma}}_j^{1/2}-\boldsymbol{\Sigma}_j^{1/2}\big\|\lesssim_{\mathbb{P}} h^{d/2}(\log n/(nh^d))^{1/4}$. For $j=0,1$, a sharper bound is available: by Bhatia (2013, Theorem X.3.8) and Lemma A.1, $\|\bar{\boldsymbol{\Sigma}}_j^{1/2}-\boldsymbol{\Sigma}_j^{1/2}\|\leq\lambda_{\min}(\boldsymbol{\Sigma}_j)^{-1/2}\|\bar{\boldsymbol{\Sigma}}_j-\boldsymbol{\Sigma}_j\|\lesssim_{\mathbb{P}} h^{d/2}\sqrt{\log n/(nh^d)}$. Thus, combining these results,

$$\mathbb{E}\Big[\sup_{\mathbf{x}\in\mathcal{X}}\Big|\frac{\boldsymbol{\gamma}_{\mathbf{q},j}(\mathbf{x})'}{\sqrt{\Omega_j(\mathbf{x})}}(\bar{\boldsymbol{\Sigma}}_j^{\frac{1}{2}}-\boldsymbol{\Sigma}_j^{\frac{1}{2}})\mathbf{N}_{K_j}\Big|\,\Big|\mathbf{X}\Big]\lesssim_{\mathbb{P}} h^{-\frac{d}{2}}\sqrt{\log n}\,\big\|\bar{\boldsymbol{\Sigma}}_j^{\frac{1}{2}}-\boldsymbol{\Sigma}_j^{\frac{1}{2}}\big\|=o_{\mathbb{P}}(r_n^{-1})$$

where the last equality holds by the additional rate restriction given in the theorem (for $j=0,1$, no additional restriction is needed). The results follow from Markov inequality and Dominated Convergence Theorem. □

Proof of Theorem II.6. It suffices to verify the conditions in Lemma 39 of Belloni, Chernozhukov, Chetverikov, and Fernandez-Val (2018). For $j=0,1,2,3$, define $\boldsymbol{\xi}_i=\frac{1}{\sqrt{n}}\boldsymbol{\Pi}_j(\mathbf{x}_i)\varepsilon_i$. Hence, $\{\boldsymbol{\xi}_i:1\leq i\leq n\}$ is an i.i.d. sequence of K_j-dimensional random vectors, and $\sum_{i=1}^n\mathbb{E}[\|\boldsymbol{\xi}_i\|^2\|\boldsymbol{\xi}_i\|_\infty]=\mathbb{E}\big[\big\|\boldsymbol{\Pi}_j(\mathbf{x}_i)\varepsilon_i\big\|^2\big\|\boldsymbol{\Pi}_j(\mathbf{x}_i)\varepsilon_i\big\|_\infty\big]/\sqrt{n}\lesssim n^{-1/2}$ using Assumption II.3, the moment condition imposed in the theorem, and Lemma A.1. On the other hand, let $\{\mathbf{g}_i:1\leq i\leq n\}$ be a sequence of independent Gaussian vectors with mean zero and variance $\frac{1}{n}\boldsymbol{\Sigma}_j$. Then, by properties of Gaussian random variables and Lemma A.1, $(\mathbb{E}[\|\mathbf{g}_i\|_\infty^2])^{1/2}\lesssim\sqrt{\log(n)/n}$, and $\sum_{i=1}^n(\mathbb{E}[\|\mathbf{g}_i\|^4])^{1/2}\lesssim\operatorname{trace}\Big(\sum_{i=1}^n\mathbb{E}[\boldsymbol{\xi}_i\boldsymbol{\xi}_i']\Big)\lesssim 1$. Thus, $L_n:=\sum_{i=1}^n\mathbb{E}[\|\boldsymbol{\xi}_i\|^2\|\boldsymbol{\xi}_i\|_\infty]+\sum_{i=1}^n\mathbb{E}[\|\mathbf{g}_i\|^2\|\mathbf{g}_i\|_\infty]\lesssim\sqrt{\frac{\log(n)}{n}}$. Then, there exists a K_j-dimensional normal vector $\mathbf{N}_{K_j}$ with variance equal to $\boldsymbol{\Sigma}_j$ such that for any $t>0$,

$$\mathbb{P}\Big(\Big\|\sum_{i=1}^n\boldsymbol{\xi}_i-\mathbf{N}_{K_j}\Big\|_\infty>\frac{3h^{\frac{d}{2}}t}{r_n}\Big)\leq\min_{\tau\geq 0}\Big(2\mathbb{P}(\|\mathbf{Z}\|_\infty>\tau)+\frac{r_n^3L_n\tau^2}{h^{\frac{3d}{2}}t^3}\Big)\lesssim\frac{r_n^3(\log n)^{\frac{3}{2}}}{\sqrt{nh^{3d}}t^3}$$

where $\mathbf{Z}$ is a K_j-dimensional standard Gaussian vector, and the second inequality follows by setting $\tau=C\sqrt{\log n}$ for a sufficiently large $C>0$. Using $\sup_{x\in\mathcal{X}}\|\boldsymbol{\gamma}_{\mathbf{q},j}(x)'/\sqrt{\Omega_j(x)}\|_\infty\lesssim h^{-d/2}$ again, the result follows. □

Proof of Theorem II.7. For each $j=0,1,2,3$,

$$\widehat{Z}_j(\mathbf{x})-Z_j(\mathbf{x})=\Big(\frac{\widehat{\boldsymbol{\gamma}}_{\mathbf{q},j}(\mathbf{x})}{\widehat{\Omega}_j^{1/2}(\mathbf{x})}-\frac{\boldsymbol{\gamma}_{\mathbf{q},j}(\mathbf{x})'}{\Omega_j^{1/2}(\mathbf{x})}\Big)\widehat{\boldsymbol{\Sigma}}_j^{\frac{1}{2}}\mathbf{N}_{K_j}+\frac{\boldsymbol{\gamma}_{\mathbf{q},j}(\mathbf{x})'}{\sqrt{\Omega_j(\mathbf{x})}}[\widehat{\boldsymbol{\Sigma}}_j^{\frac{1}{2}}-\boldsymbol{\Sigma}_j^{\frac{1}{2}}]\mathbf{N}_{K_j}\,.$$

Conditional on the data, each term in the above is a mean-zero Gaussian process. The desired results can be obtained by applying Gaussian maximal inequality to each term as in the proof of Lemma A.3. □

Proof of Theorem II.8. In view of Theorem II.5 and II.6, there exists a sequence of constants η_n such that $\eta_n = o(1)$ and $\mathbb{P}(|\sup_{\mathbf{x}\in\mathcal{X}} |\widehat{T}_j(\mathbf{x})| - \sup_{\mathbf{x}\in\mathcal{X}} |Z_j(\mathbf{x})|| > \eta_n/r_n) = o(1)$. Therefore, for any $u \in \mathbb{R}$,

$$\begin{aligned}\mathbb{P}\Big[\sup_{\mathbf{x}\in\mathcal{X}} |\widehat{T}_j(\mathbf{x})| \leq u\Big] &\leq \mathbb{P}\Big[\Big\{\sup_{\mathbf{x}\in\mathcal{X}} |\widehat{T}_j(\mathbf{x})| \leq u\Big\} \cap \Big\{\Big|\sup_{\mathbf{x}\in\mathcal{X}} |\widehat{T}_j(\mathbf{x})| - \sup_{\mathbf{x}\in\mathcal{X}} |Z_j(\mathbf{x})|\Big| \leq \eta_n/r_n\Big\}\Big] \\ &\quad + \mathbb{P}\Big[\Big\{\Big|\sup_{\mathbf{x}\in\mathcal{X}} |\widehat{T}_j(\mathbf{x})| - \sup_{\mathbf{x}\in\mathcal{X}} |Z_j(\mathbf{x})|\Big| > \eta_n/r_n\Big\}\Big] \\ &\leq \mathbb{P}\left[\sup_{\mathbf{x}\in\mathcal{X}} |Z_j(\mathbf{x})| \leq u + \eta_n/r_n\right] + o(1) \\ &\leq \mathbb{P}\left[\sup_{\mathbf{x}\in\mathcal{X}} |Z_j(\mathbf{x})| \leq u\right] + Cr_n^{-1}\eta_n\mathbb{E}[\sup_{\mathbf{x}\in\mathcal{X}} |Z_j(\mathbf{x})|] + o(1)\end{aligned}$$

for some constant $C > 0$ where the last line holds by the Anti-Concentration Inequality due to Chernozhukov, Chetverikov, and Kato (2014b). By Gaussian maximal inequality, $\mathbb{E}[\sup_{\mathbf{x}\in\mathcal{X}} |Z_j(\mathbf{x})|] \lesssim \sqrt{\log n}$. Since we assume $r_n = \sqrt{\log n}$, the two terms on the far right of the last line is $o(1)$ and do not depend on u. The reverse of the inequality follows similarly, and we conclude that $\sup_{u\in\mathbb{R}} \big|\mathbb{P}\big[\sup_{\mathbf{x}\in\mathcal{X}} |\widehat{T}_j(\mathbf{x})| \leq u\big] - \mathbb{P}\big[\sup_{\mathbf{x}\in\mathcal{X}} |Z_j(\mathbf{x})| \leq u\big]\big| = o(1)$. On the other hand, by Theorem II.7, $\widehat{Z}_j(\cdot)$ is approximated by the same Gaussian process conditional on the data. Thus, using the same argument given above, the result follows. □

Appendix B

Proof for Chapter III

Notation

We introduce more notation for this appendix. We employ standard empirical process notation: $\mathbb{E}_n[g(\mathbf{x}_i)] = \frac{1}{n}\sum_{i=1}^{n} g(\mathbf{x}_i)$, and $\mathbb{G}_n[g(\mathbf{x}_i)] = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(g(\mathbf{x}_i) - \mathbb{E}[g(\mathbf{x}_i)])$. In addition, we employ the notion of covering number extensively in the proofs. Specifically, given a measurable space $(S, \mathcal{S})$ and a suitably measurable class of functions $\mathcal{G}$ mapping S to $\mathbb{R}$ equipped with a measurable envelop function $\bar{G}(z) \geq \sup_{g\in\mathcal{G}} |g(z)|$. The *covering number* of $N(\mathcal{G}, L_2(Q), \varepsilon)$ is the minimal number of $L_2(Q)$-balls of radius ε needed to cover $\mathcal{G}$. The covering number of $\mathcal{G}$ relative to the envelop is denoted as $N(\mathcal{G}, L_2(Q), \varepsilon\|\bar{G}\|_{Q,2})$.

Given the partition $\widehat{\Delta}$ described in Chapter III, redefine $\widehat{\mathbf{b}}(x)$ as a *standardized rotated* basis for convenience of analysis. Specifically, for each $\alpha = 0, \ldots, p$, and $j = 1, \ldots, J$, the polynomial basis of degree α supported on $\widehat{\mathcal{B}}_j$ is rotated and rescaled:

$$\mathbb{1}_{\widehat{\mathcal{B}}_j}(x)x^{\alpha} \quad \mapsto \quad \sqrt{J}\cdot \mathbb{1}_{\widehat{\mathcal{B}}_j}(x)\Big(\frac{x - x_{(\lfloor (j-1)n/J\rfloor)}}{\hat{h}_j}\Big)^{\alpha},$$

where $\hat{h}_j = x_{(\lfloor jn/J\rfloor)} - x_{(\lfloor (j-1)n/J\rfloor)}$.

Given the random partition $\widehat{\Delta}$, we will use the notation $\mathbb{E}_{\widehat{\Delta}}[\cdot]$ to denote that the expectation is taken with the partition $\widehat{\Delta}$ understood as fixed. To further simplify notation, we let $\{\hat{\tau}_0 \leq \hat{\tau}_1 \leq \cdots \leq \hat{\tau}_J\}$ denote the empirical quantile sequence employed by $\widehat{\Delta}$. Accordingly, let $\{\tau_0 \leq \cdots \leq \tau_J\}$ be the population quantile sequence, i.e., $\tau_j = F^{-1}(j/J)$ for $0 \leq j \leq J$. Then $\Delta = \{\mathcal{B}_1, \ldots, \mathcal{B}_J\}$ denotes the partition based on population quantiles, i.e.,

$$\mathcal{B}_j = \begin{cases} \big[\tau_0, \tau_1\big) & \text{if } j = 1 \\ \big[\tau_{j-1}, \tau_j\big) & \text{if } j = 2, 3, \ldots, J-1\, . \\ \big[\tau_{J-1}, \tau_J\big] & \text{if } j = J \end{cases}$$

Let $h_j = F^{-1}(j/J) - F^{-1}((j-1)/J)$ be the width of $\mathcal{B}_j$. $\mathbf{b}_s(x)$ denotes the (smooth) binscatter basis based on the *nonrandom* partition Δ. Moreover, x_i's are collected in a matrix $\mathbf{X}=[x_1,\ldots,x_n]'$, all the data are collected in $\mathbf{D}=\{(y_i,x_i,\mathbf{w}_i'):1\leq i\leq n\}$.

We sometimes write $\mathbf{b}_s(x;\bar{\Delta})=(b_{s,1}(x;\bar{\Delta}),\ldots,b_{s,K_s}(x;\bar{\Delta}))'$ to emphasize a binscatter basis is constructed based on a particular partition $\bar{\Delta}$. Clearly, $\widehat{\mathbf{b}}_s(x)=\mathbf{b}_s(x;\widehat{\Delta})$ and $\mathbf{b}_s(x)=\mathbf{b}_s(x;\Delta)$.

The following expression of the coefficient estimators, also known as "backfitting" in statistics literature, will be convenient for theoretical analysis:

$$\widehat{\boldsymbol{\beta}}=(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'(\mathbf{Y}-\mathbf{W}\widehat{\boldsymbol{\gamma}}),\qquad \widehat{\boldsymbol{\gamma}}=(\mathbf{W}'\mathbf{M}_{\mathbf{B}}\mathbf{W})^{-1}(\mathbf{W}'\mathbf{M}_{\mathbf{B}}\mathbf{Y})$$

where $\mathbf{Y}=(y_1,\ldots,y_n)'$, $\mathbf{B}=(\widehat{\mathbf{b}}_s(x_1),\ldots,\widehat{\mathbf{b}}_s(x_n))'$, $\mathbf{W}=(\mathbf{w}_1,\ldots,\mathbf{w}_n)'$, $\mathbf{M}_{\mathbf{B}}=\mathbf{I}_n-\mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'$. It is well known that the least squares estimator provides a best linear approximation to the target function. For any given partition $\bar{\Delta}$, the population least squares estimator is defined as

$$\boldsymbol{\beta}_\mu(\bar{\Delta}):=\arg\min_{\boldsymbol{\beta}}\mathbb{E}[(\mu(x_i)-\mathbf{b}_s(x_i;\bar{\Delta})'\boldsymbol{\beta})^2].$$

Accordingly, $r_\mu(x;\bar{\Delta})=\mu(x)-\mathbf{b}_s(x;\bar{\Delta})'\boldsymbol{\beta}_\mu(\bar{\Delta})$ denotes the L_2 approximation error. We let $\widehat{\boldsymbol{\beta}}_\mu:=\boldsymbol{\beta}_\mu(\widehat{\Delta})$, $\boldsymbol{\beta}_\mu:=\boldsymbol{\beta}_\mu(\Delta)$, $\widehat{r}_\mu(x):=r_\mu(x;\widehat{\Delta})$ and $r_\mu(x):=r_\mu(x;\Delta)$.

In addition, we introduce the following matrices:

$$\begin{aligned}
&\widehat{\mathbf{Q}}:=\widehat{\mathbf{Q}}(\widehat{\Delta}):=\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'],\quad \mathbf{Q}:=\mathbf{Q}(\Delta):=\mathbb{E}[\mathbf{b}_s(x_i)\mathbf{b}_s(x_i)'],\\
&\widehat{\boldsymbol{\Sigma}}:=\widehat{\boldsymbol{\Sigma}}(\widehat{\Delta}):=\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'(y_i-\widehat{\mathbf{b}}_s(x_i)'\widehat{\boldsymbol{\beta}}-\mathbf{w}_i'\widehat{\boldsymbol{\gamma}})^2],\\
&\bar{\boldsymbol{\Sigma}}:=\bar{\boldsymbol{\Sigma}}(\widehat{\Delta}):=\mathbb{E}_n\Big[\mathbb{E}\Big[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'\epsilon_i^2\Big|\mathbf{X}\Big]\Big],\quad \boldsymbol{\Sigma}:=\boldsymbol{\Sigma}(\Delta):=\mathbb{E}\Big[\mathbf{b}_s(x_i)\mathbf{b}_s(x_i)'\epsilon_i^2\Big],\\
&\bar{\Omega}(x):=\bar{\Omega}(x;\widehat{\Delta}):=\widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\mathbf{Q}}^{-1}\bar{\boldsymbol{\Sigma}}\widehat{\mathbf{Q}}^{-1}\widehat{\mathbf{b}}_s^{(v)}(x),\quad\text{and}\\
&\Omega(x):=\Omega(x;\widehat{\Delta}):=\widehat{\mathbf{b}}_s^{(v)}(x)'\mathbf{Q}^{-1}\boldsymbol{\Sigma}\mathbf{Q}^{-1}\widehat{\mathbf{b}}_s^{(v)}(x).
\end{aligned}$$

All quantities with $\widehat{\ }$ or $\bar{\ }$ depend on the random partition $\widehat{\Delta}$, and those without any accents are nonrandom with the only exception of $\Omega(x)$, where the basis $\widehat{\mathbf{b}}_s^{(v)}(x)$ still depends on $\widehat{\Delta}$.

Finally, we let $\bar{f}=\sup_{x\in\mathcal{X}}f(x)$ and $\underline{f}=\inf_{x\in\mathcal{X}}f(x)$, and for any partition $\bar{\Delta}$ with J bins, we let $h_j(\bar{\Delta})$ denote the length of the jth bin in $\bar{\Delta}$. Then, we introduce a family of partitions:

$$\Pi=\left\{\bar{\Delta}:\frac{\max_{1\leq j\leq J}h_j(\bar{\Delta})}{\min_{1\leq j\leq J}h_j(\bar{\Delta})}\leq\frac{3\bar{f}}{\underline{f}}\right\}.\tag{B.1}$$

Intuitively, if a partition belongs to Π, then the lengths of its bins do not differ "too" much, a property usually referred to as *quasi-uniformity* in approximation theory.

Our first lemma shows that a quantile-spaced partition possesses this property with probability approaching one.

Preliminary Lemmas

We first prepare some preliminary lemmas. The detailed proofs can be found in the online appendix to Cattaneo, Crump, Farrell, and Feng (2019a).

Lemma B.1. *Under Assumption III.1, if* $\frac{J\log J}{n}=o(1)$ *and* $\frac{\log n}{J}=o(1)$*, then*

$$
\begin{array}{ll}
\text{(i)} & \max_{1\le j\le J}|\hat{h}_j-h_j|\lesssim_{\mathbb{P}} J^{-1}\sqrt{J\log J/n}, \\
 & \widehat{\Delta}\in\Pi \quad \textit{with probability approaching one}; \\
\text{(ii)} & \|\widehat{\mathbf{T}}_s\|_\infty\lesssim_{\mathbb{P}}1, \quad \|\widehat{\mathbf{T}}_s\|\lesssim_{\mathbb{P}}1, \\
 & \|\widehat{\mathbf{T}}_s-\mathbf{T}_s\|_\infty\lesssim_{\mathbb{P}}\sqrt{J\log J/n}, \quad \|\widehat{\mathbf{T}}_s-\mathbf{T}_s\|\lesssim_{\mathbb{P}}\sqrt{J\log J/n}; \\
\text{(iii)} & \sup_{x\in\mathcal{X}}\|\widehat{\mathbf{b}}_s^{(v)}(x)\|_0\le(p+1)^2, \quad \sup_{x\in\mathcal{X}}\|\widehat{\mathbf{b}}_s^{(v)}(x)\|\lesssim_{\mathbb{P}}J^{\frac{1}{2}+v}; \\
\text{(iv)} & 1\lesssim\lambda_{\min}(\mathbf{Q})\le\lambda_{\max}(\mathbf{Q})\lesssim 1, \quad \|\widehat{\mathbf{Q}}-\mathbf{Q}\|\lesssim_{\mathbb{P}}\sqrt{J\log J/n}, \\
 & \|\widehat{\mathbf{Q}}^{-1}\|_\infty\lesssim_{\mathbb{P}}1, \quad \|\widehat{\mathbf{Q}}^{-1}-\mathbf{Q}^{-1}\|_\infty\lesssim_{\mathbb{P}}\sqrt{J\log J/n}; \\
\text{(v)} & J^{1+2v}\lesssim_{\mathbb{P}}\inf_{x\in\mathcal{X}}\bar{\Omega}(x)\le\sup_{x\in\mathcal{X}}\bar{\Omega}(x)\lesssim_{\mathbb{P}}J^{1+2v}, \\
 & J^{1+2v}\lesssim\inf_{x\in\mathcal{X}}\Omega(x)\le\sup_{x\in\mathcal{X}}\Omega(x)\lesssim J^{1+2v}; \\
\text{(vi)} & \sup_{x\in\mathcal{X}}|\widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\boldsymbol{\beta}}_\mu-\mu^{(v)}(x)|\lesssim_{\mathbb{P}}J^{-p-1+v}, \\
 & \sup_{x\in\mathcal{X}}|\widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\mathbf{Q}}^{-1}\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{r}_\mu(x_i)]|\lesssim_{\mathbb{P}}J^{-p-1+v}\sqrt{J\log J/n}.
\end{array}
$$

Proof. The detailed proofs of these results are available in Section SA-6.1, SA-6.2, SA-6.3, SA-6.4, SA-6.5, and SA-6.6 of the online appendix to Cattaneo, Crump, Farrell, and Feng (2019b). Here we only provide the proof for (iii) and (iv).

The sparsity of the basis follows by construction. The upper bound on the maximum eigenvalue of $\mathbf{Q}$ follows from part (i) and (ii). Also, in view of part (i), the lower bound on the minimum eigenvalue of $\mathbf{Q}$ follows from Schumaker (2007, Theorem 4.41), by which the minimum eigenvalue of $\mathbf{Q}/J$ (the scaling factor dropped) is bounded by $\min_{1\le j\le J}h_j$ up to some universal constant. To show the bound on $\|\widehat{\mathbf{b}}_s^{(v)}(x)\|$, notice that when $s=0$, for any $x\in\mathcal{X}$ and any $j=1,\dots,J(p+1)$, $0\le\widehat{b}_{0,j}(x)\le\sqrt{J}$. Define $\varphi_{j,\alpha}(x)$ as

$$\varphi_{j,\alpha}(x)=\sqrt{J}\cdot\mathbb{1}_{\widehat{\mathcal{B}}_j}(x)\Big(\frac{x-\hat{\tau}_{j-1}}{\hat{h}_j}\Big)^\alpha, \quad 1\le\alpha\le p, \quad 1\le j\le J.$$

Since

$$\varphi^{(v)}_{j,\alpha} = \sqrt{J}\alpha(\alpha-1)\cdots(\alpha-v+1)\hat{h}_j^{-v}\mathbb{1}_{\widehat{\mathcal{B}}_j}(x)\Big(\frac{x-\hat{\tau}_{j-1}}{\hat{h}_j}\Big)^{\alpha-v} \lesssim \sqrt{J}\hat{h}_j^{-v},$$

the bound on $\|\widehat{\mathbf{b}}^{(v)}_s(x)\|$ simply follows from part (i) and (ii).

Now, we prove the convergence of $\widehat{\mathbf{Q}}$. In view of part (ii), it suffices to show the convergence of $\widehat{\mathbf{Q}}$ when $s=0$, i.e., $\|\mathbb{E}_n[\widehat{\mathbf{b}}_0(x_i)\widehat{\mathbf{b}}_0(x_i)'] - \mathbb{E}[\mathbf{b}_0(x_i)\mathbf{b}_0(x_i)']\| \lesssim_{\mathbb{P}} \sqrt{J\log J/n}$. By part (i), with probability approach 1, $\widehat{\Delta}$ ranges within the family of partitions Π. Let $\mathcal{A}_n$ denote the event on which $\widehat{\Delta}\in\Pi$. Thus, $\mathbb{P}(\mathcal{A}_n^c)=o(1)$. On $\mathcal{A}_n$,

$$\begin{aligned}&\Big\|\mathbb{E}_n[\widehat{\mathbf{b}}_0(x_i)\widehat{\mathbf{b}}_0(x_i)'] - \mathbb{E}_{\widehat{\Delta}}[\widehat{\mathbf{b}}_0(x_i)\widehat{\mathbf{b}}_0(x_i)']\Big\|\\ \leq&\sup_{\bar{\Delta}\in\Pi}\Big\|\mathbb{E}_n[\mathbf{b}_0(x_i;\bar{\Delta})\mathbf{b}_0(x_i;\bar{\Delta})'] - \mathbb{E}[\mathbf{b}_0(x_i;\bar{\Delta})\mathbf{b}_0(x_i;\bar{\Delta})']\Big\|.\end{aligned}$$

By the relation between matrix norms, the right-hand-side of the above inequality is further bounded by

$$\sup_{\bar{\Delta}\in\Pi}\Big\|\mathbb{E}_n[\mathbf{b}_0(x_i;\bar{\Delta})\mathbf{b}_0(x_i;\bar{\Delta})'] - \mathbb{E}[\mathbf{b}_0(x_i;\bar{\Delta})\mathbf{b}_0(x_i;\bar{\Delta})']\Big\|_\infty.$$

Let a_{kl} be a generic (k,l)th entry of the matrix inside the matrix norm, i.e.,

$$|a_{kl}| = \Big|\mathbb{E}_n[b_{0,k}(x_i;\bar{\Delta})b_{0,l}(x_i;\bar{\Delta})'] - \mathbb{E}\Big[b_{0,k}(x_i;\bar{\Delta})b_{0,l}(x_i;\bar{\Delta})'\Big]\Big|$$

Clearly, if $b_{0,k}(\cdot\,;\bar{\Delta})$ and $b_{0,l}(\cdot\,;\bar{\Delta})$ are basis functions with different supports, a_{kl} is zero. Now define the following function class

$$\mathcal{G} = \Big\{x\mapsto b_{0,k}(x;\bar{\Delta})b_{0,l}(x;\bar{\Delta}) : 1\leq k,l\leq J(p+1), \bar{\Delta}\in\Pi\Big\}.$$

For such a class, $\sup_{g\in\mathcal{G}}|g|_\infty \lesssim J$ and $\sup_{g\in\mathcal{G}}\mathbb{V}[g]\leq\sup_{g\in\mathcal{G}}\mathbb{E}[g^2]\lesssim J$ where the second result follows from the fact that the supports of $b_{0,k}(\cdot;\bar{\Delta})$ and $b_{0,l}(\cdot;\bar{\Delta})$ shrink at the rate of J^{-1}. In addition, each function in $\mathcal{G}$ is simply a dilation and translation of a polynomial function supported on $[0,1]$, plus a zero function, and the number of polynomial degree is finite. Then, by Proposition 3.6.12 of Giné and Nickl (2016), the collection $\mathcal{G}$ of such functions is of VC type, i.e., there exists some constant C_z and $z>6$ such that

$$N(\mathcal{G}, L_2(Q), \varepsilon\|\bar{G}\|_{L_2(Q)}) \leq \Big(\frac{C_z}{\varepsilon}\Big)^{2z},$$

for ε small enough where we take $\bar{G}=CJ$ for some constant $C>0$ large enough.

Theorem 6.1 of Belloni, Chernozhukov, Chetverikov, and Kato (2015),

$$\mathbb{E}\Big[\sup_{g\in\mathcal{G}}\Big|\sum_{i=1}^{n} g(x_i)-\sum_{i=1}^{n}\mathbb{E}[g(x_i)]\Big|\Big]\lesssim\sqrt{nJ\log J}+J\log J,$$

implying that

$$\sup_{g\in\mathcal{G}}\Big|\frac{1}{n}\sum_{i=1}^{n} g(x_i)-\mathbb{E}[g(x_i)]\Big|\lesssim_{\mathbb{P}}\sqrt{J\log J/n}.$$

Since any row or column of the matrix (a_{kl}) only contains a finite number of nonzero entries, only depending on p, the above result suffices to show that

$$\Big\|\mathbb{E}_n[\widehat{\mathbf{b}}_0(x_i)\widehat{\mathbf{b}}_0(x_i)']-\mathbb{E}_{\widehat{\Delta}}[\widehat{\mathbf{b}}_0(x_i)\widehat{\mathbf{b}}_0(x_i)']\Big\|\lesssim_{\mathbb{P}}\sqrt{J\log J/n}.$$

Next, let α_{kl} denote the (k,l)th entry of $\mathbb{E}_{\widehat{\Delta}}\Big[\widehat{\mathbf{b}}_0(x_i)\widehat{\mathbf{b}}_0(x_i)'\Big]/J-\mathbb{E}\Big[\mathbf{b}_0(x_i)\mathbf{b}_0(x_i)'\Big]/J$, where by dividing them by J we drop the normalizing constant for notational simplicity. By definition, it is either equal to zero, or can be rewritten as

$$\begin{aligned}\alpha_{kl}=&\int_{\widehat{\mathcal{B}}_j}\Big(\frac{x-\hat{\tau}_j}{\hat{h}_j}\Big)^{\ell}f(x)dx-\int_{\widehat{\mathcal{B}}_j}\Big(\frac{x-\tau_j}{h_j}\Big)^{\ell}f(x)dx\\ =&\hat{h}_j\int_0^1 z^{\ell}f(z\hat{h}_j+\hat{\tau}_j)dz-h_j\int_0^1 z^{\ell}f(zh_j+\tau_j)dz\\ =&(\hat{h}_j-h_j)\int_0^1 z^{\ell}f(z\hat{h}_j+\hat{\tau}_j)dz+h_j\int_0^1 z^{\ell}\Big(f(z\hat{h}_j+\hat{\tau}_j)-f(zh_j+\tau_j)\Big)dz \qquad \text{(B.2)}\end{aligned}$$

for some $1\leq j\leq J$ and $0\leq\ell\leq 2p$. By Assumption III.1 and Lemma SA2 of Calonico, Cattaneo, and Titiunik (2015), $\max_{1\leq j\leq J}f(\hat{\tau}_j)\lesssim 1$ and $\max_{1\leq j\leq J}|\hat{h}_j-h_j|\lesssim_{\mathbb{P}} J^{-1}\sqrt{J\log J/n}$. Also, Lemma SA2 of Calonico, Cattaneo, and Titiunik (2015) implies that

$$\sup_{z\in[0,1]}\max_{1\leq j\leq J}|\hat{\tau}_j+z\hat{h}_j-(\tau_j+zh_j)|\lesssim_{\mathbb{P}}\sqrt{J\log J/n}.$$

Since $f(\cdot)$ is uniformly continuous on $\mathcal{X}$, the second term in (B.2) is also $O_{\mathbb{P}}(J^{-1}\sqrt{J\log J/n})$. Again, using the sparsity structure of the matrix $[\alpha_{kl}]$, the above result suffices to show that $\|\mathbb{E}_{\widehat{\Delta}}[\widehat{\mathbf{b}}_0(x_i)\widehat{\mathbf{b}}_0(x_i)']-\mathbf{Q}\|\lesssim_{\mathbb{P}}\sqrt{J\log J/n}$.

Given the above fact, it follows that $\|\widehat{\mathbf{Q}}^{-1}\|\lesssim_{\mathbb{P}} 1$. Notice that $\widehat{\mathbf{Q}}$ and $\mathbf{Q}$ are banded matrices with finite band width. Then the bounds on $\|\widehat{\mathbf{Q}}\|_\infty$ and $\|\widehat{\mathbf{Q}}^{-1}-\mathbf{Q}^{-1}\|_\infty$ hold by Theorem 2.2 of Demko (1977). This completes the proof. □

Lemma B.2 (Uniform Convergence: Variance). *Suppose that Assumption III.1 holds.*

If $\frac{J^2 \log J}{n} = o(1)$ *and* $\frac{\log n}{J} = o(1)$*, then*

$$\sup_{x\in\mathcal{X}} |\widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\mathbf{Q}}^{-1}\mathbb{E}_n[\mathbf{b}_s(x_i)\epsilon_i]| \lesssim J^v\sqrt{J\log J/n}.$$

Proof. By Lemma B.1, $\sup_{x\in\mathcal{X}}\|\widehat{\mathbf{b}}^{(v)}(x)'\|_\infty \lesssim_{\mathbb{P}} J^{1/2+v}$, $\|\widehat{\mathbf{Q}}^{-1}\|_\infty \lesssim_{\mathbb{P}} 1$ and $\|\widehat{\mathbf{T}}_s\|_\infty \lesssim_{\mathbb{P}} 1$. Define a function class

$$\mathcal{G} = \Big\{(x_1, \epsilon_1) \mapsto b_{0,l}(x_1; \bar{\Delta})\epsilon_1 : 1 \le l \le J(p+1), \bar{\Delta} \in \Pi\Big\}.$$

Then, $\sup_{g\in\mathcal{G}}|g| \lesssim \sqrt{J}|\epsilon_1|$, and hence take an envelop $\bar{G} = C\sqrt{J}|\epsilon_1|$ for some C large enough. Moreover, $\sup_{g\in\mathcal{G}}\mathbb{V}[g] \lesssim 1$ and, as in the proof of Lemma B.1, $\mathcal{G}$ is of VC-type. By Proposition 6.1 of Belloni, Chernozhukov, Chetverikov, and Kato (2015),

$$\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^n g(x_i,\epsilon_i)\right| \lesssim_{\mathbb{P}} \sqrt{\frac{\log J}{n}} + \frac{J\log J}{n} \lesssim \sqrt{\frac{\log J}{n}},$$

and the desired result follows. □

Let $\{a_n : n \ge 1\}$ be a sequence of non-vanishing constants, which will be used later to characterize the strong approximation rate. The next theorem shows that under certain conditions the estimation of $\boldsymbol{\gamma}$ does not impact the asymptotic inference on the nonparametric component.

Lemma B.3 (Covariate Adjustment)**.** *Suppose that Assumption III.1 holds. If* $\frac{J\log J}{n} = o(1)$, $\frac{a_n}{\sqrt{J}} = o(1)$, $a_n\sqrt{n}J^{-2p-\frac{5}{2}} = o(1)$*, then*

$$\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\| = o_{\mathbb{P}}(a_n^{-1}\sqrt{J/n}), \quad \|\widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\mathbf{Q}}^{-1}\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\mathbf{w}_i']\|_\infty \lesssim_{\mathbb{P}} J^v \ \textit{ for each } x \in \mathcal{X}.$$

If, in addition, $\frac{J^2\log J}{n} \lesssim 1$*, then* $\sup_{x\in\mathcal{X}}\|\widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\mathbf{Q}}^{-1}\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\mathbf{w}_i']\|_\infty \lesssim_{\mathbb{P}} J^v$.

Proof. This result follows from Lemma SA-1, Lemma SA-2 of Cattaneo, Jansson, and Newey (2018b), Lemma 2 of Cattaneo, Jansson, and Newey (2018a) and Lemma B.1. See more details in Section SA-6.8 of the online appendix to Cattaneo, Crump, Farrell, and Feng (2019b). □

Using the previous results, the next lemma constructs the rate of uniform convergence for binscatter estimators.

Lemma B.4 (Uniform Convergence)**.** *Suppose that Assumption III.1 holds. If* $\sqrt{n}J^{-2p-\frac{5}{2}} = o(1)$ *and* $\frac{J^2\log J}{n} \lesssim 1$*, then*

$$\sup_{x\in\mathcal{X}}|\widehat{\mu}^{(v)}(x) - \mu^{(v)}(x)| \lesssim_{\mathbb{P}} J^v\sqrt{J\log J/n} + J^{-p-1+v}$$

Proof. Noticing that

$$\begin{aligned}\widehat{\mu}^{(v)}(x) - \mu^{(v)}(x) =&\widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\mathbf{Q}}^{-1}\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\epsilon_i] + \widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\mathbf{Q}}^{-1}\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{r}_\mu(x_i)]+ \\ &\Big(\widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\boldsymbol{\beta}}_\mu - \mu^{(v)}(x)\Big) - \widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\mathbf{Q}}^{-1}\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\mathbf{w}_i'](\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}).\end{aligned} \tag{B.3}$$

Then the result follows by Lemma B.1, B.2 and B.3. □

The last lemma shows that the proposed variance estimator is consistent.

Lemma B.5 (Variance Estimate)**.** *Suppose that Assumption III.1 holds. If* $\frac{J^2(\log J)^2}{n} = o(1)$ *and* $\sqrt{n}J^{-2p-\frac{5}{2}} = o(1)$*, then*

$$\begin{aligned}&\left\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right\| \lesssim_{\mathbb{P}} J^{-p-1} + \sqrt{J\log J/n^{1/2}}, \\ &\sup_{x\in\mathcal{X}} |\widehat{\Omega}(x) - \Omega(x)| \lesssim_{\mathbb{P}} J^{1+2v}\Big(J^{-p-1} + \sqrt{J\log J/n^{1/2}}\Big).\end{aligned}$$

Proof. Since $\widehat{\epsilon}_i := y_i - \widehat{\mathbf{b}}_s(x_i)'\widehat{\boldsymbol{\beta}} - \mathbf{w}_i'\widehat{\boldsymbol{\gamma}} = \epsilon_i + \mu(x_i) - \widehat{\mathbf{b}}_s(x_i)'\widehat{\boldsymbol{\beta}} - \mathbf{w}_i'(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) =: \epsilon_i + u_i$, we can write

$$\begin{aligned}&\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'\widehat{\epsilon}_i^2] - \mathbb{E}[\mathbf{b}_s(x_i)\mathbf{b}_s(x_i)'\sigma^2(x_i)] \\ =&\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'u_i^2] + 2\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'u_i\epsilon_i] + \mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'(\epsilon_i^2 - \sigma^2(x_i))] \\ &+ \Big(\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'\sigma^2(x_i)] - \mathbb{E}\Big[\mathbf{b}_s(x_i)\mathbf{b}_s(x_i)'\sigma^2(x_i)\Big]\Big) \\ =:&\mathbf{V}_1 + \mathbf{V}_2 + \mathbf{V}_3 + \mathbf{V}_4.\end{aligned}$$

Now we bound each term in the following.

Step 1: For $\mathbf{V}_1$, we further write $u_i = (\mu(x_i) - \widehat{\mathbf{b}}_s(x_i)'\widehat{\boldsymbol{\beta}}) - \mathbf{w}_i'(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) =: u_{i1} - u_{i2}$. Then $\mathbf{V}_1 = \mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'(u_{i1}^2 + u_{i2}^2 - 2u_{i1}u_{i2})] =: \mathbf{V}_{11} + \mathbf{V}_{12} - \mathbf{V}_{13}$. Since $\|2\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'u_{i1}u_{i2}]\| \leq \|\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'(u_{i1}^2 + u_{i2}^2)]\|$, it suffices to bound $\mathbf{V}_{11}$ and $\mathbf{V}_{12}$. For $\mathbf{V}_{11}$,

$$\|\mathbf{V}_{11}\| \leq \max_{1\leq i\leq n} |u_{i1}|^2\left\|\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)']\right\| \lesssim_{\mathbb{P}} \frac{J\log J}{n} + J^{-2(p+1)}$$

where the last inequality holds by Lemma B.1 and B.4. On the other hand,

$$\begin{aligned}\|\mathbf{V}_{12}\| &= \left\|\mathbb{E}_n\Big[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'\Big(\sum_\ell^d w_{i\ell}^2(\widehat{\gamma}_\ell - \gamma_\ell)^2 + \sum_{\ell\neq\ell'} w_{i\ell}w_{i\ell'}(\widehat{\gamma}_\ell - \gamma)(\widehat{\gamma}_{\ell'} - \gamma_{\ell'})\Big)\Big]\right\| \\ &\lesssim \left\|\mathbb{E}_n\Big[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'\Big(\sum_\ell^d w_{i\ell}^2(\widehat{\gamma}_\ell - \gamma_\ell)^2\Big)\Big]\right\|\end{aligned}$$

by CR-inequality. By Lemma B.3, $\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|^2 = o_{\mathbb{P}}(J/n)$. Then it suffices to show that for every $\ell = 1, \ldots, d$, $\|\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'w_{i\ell}^2]\| \lesssim_{\mathbb{P}} 1$. Under the conditions given in the theorem, this bound can be established using the argument that will be given in Step 3 and 4.

Step 2: For $\mathbf{V}_2$, we have $\mathbf{V}_2 = 2\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'\epsilon_i(u_{i1}-u_{i2})] =: \mathbf{V}_{21} - \mathbf{V}_{22}$. Then,

$$\|\mathbf{V}_{21}\| \le \max_{1\le i\le n}|u_{i1}|(\|\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)']\| + \|\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'\epsilon_i^2]\|) \lesssim_{\mathbb{P}} \sqrt{\frac{J\log J}{n}} + J^{-p-1}$$

where the last step follows from Lemma B.1 and the result given in the next step. In addition, $\|\mathbf{V}_{22}\| = \|2\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'\epsilon_i\sum_{\ell=1}^{d} w_{i\ell}(\widehat{\gamma}_\ell - \gamma_\ell)]\|$. Then, since $\|2\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'\epsilon_i w_{i\ell}]\| \le \|\mathbb{E}_n[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'(\epsilon_i^2 + w_{i\ell}^2)]\|$, the result can be established using the strategy given in the next step.

Step 3: For $\mathbf{V}_3$, in view of Lemma B.1, it suffices to show that

$$\sup_{\bar{\Delta}\in\Pi}\left\|\mathbb{E}_n[\mathbf{b}_0(x_i;\bar{\Delta})\mathbf{b}_0(x_i;\bar{\Delta})'(\epsilon_i^2 - \sigma^2(x_i))]\right\| \lesssim_{\mathbb{P}} \sqrt{J\log J/n^{1/2}}.$$

For notational simplicity, we write $\eta_i = \epsilon_i^2 - \sigma^2(x_i)$, $\eta_i^- = \eta_i\mathbb{1}(|\eta_i| \le M) - \mathbb{E}[\eta_i\mathbb{1}(|\eta_i| \le M)|x_i]$, $\eta_i^+ = \eta_i\mathbb{1}(|\eta_i| > M) - \mathbb{E}[\eta_i\mathbb{1}(|\eta_i| > M)|x_i]$ for some $M > 0$ to be specified later. Since $\mathbb{E}[\eta_i|x_i] = 0$, $\eta_i = \eta_i^- + \eta_i^+$. Then define a function class

$$\mathcal{G} = \Big\{(x_1,\eta_1) \mapsto b_{0,l}(x_1;\bar{\Delta})b_{0,k}(x_1;\bar{\Delta})\eta_1 : 1 \le l \le J(p+1), 1 \le k \le J(p+1), \bar{\Delta} \in \Pi\Big\}.$$

Then for $g \in \mathcal{G}$, $\sum_{i=1}^n g(x_1,\eta_1) = \sum_{i=1}^n g(x_1,\eta_1^+) + \sum_{i=1}^n g(x_1,\eta_1^-)$.

Now, for the truncated piece, we have $\sup_{g\in\mathcal{G}}|g(x_1,\eta_1^-)| \lesssim JM$, and

$$\begin{aligned}\sup_{g\in\mathcal{G}}\mathbb{V}[g(x_1,\eta_1^-)] &\lesssim \sup_{x\in\mathcal{X}}\mathbb{E}[\eta_1^2|x_1 = x]\sup_{\bar{\Delta}\in\Pi}\sup_{1\le l,k\le J(p+1)}\mathbb{E}[b_{0,l}^2(x_1;\bar{\Delta})b_{0,k}^2(x_1;\bar{\Delta})]\\ &\lesssim JM\sup_{x\in\mathcal{X}}\mathbb{E}\Big[|\eta_1|\Big|x_i = x\Big] \lesssim JM.\end{aligned}$$

The VC condition holds by the same argument given in the proof of Lemma B.1. Then using Proposition 6.2 of Belloni, Chernozhukov, Chetverikov, and Kato (2015),

$$\mathbb{E}\Big[\sup_{g\in\mathcal{G}}\Big|\mathbb{E}_n[g(x_i,\eta_i^-)]\Big|\Big] \lesssim \sqrt{\frac{JM\log(JM)}{n}} + \frac{JM\log(JM)}{n}.$$

Regarding the tail, we apply Theorem 2.14.1 of van der vaart and Wellner (1996) and obtain

$$\begin{aligned}\mathbb{E}\Big[\sup_{g\in\mathcal{G}}\Big|\mathbb{E}_n[g(x_i,\eta_i^+)]\Big|\Big] &\lesssim \frac{1}{\sqrt{n}}J\sqrt{\log J}\mathbb{E}\Big[\sqrt{\mathbb{E}_n[|\eta_i^+|^2]}\Big]\\ &\le \frac{1}{\sqrt{n}}J\sqrt{\log J}(\mathbb{E}[\max_{1\le i\le n}|\eta_i^+|])^{1/2}(\mathbb{E}[\mathbb{E}_n[|\eta_i^+|])^{1/2}\\ &\lesssim \frac{J\sqrt{\log J}}{\sqrt{n}}\cdot\frac{n^{\frac{1}{4}}}{M^{1/2}}\end{aligned}$$

where the second line follows from Cauchy-Schwarz inequality and the third line uses the fact that

$$\begin{aligned}&\mathbb{E}[\max_{1\le i\le n}|\eta_i^+|]\lesssim\mathbb{E}[\max_{1\le i\le n}\epsilon_i^2]\lesssim n^{1/2}, \quad \text{and}\\ &\mathbb{E}[\mathbb{E}_n[|\eta_i^+|]]\le\mathbb{E}[|\eta_1|^+|]\lesssim\frac{\mathbb{E}[|\epsilon|^4]}{M}.\end{aligned}$$

Then the desired result follows simply by setting $M=J$ and the sparsity of the basis.

Step 4: For $\mathbf{V}_4$, since by Assumption III.1, $\sup_{x\in\mathcal{X}}\mathbb{E}[\epsilon_i^2|x_i=x]\lesssim 1$. Then, by the same argument given in the proof of Lemma B.1,

$$\sup_{\bar\Delta\in\Pi}\Big\|\mathbb{E}_n[\mathbf{b}_s(x_i;\bar\Delta)\mathbf{b}_s(x_i;\bar\Delta)'\sigma^2(x_i)]-\mathbb{E}\Big[\mathbf{b}_s(x_i;\bar\Delta)\mathbf{b}_s(x_i;\bar\Delta)'\epsilon_i^2\Big]\Big\|\lesssim_{\mathbb{P}}\sqrt{J\log J/n}$$

and

$$\Big\|\mathbb{E}_{\widehat\Delta}\Big[\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'\epsilon_i^2\Big]-\mathbb{E}\Big[\mathbf{b}_s(x_i)\mathbf{b}_s(x_i)'\epsilon_i^2\Big]\Big\|\lesssim_{\mathbb{P}}\sqrt{J\log J/n}.$$

Then the proof is complete. □

Main Proofs

Proof of Theorem III.1. This result follows from Theorem II.1, Lemma B.1, Lemma B.3 and Lemma IV.3. See Section SA-6.11 of the supplemental appendix to Cattaneo, Crump, Farrell, and Feng (2019b) for more details. □

Proof of Lemma III.1. Let $\bar\Omega(x)^{-1/2}\widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\mathbf{Q}}^{-1}\mathbb{G}_n[\widehat{\mathbf{b}}_s(x_i)\epsilon_i]=:\mathbb{G}_n[a_i\epsilon_i]$. Conditional on $\mathbf{X}$, it is a mean zero sequence independent over i with variance equal to 1. Then by Berry-Esseen inequality,

$$\sup_{u\in\mathbb{R}}\Big|\mathbb{P}(\mathbb{G}_n[a_i\epsilon_i]\le u|\mathbf{X})-\Phi(u)\Big|\le\min\left(1,\frac{\sum_{i=1}^n\mathbb{E}[|a_i\epsilon_i|^3|\mathbf{X}]}{n^{3/2}}\right).$$

Now, using Lemma B.1,

$$\begin{aligned}
&\frac{1}{n^{3/2}}\sum_{i=1}^{n}\mathbb{E}\Big[|a_i\epsilon_i|^3\Big|\mathbf{X}\Big]\\
&\lesssim \bar{\Omega}(x)^{-3/2}\frac{1}{n^{3/2}}\sum_{i=1}^{n}|\widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\mathbf{Q}}^{-1}\widehat{\mathbf{b}}_s(x_i)|^3\\
&\leq \bar{\Omega}(x)^{-3/2}\frac{\sup_{x\in\mathcal{X}}\sup_{z\in\mathcal{X}}|\widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\mathbf{Q}}^{-1}\widehat{\mathbf{b}}_s(z)|}{n^{3/2}}\sum_{i=1}^{n}|\widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\mathbf{Q}}^{-1}\widehat{\mathbf{b}}_s(x_i)|^2\\
&\lesssim_{\mathbb{P}}\frac{1}{J^{3/2+3v}}\cdot\frac{J^{1+v}}{\sqrt{n}}\cdot J^{1+2v}\to 0
\end{aligned}$$

since $J/n = o(1)$. By Lemma B.5, the above weak convergence still holds if $\bar{\Omega}(x)$ is replaced by $\widehat{\Omega}(x)$. Now, the desired result follows by Lemma B.1 and B.3. □

Proof of Theorem III.2. The result follows by Lemma III.1 and the rate restrictions. □

Proof of Lemma III.2. Define $Z_p(x) = \frac{\widehat{\mathbf{b}}_0(x)'\mathbf{T}_s'\mathbf{Q}^{-1}\mathbf{\Sigma}^{1/2}}{\sqrt{\Omega(x)}}\mathbf{N}_{K_s}$ where $\mathbf{N}_{K_s}$ is K_s-dimensional standard normal vector defined on a sufficiently enriched probability space. Since we aim at distributional approximation, $\mathbf{N}_{K_s}$ could represent different normal vectors and should be understood in context. The proof is divided into several steps.

Step 1: Note that

$$\begin{aligned}
&\sup_{x\in\mathcal{X}}\left|\frac{\widehat{\mu}^{(v)}(x)-\mu^{(v)}(x)}{\sqrt{\widehat{\Omega}(x)/n}}-\frac{\widehat{\mu}^{(v)}(x)-\mu^{(v)}(x)}{\sqrt{\Omega(x)/n}}\right|\\
\leq&\sup_{x\in\mathcal{X}}\left|\frac{\widehat{\mu}^{(v)}(x)-\mu^{(v)}(x)}{\sqrt{\Omega(x)/n}}\right|\sup_{x\in\mathcal{X}}\left|\frac{\widehat{\Omega}(x)^{1/2}-\Omega(x)^{1/2}}{\widehat{\Omega}(x)^{1/2}}\right|\\
\lesssim_{\mathbb{P}}&\Big(\sqrt{\log J}+\sqrt{n}J^{-p-1-1/2}\Big)\left(J^{-p-1}+\sqrt{\frac{J\log J}{n^{1/2}}}\right)
\end{aligned}$$

where the last step uses Lemma B.1 and B.4. Then, in view of Lemma B.1, B.3 and B.5 and the rate restriction given in the lemma, we have

$$\sup_{x\in\mathcal{X}}\left|\frac{\widehat{\mu}^{(v)}(x)-\mu(x)}{\sqrt{\widehat{\Omega}(x)/n}}-\frac{\widehat{\mathbf{b}}_s(x)'\widehat{\mathbf{Q}}^{-1}}{\sqrt{\Omega(x)}}\mathbb{G}_n[\widehat{\mathbf{b}}_s(x_i)\epsilon_i]\right| = o_{\mathbb{P}}(1/\sqrt{\log J}).$$

Step 2: Write $\mathscr{K}(x,x_i) = \Omega(x)^{-1/2}\widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\mathbf{Q}}^{-1}\mathbf{b}_s(x_i)$. Using Lemma A.3 given in Appendix A, it follows that for any $\eta > 0$, $\mathbb{P}\Big(\sup_{x\in\mathcal{X}}|\mathbb{G}_n[\mathscr{K}(x,x_i)(\epsilon_i-\sigma_i\zeta_i)]| > \eta/\sqrt{\log J}|\mathbf{X}\Big) = o_{\mathbb{P}}(1)$, where $\sigma_i^2 = \sigma^2(x_i)$. Since $\mathbb{G}_n[\widehat{\mathbf{b}}(x_i)\zeta_i\sigma_i] =_{d|\mathbf{X}} \mathbf{N}(0,\bar{\mathbf{\Sigma}})$ $(=_{d|\mathbf{X}}$

denotes "equal in distribution conditional on $\mathbf{X}$"), the above steps construct the following approximating process: $\bar{Z}_p(x) := \frac{\widehat{\mathbf{b}}^{(v)}(x)'\widehat{\mathbf{Q}}^{-1}}{\sqrt{\Omega(x)}}\bar{\mathbf{\Sigma}}^{1/2}\mathbf{N}_{K_s}$.

Step 3: Since $\sup_{x\in\mathcal{X}}|\bar{Z}_p(x) - Z_p(x)| \leq \sup_{x\in\mathcal{X}}\left|\frac{\widehat{\mathbf{b}}^{(v)}(x)'(\widehat{\mathbf{Q}}^{-1}-\mathbf{Q}^{-1})}{\sqrt{\Omega(x)}}\bar{\mathbf{\Sigma}}^{1/2}\mathbf{N}_{K_s}\right| + \sup_{x\in\mathcal{X}}\left|\frac{\widehat{\mathbf{b}}^{(v)}(x)'\mathbf{Q}^{-1}}{\sqrt{\Omega(x)}}\left(\bar{\mathbf{\Sigma}}^{1/2}-\mathbf{\Sigma}^{1/2}\right)\mathbf{N}_{K_s}\right|+\sup_{x\in\mathcal{X}}\left|\frac{\widehat{\mathbf{b}}_0^{(v)}(x)'(\widehat{\mathbf{T}}_s-\mathbf{T}_s)\mathbf{Q}^{-1}}{\sqrt{\Omega(x)}}\mathbf{\Sigma}^{1/2}\mathbf{N}_{K_s}\right|$, where each term on the right-hand-side is a mean-zero Gaussian process conditional on $\mathbf{X}$. Using Lemma B.1, Theorem X.3.8 of Bhatia (2013) and applying Gaussian Maximal Inequality (see Chernozhukov, Lee, and Rosen, 2013, Lemma 13) as in the proof of Theorem II.5, we have $\mathbb{E}\left[\sup_{x\in\mathcal{X}}|\bar{Z}_p(x)-Z_p(x)|\middle|\mathbf{X}\right] \lesssim_{\mathbb{P}} \sqrt{\log J}(\|\bar{\mathbf{\Sigma}}^{1/2}-\mathbf{\Sigma}^{1/2}\|+\|\widehat{\mathbf{Q}}^{-1}-\mathbf{Q}^{-1}\|+\|\widehat{\mathbf{T}}_s-\mathbf{T}_s\|) = o_{\mathbb{P}}(1/\sqrt{\log J})$.

Step 4: It follows from the same argument as given in Step 3 that on a properly enriched probability space, there exists K_s-dimensional standard normal vector $\mathbf{N}_{K_s}$ independent of $\mathbf{D}$ such that for any $\eta > 0$, $\mathbb{P}\left[\sup_{x\in\mathcal{X}}|\widehat{Z}_p(x)-Z_p(x)| > \frac{\eta}{\sqrt{\log J}}\middle|\mathbf{D}\right] = o_{\mathbb{P}}(1)$.

Step 5: Using the similar argument as in the proof of Theorem II.8, we have

$$\sup_{u\in\mathbb{R}}\left|\mathbb{P}\Big(\sup_{x\in\mathcal{X}}|\widehat{T}_p(x)| \leq u\Big) - \mathbb{P}\Big(\sup_{x\in\mathcal{X}}|Z_p(x)| \leq u\Big)\right| = o(1),$$

and

$$\sup_{u\in\mathbb{R}}\left|\mathbb{P}\Big(\sup_{x\in\mathcal{X}}|\widehat{Z}_p(x)| \leq u\Big|\mathbf{X}\Big) - \mathbb{P}\Big(\sup_{x\in\mathcal{X}}|Z_p(x)| \leq u\Big|\mathbf{X}\Big)\right| = o_{\mathbb{P}}(1).$$

Then it remains to show that

$$\sup_{u\in\mathbb{R}}\left|\mathbb{P}\Big(\sup_{x\in\mathcal{X}}|Z_p(x)| \leq u\Big) - \mathbb{P}\Big(\sup_{x\in\mathcal{X}}|Z_p(x)| \leq u|\mathbf{X}\Big)\right| = o_{\mathbb{P}}(1). \tag{B.4}$$

We can write $Z_p(x) = \frac{\widehat{\mathbf{b}}_0^{(v)}(x)'}{\sqrt{\widehat{\mathbf{b}}_0^{(v)}(x)'\mathbf{V}\widehat{\mathbf{b}}_0^{(v)}(x)}}\breve{N}_{K_s}$ where $\mathbf{V} = \mathbf{T}_s'\mathbf{Q}^{-1}\mathbf{\Sigma}\mathbf{Q}^{-1}\mathbf{T}_s$ and $\breve{\mathbf{N}}_{K_s} := \mathbf{T}_s'\mathbf{Q}^{-1}\mathbf{\Sigma}^{1/2}\mathbf{N}_{K_s}$ is a K_s-dimensional normal random vector. Importantly, by this construction, $\breve{\mathbf{N}}_{K_s}$ and $\mathbf{V}$ do not depend on $\widehat{\Delta}$ and x, and they are only determined by the deterministic partition Δ.

Now, first consider $v = 0$. For any two partitions $\Delta_1, \Delta_2 \in \Pi$, for any $x \in \mathcal{X}$, there exists $\check{x} \in \mathcal{X}$ such that $\mathbf{b}_0^{(v)}(x;\Delta_1) = \mathbf{b}_0^{(v)}(\check{x};\Delta_2)$, and vice versa. Therefore, the following two events are equivalent: $\{\omega : \sup_{x\in\mathcal{X}}|Z_p(x;\Delta_1)| \leq u\} = \{\omega : \sup_{x\in\mathcal{X}}|Z_p(x;\Delta_2)| \leq u\}$ for any u. Thus, $\mathbb{E}\left[\mathbb{P}\Big(\sup_{x\in\mathcal{X}}|Z_p(x)| \leq u\Big|\mathbf{X}\Big)\right] = \mathbb{P}\Big(\sup_{x\in\mathcal{X}}|Z_p(x)| \leq u\Big|\mathbf{X}\Big)$. Then for $v = 0$, the desired result follows.

For $v > 0$, simply notice that $\widehat{\mathbf{b}}_0^{(v)}(x) = \widehat{\mathfrak{T}}_v\widehat{\mathbf{b}}_0(x)$ for some transformation matrix $\widehat{\mathfrak{T}}_v$. Clearly, $\widehat{\mathfrak{T}}_v$ takes a similar structure as $\widehat{\mathbf{T}}_s$: each row and each column only have a finite number of nonzeros. Each nonzero element is simply $\hat{h}_j^{-v}$ up to some constants.

Using the argument given in the proof of Lemma SA-2.2 of Cattaneo, Crump, Farrell, and Feng (2019b), $\|\widehat{\mathfrak{T}}_v - \mathfrak{T}_v\| \lesssim \sqrt{J\log J/n}$ where $\mathfrak{T}_v$ is the population analogue ($\hat{h}_j$ replaced by h_j). Repeating the argument given in Step 3 and 4, we can replace $\widehat{\mathfrak{T}}_v$ in $Z_p(x)$ by $\mathfrak{T}_v$ without affecting the approximation rate. Then the desired result follows by repeating the argument given for $v = 0$ above. □

We introduce some notation for the following proofs. Let $\eta_{1,n} = o(1)$, $\eta_{2,n} = o(1)$ and $\eta_{3,n} = o(1)$ be sequences of vanishing constants. Moreover, let A_n be a sequence of diverging constants such that $\sqrt{\log J}A_n \leq \sqrt{\frac{n}{J^{1+2v}}}$.

Proof of Theorem III.3. Given $J = J_{\texttt{IMSE}} \asymp n^{\frac{1}{2p+3}}$, the rate restrictions required in Theorem III.2 are satisfied. Then,

$$\begin{aligned}
\mathbb{P}\left[\sup_{x\in\mathcal{X}}|\widehat{T}_{p+q}(x)| \leq \mathfrak{c}\right] &\leq \mathbb{P}\left[\sup_{x\in\mathcal{X}}|Z_{p+q}(x)| \leq \mathfrak{c} + \eta_{1,n}/\sqrt{\log J}\right] + o(1)\\
&\leq \mathbb{P}\left[\sup_{x\in\mathcal{X}}|Z_{p+q}(x)| \leq c^0(1-\alpha+\eta_{3,n}) + \frac{\eta_{1,n}+\eta_{2,n}}{\sqrt{\log J}}\right] + o(1)\\
&\leq \mathbb{P}\left[\sup_{x\in\mathcal{X}}|Z_{p+q}(x)| \leq c^0(1-\alpha+\eta_{3,n})\right] + o(1) \to 1-\alpha,
\end{aligned}$$

where $c^0(1-\alpha+\eta_{3,n})$ denotes the $(1-\alpha+\eta_{3,n})$-quantile of $\sup_{x\in\mathcal{X}}|Z_{p+q}(x)|$, the second follows by Lemma A.1 of Belloni, Chernozhukov, Chetverikov, and Kato (2015), and the third by Anti-Concentration Inequality in Chernozhukov, Chetverikov, and Kato (2014b). The other side of the bound follows similarly. □

Proof of Theorem III.4. Note that under $\ddot{\mathsf{H}}_0$,

$$\sup_{x\in\mathcal{X}}|\ddot{T}_p(x)| \leq \sup_{x\in\mathcal{X}}\left|\frac{\widehat{\mu}^{(v)}(x)-\mu^{(v)}(x)}{\sqrt{\widehat{\Omega}(x)/n}}\right| + \sup_{x\in\mathcal{X}}\left|\frac{\mu^{(v)}(x)-m^{(v)}(x;\widehat{\boldsymbol{\theta}})}{\sqrt{\widehat{\Omega}(x)/n}}\right|.$$

Therefore,

$$\mathbb{P}\left[\sup_{x\in\mathcal{X}}|\ddot{T}_p(x)| > \mathfrak{c}\right] \leq \mathbb{P}\left[\sup_{x\in\mathcal{X}}|\widehat{T}_p(x)| > \mathfrak{c} - \sup_{x\in\mathcal{X}}\left|\frac{\mu^{(v)}(x)-m^{(v)}(x;\widehat{\boldsymbol{\theta}})}{\sqrt{\widehat{\Omega}(x)/n}}\right|\right]$$

$$\leq \mathbb{P}\left[\sup_{x\in\mathcal{X}}|Z_p(x)| > \mathfrak{c} - \frac{\eta_{1,n}}{\sqrt{\log J}} - \sup_{x\in\mathcal{X}}\left|\frac{\mu^{(v)}(x)-m^{(v)}(x;\widehat{\boldsymbol{\theta}})}{\sqrt{\widehat{\Omega}(x)/n}}\right|\right] + o(1)$$

$$\begin{aligned}
\leq \mathbb{P}\Bigg[\sup_{x\in\mathcal{X}}|Z_p(x)| > c^0(1-\alpha-\eta_{3,n}) - \frac{\eta_{1,n}+\eta_{2,n}}{\sqrt{\log J}} -\\
\sup_{x\in\mathcal{X}}\left|\frac{\mu^{(v)}(x)-m^{(v)}(x,\widehat{\boldsymbol{\theta}})}{\sqrt{\widehat{\Omega}(x)/n}}\right|\Bigg] + o(1)
\end{aligned}$$

$$
\begin{aligned}
&\leq \mathbb{P}\Big[\sup_{x\in\mathcal{X}}|Z_p(x)| > c^0(1-\alpha-\eta_{3,n})\Big] + o(1)\\
&= \alpha + o(1)
\end{aligned}
$$

where $c^0(1-\alpha-\eta_{3,n})$ denotes the $(1-\alpha-\eta_{3,n})$-quantile of $\sup_{x\in\mathcal{X}}|Z_p(x)|$, the second inequality holds by Lemma III.2, the third by Lemma A.1 of Belloni, Chernozhukov, Chetverikov, and Kato (2015), the fourth by the condition that $\sup_{x\in\mathcal{X}}\big|\frac{\mu^{(v)}(x)-m^{(v)}(x,\widehat{\boldsymbol{\theta}})}{\sqrt{\widehat{\Omega}(x)/n}}\big| = o_{\mathbb{P}}(\frac{1}{\sqrt{\log J}})$ and Anti-Concentration Inequality in Chernozhukov, Chetverikov, and Kato (2014b). The other side of the bound follows similarly.

On the other hand, under $\ddot{\mathsf{H}}_A$,

$$
\begin{aligned}
&\mathbb{P}\Big[\sup_{x\in\mathcal{X}}|\ddot{T}_p(x)| > \mathfrak{c}\Big]\\
\geq\,&\mathbb{P}\Big[\sup_{x\in\mathcal{X}}|\widehat{T}_p(x)| \leq \sup_{x\in\mathcal{X}}\Big|\frac{\mu^{(v)}(x)-m^{(v)}(x,\bar{\boldsymbol{\theta}})}{\sqrt{\widehat{\Omega}(x)/n}} + \frac{m^{(v)}(x,\bar{\boldsymbol{\theta}})-m^{(v)}(x,\widehat{\boldsymbol{\theta}})}{\sqrt{\widehat{\Omega}(x)/n}}\Big| - \mathfrak{c}\Big] - o(1)\\
\geq\,&\mathbb{P}\Big[\sup_{x\in\mathcal{X}}|Z_p(x)| \leq \sqrt{\log J}A_n - \eta_{1,n}/\sqrt{\log J}\Big] - o(1)\\
\geq\,&1-o(1).
\end{aligned}
$$

where the second line holds by Lemma B.1, Lemma B.5, Lemma A.1 of Belloni, Chernozhukov, Chetverikov, and Kato (2015) and $J^v\sqrt{J\log J/n} = o(1)$, the third by definition of A_n and Lemma III.2, and the last by Concentration Inequality given in Lemma 12 of Chernozhukov, Lee, and Rosen (2013). □

Proof of Theorem III.5. The proof is similar to that for Theorem III.4, and omitted here. See Section SA-6.19 of Cattaneo, Crump, Farrell, and Feng (2019b) for more details. □

Simulated Dataset

Section 3.2 uses a simulated dataset to illustrate our main methods. The data generating process is constructed based on the real survey dataset on the Gini index and household income. Specifically, we set the sample size $n = 1,000$, and the independent variable of interest $x_i \sim \texttt{beta}(2,4)$ where $\texttt{beta}(2,4)$ is *beta* distribution with parameters 2 and 4. The regression function of interest is

$$\mu(x) = 24x^4 - 98.8x^3 + 112.4x^2 - 44.4x + 3.6.$$

The basic model for the outcome variable y is constructed as

$$y_i = \mu(x_i) + w_i + \epsilon_i, \quad \epsilon_i \sim \mathtt{N}(0, 0.5^2), \quad w_i \sim \mathtt{U}(-1, 1)$$

where $\mathtt{N}(0, 0.5^2)$ is the normal distribution with mean 0 and variance 0.5^2 and $\mathtt{U}(-1, 1)$ is the uniform distribution over $[-1, 1]$. x_i, w_i and ϵ_i are independent of each other.

The basic model is modified in several scenarios. In the discussion of data variability and heteroskedasticity, we change the conditional variance of ϵ_i. In the discussion of covariate adjustment, we allow the dependence between x_i and w_i. In that case, $w_i = 3(x - 0.5) + \mathtt{U}(-0.5, 0.5)$ where the errors are independent of x_i.

Appendix C

Proof for Chapter IV

Proof of Lemma IV.1.

(1) Assumption II.3(1) directly follows from the construction of tensor-product B-splines. Assumption II.3(2) follows from Schumaker (2007, Theorem 12.5). To prove Assumption II.3(3), notice that given univariate B-splines $\{p_{l_\ell}(x_\ell)\}_{l_\ell=1}^{K_\ell}$, there exists a universal constant $C > 0$ such that for any $\varsigma_\ell \leq m$, $\delta \in \Delta$, $\sup_{x_\ell \in \mathrm{clo}(\delta)} \left| d^{\varsigma_\ell} p_{l_\ell}(x_\ell)/dx_\ell^{\varsigma_\ell} \right| \lesssim h^{-\varsigma_\ell}$. Since there are only a fixed number of nonzero elements in $\mathbf{p}$, we have for $[\boldsymbol{\varsigma}] \leq m$, $\sup_{\delta \in \Delta} \sup_{\mathbf{x} \in \mathrm{clo}(\delta)} \|\partial^{\boldsymbol{\varsigma}} \mathbf{p}(\mathbf{x})\| \lesssim h^{-[\boldsymbol{\varsigma}]}$. To derive the other side of the bound, notice that the proof of Zhou and Wolfe (2000, Lemma 5.4) shows that for a univariate B-spline basis $\breve{\mathbf{p}}_\ell(x_\ell) := (\breve{p}_1(x_\ell), \cdots, \breve{p}_{K_\ell}(x_\ell))'$, for any $\varsigma_\ell \leq m-1$, $x_\ell \in \mathcal{X}_\ell$, $\left\| d^{\varsigma_\ell} \breve{\mathbf{p}}_\ell(x_\ell)/dx_\ell^{\varsigma_\ell} \right\| \gtrsim h^{-\varsigma_\ell}$. Since for any x_ℓ, there are only m nonzero elements in $\breve{\mathbf{p}}_\ell(x_\ell)$, this suffices to show that for any $x_\ell \in \mathcal{X}_\ell$, there exists some $\breve{p}_{l_\ell}(x_\ell)$ such that $\left| d^{\varsigma_\ell} p_{l_\ell}(x_\ell)/dx_\ell^{\varsigma_\ell} \right| \gtrsim h^{-\varsigma_\ell}$. Then the direct lower bound directly follows from the construction of tensor-product B-splines.

(2) The proof of orthogonality between the constructed leading error and B-splines can be found in Barrow and Smith (1978). Regarding the bias expansion, we first consider $\boldsymbol{\varsigma} = \mathbf{0}$. Noticing that

$$\mathscr{B}_{m,\mathbf{0}}(\mathbf{x}) = -\sum_{\ell=1}^{d} \frac{\partial^m \theta(\mathbf{t}_{\mathbf{x}}^L)}{\partial x_\ell^m} \frac{b_{\ell,l_\ell}^m}{m!} B_m\left(\frac{x_\ell - t_{\ell,l_\ell}}{b_{\ell,l_\ell}}\right) + O(h^{m+\varrho}) \quad \text{for } \mathbf{x} \in \delta_{l_1 \ldots l_d}$$

where $B_m(\cdot)$ is the mth Bernoulli polynomial, we only need to work with the first term on the RHS, denoted by $\bar{\mathscr{B}}_m(\mathbf{x})$. By construction, $\bar{\mathscr{B}}_m(\mathbf{x})$ is continuous on the interior of each subrectangle $\delta_{l_1 \ldots l_d}$, and the discontinuity only takes place at boundaries of subrectangles. Let $J_{\mathbf{0}}$ denote the magnitude of jump of $\bar{\mathscr{B}}_m(\mathbf{x})$. By Assumption II.1, $J_{\mathbf{0}}$ is also the jump of $\bar{\theta} := \theta + \bar{\mathscr{B}}_m$. We first check the magnitude of the jump as Barrow and Smith (1978) did in their proof. We introduce the following notation:

i. $\boldsymbol{\tau} := (\tau_1, \cdots, \tau_d)$ is a point on the boundary of a generic rectangle $\delta_{l_1 \ldots l_d}$;

ii. $\boldsymbol{\tau}^- := (\tau_1^-, \cdots, \tau_d^-)$ and $\boldsymbol{\tau}^+ := (\tau_1^+, \cdots, \tau_d^+)$ are two points close to $\boldsymbol{\tau}$ but belong to two different subrectangles $\delta^-_{l_1\dots l_d} := \{\mathbf{x}\colon t^-_{\ell,l_\ell} \le x_\ell < t^-_{\ell,l_\ell+1}\}$ and $\delta^+_{l_1\dots l_d} := \{\mathbf{x}\colon t^+_{\ell,l_\ell} \le x_\ell < t^+_{\ell,l_\ell+1}\}$;

iii. $\mathbf{t}^L_-$ and $\mathbf{t}^L_+$ are the starting points of $\delta^-_{l_1\dots l_d}$ and $\delta^+_{l_1\dots l_d}$;

iv. $(b_{1,-},\cdots,b_{d,-})$ and $(b_{1,+},\cdots,b_{d,+})$ are the corresponding mesh widths of $\delta^-_{l_1\dots l_d}$ and $\delta^+_{l_1\dots l_d}$;

v. $\Xi := \{\ell\colon\ \tau^-_\ell - \tau_\ell$ and $\tau^+_\ell - \tau_\ell$ differ in signs$\}$.

In words, the index set Ξ indicates the directions in which we cross boundaries when we move from $\boldsymbol{\tau}^-_\ell$ to $\boldsymbol{\tau}^+_\ell$. To further simplify notation, we write $\bar{\theta}(\boldsymbol{\tau}^-) := \lim_{\mathbf{x}\to\boldsymbol{\tau},\mathbf{x}\in\delta^-_{l_1\dots l_d}} \bar{\theta}(\mathbf{x})$ and $\bar{\theta}(\boldsymbol{\tau}^+) := \lim_{\mathbf{x}\to\boldsymbol{\tau},\mathbf{x}\in\delta^+_{l_1\dots l_d}} \bar{\theta}(\mathbf{x})$. Then we have

$$\begin{aligned}
J_{\mathbf{0}} &= |\bar{\theta}(\boldsymbol{\tau}^+) - \bar{\theta}(\boldsymbol{\tau}^-)| = \Big|\bar{\mathscr{B}}_m(\boldsymbol{\tau}^+) - \bar{\mathscr{B}}_m(\boldsymbol{\tau}^-))\Big| \\
&= \sum_{\ell\in\Xi}(B_m(0)|/m!)\Big|\frac{\partial^m\theta(\mathbf{t}^L_+)}{\partial x^m_\ell} b^m_{\ell,+} - \frac{\partial^m\theta(\mathbf{t}^L_-)}{\partial x^m_\ell} b^m_{\ell,-}\Big| \\
&= \sum_{\ell\in\Xi}(B_m(0)|/m!)\Big|\Big(\frac{\partial^m\theta(\mathbf{t}^L_+)}{\partial x^m_\ell} - \frac{\partial^m\theta(\mathbf{t}^L_-)}{\partial x^m_\ell}\Big) b^m_{\ell,+} + \frac{\partial^m\theta(\mathbf{t}^L_-)}{\partial x^m_\ell}(b^m_{\ell,+} - b^m_{\ell,-})\Big| \\
&\le \sum_{\ell\in\Xi}(B_m(0)|/m!)\Big[O(h^{m+\varrho}) + Ch^{m-1}|b_{\ell,+} - b_{\ell,-}|\Big] \\
&\le \sum_{\ell\in\Xi}(B_m(0)|/m!)\Big[O(h^{m+\varrho}) + Ch^{m-1}O(h^{1+\varrho})\Big]
\end{aligned}$$

where the fourth line follows from Assumption II.1 and the last line follows from the stronger quasi-uniformity condition given in the Lemma. This suffices to show that $J_{\mathbf{0}}$ is $O(h^{m+\varrho})$.

Then we mimic the proof strategy used in Schumaker (2007, Theorem 12.7). By Schumaker (2007, Theorem 12.6), we can construct a bounded linear operator $\mathscr{L}[\cdot]$ mapping $\mathcal{L}_1(\mathcal{X})$ onto $\mathcal{S}_{\Delta,m}$ with $\mathscr{L}[s] = s$ for all $s \in \mathcal{S}_{\Delta,m}$. Specifically, $\mathscr{L}[\cdot]$ is defined as

$$\mathscr{L}[\theta](\mathbf{x}) := \sum_{l_1=1}^{K_1}\cdots\sum_{l_d=1}^{K_d}(\psi_{l_1\dots l_d}\theta)p_{l_1\dots l_d}(\mathbf{x})$$

where $\{\psi_{l_1\dots l_d}\}^{K_1,\dots,K_d}_{l_1=1,\dots,l_d=1}$ is a dual basis defined as Schumaker (2007, Equation 12.24). By multi-dimensional Taylor expansion, there exists a polynomial $\varphi_{l_1\dots l_d}$ such that $\|\bar{\theta} - \varphi_{l_1\dots l_d}\|_{L_\infty(\delta_{l_1\dots l_d})} \lesssim h^{m+\varrho}$, and the degree of $\varphi_{l_1\dots l_d}$ is no greater than $m-1$. Since

$\mathscr{L}$ reproduces polynomials, we have

$$\begin{aligned}\|\bar{\theta} - \mathscr{L}[\bar{\theta}]\|_{L_\infty(\delta_{l_1 \dots l_d})} &\leq \|\bar{\theta} - \varphi_{l_1 \dots l_d}\|_{L_\infty(\delta_{l_1 \dots l_d})} + \|\mathscr{L}[\bar{\theta} - \varphi_{l_1 \dots l_d}]\|_{L_\infty(\delta_{l_1 \dots l_d})} \\ &\leq C\|\bar{\theta} - \varphi_{l_1 \dots l_d}\|_{L_\infty(\delta_{l_1 \dots l_d})} \lesssim h^{m+\varrho}.\end{aligned}$$

With the jump of $\bar{\theta}$ along boundaries taken account of, the approximation error of $\mathscr{L}[\bar{\theta}]$ is still $O(h^{m+\varrho})$. Evaluate the L_∞ norm on all subrectangles, and then we conclude that there exists some $s^* \in \mathcal{S}_{\Delta,m}$ such that $\|\theta + \bar{\mathscr{B}}_m - s^*\|_{L_\infty(\mathcal{X})} \lesssim h^{m+\varrho}$.

For other $\boldsymbol{\varsigma}$, we only need to show that the desired result holds for $s^* = \mathscr{L}[\bar{\theta}]$. By construction of $\mathscr{L}$,

$$|\partial^{\varsigma}(\mathscr{L}[\bar{\theta}])| \leq \sum_{l_1=1}^{m+\kappa_1} \cdots \sum_{l_d=1}^{m+\kappa_d} |\psi_{l_1 \dots l_d}\bar{\theta}||\partial^{\varsigma} p_{l_1 \dots l_d}(\mathbf{x})| \leq Ch^{-[\varsigma]}\|\bar{\theta}\|_{L_\infty(\delta_{l_1 \dots l_d})} \tag{C.1}$$

where the last line follows from (Schumaker, 2007, Theorem 12.5). Then we have

$$\begin{aligned}&\|\partial^{\varsigma}\theta + \partial^{\varsigma}\bar{\mathscr{B}}_m - \partial^{\varsigma}(\mathscr{L}[\bar{\theta}])\|_{L_\infty(\delta_{l_1 \dots l_d})} \\ &\leq \|\partial^{\varsigma}\theta + \partial^{\varsigma}\mathscr{B}_m^* - \partial^{\varsigma}\varphi_{l_1 \dots l_d}\|_{L_\infty(\delta_{l_1 \dots l_d})} + \|\partial^{\varsigma}(\mathscr{L}[\bar{\theta} - \varphi_{l_1 \dots l_d}])\|_{L_\infty(\delta_{l_1 \dots l_d})} \\ &\leq O(h^{m+\varrho-[\varsigma]}) + Ch^{-[\varsigma]}\|\bar{\theta} - \varphi_{l_1 \dots l_d}\|_{L_\infty(\delta_{l_1 \dots l_d})} \lesssim h^{m+\varrho-[\varsigma]}\end{aligned}$$

where the second inequality follows from Taylor expansion and Equation (C.1). By the similar argument for $J_{\mathbf{0}}$, the jump of $\partial^{\varsigma}\bar{\mathscr{B}}_m$ is $O(h^{m+\varrho-[\varsigma]})$.

(3) By construction of $\tilde{\mathbf{p}}$, $\rho = 1$. It follows from the same argument in part (1) and (2) that $\tilde{\mathbf{p}}$ satisfies Assumption II.3 and II.4. Finally, by definition of tensor-product splines, both $\mathbf{p}$ and $\tilde{\mathbf{p}}$ reproduce polynomials of degree no greater than $m-1$. Then the proof is complete. □

Proof of Lemma IV.2.

(1) Assumption II.3(1) directly follows from the fact that the father wavelet is compactly supported and $\{\phi_{sl}\}$ is generated by translation and dilation. Assumption II.1(2) follows from the fact that $\{\phi_{sl}\}$ is an orthonormal basis with respect to the Lebesgue measure. For Assumption II.3(3), notice that

$$\frac{d^{\varsigma_\ell}\phi(2^s x_\ell - l_\ell)}{dx_\ell^{\varsigma_\ell}} = 2^{s\varsigma_\ell}\frac{d^{\varsigma_\ell}\phi(z)}{dz^{\varsigma_\ell}}\Big|_{z=2^s x_\ell - l_\ell} = b^{-\varsigma_\ell}\frac{d^{\varsigma_\ell}\phi(z)}{dz^{\varsigma_\ell}}\Big|_{z=2^s x_\ell - l_\ell}.$$

Since the wavelet basis reproduces polynomials of degree no greater than $m-1$ and ϕ is assumed to have $q+1$ continuous derivatives, the desired bounds follow.

(2) We follow the same strategy used in Sweldens and Piessens (1994), and extend their proof to the multidimensional case. First, we denote by V_s^ℓ the closure of the level-s subspace spanned by $\{\phi_{sl}(x_\ell)\}$ and W_s^ℓ the orthogonal complement of V_s^ℓ in V_{s+1}^ℓ. Then we write $\mathcal{V}_s := \otimes_{\ell=1}^d V_s^\ell$ for the space spanned by the tensor-product level-s father wavelets, and $\mathcal{W}_s$ as the orthogonal complement of $\mathcal{V}_s$ in $\mathcal{V}_{s+1}$. We use the following fact: $\mathcal{W}_s = \oplus_{i=1}^{2^d-1}\mathcal{W}_{s,i}$ where $\oplus$ denotes "direct sum", and each $\mathcal{W}_{s,i}$ takes the following form: $\mathcal{W}_{s,i} = \otimes_{\ell=1}^d Z_s^\ell$. Each Z_s^ℓ is either V_s^ℓ or W_s^ℓ, but $\{Z_s^\ell\}_{\ell=1}^d$ cannot be identical to $\{V_s^\ell\}_{\ell=1}^d$. There are in total (2^d-1) such subspaces. Accordingly, a typical element in a basis vector of $\mathcal{W}_s$ can be written as

$$\bar{\psi}_{s\mathbf{l}\boldsymbol{\alpha}}(\mathbf{x}) = \prod_{\ell=1}^{d}\left[\alpha_\ell \phi_{sl_\ell}(x_\ell) + (1-\alpha_\ell)\psi_{sl_\ell}(x_\ell)\right]$$

where $\mathbf{l} = (l_1, \ldots, l_d)$ and $\alpha_\ell = 0$ or 1, but $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d) \neq (1, \ldots, 1)$. Then it directly follows from the properties of wavelet basis that for $\bar{\psi}_{s\mathbf{l}\boldsymbol{\alpha}}$, $s \geq m$,

$$\langle \mathbf{x}^{\boldsymbol{\varsigma}}, \bar{\psi}_{s\mathbf{l}\boldsymbol{\alpha}}(\mathbf{x})\rangle := \int_{\mathcal{X}} \mathbf{x}^{\boldsymbol{\varsigma}}\bar{\psi}_{s\mathbf{l}\boldsymbol{\alpha}}(\mathbf{x})d\mathbf{x} = 0, \quad \text{for } \boldsymbol{\varsigma} \text{ such that } [\boldsymbol{\varsigma}] \leq m, \varsigma_\ell \neq m \ \forall \ell. \tag{C.2}$$

Denote by $\mathscr{L}_s[\cdot]$ the orthogonal projection operator onto $\mathcal{W}_s$. Then the approximation error of the tensor-product wavelet space $\mathcal{V}_{s_n}$ can be written as

$$\begin{aligned}\sum_{s=s_n}^{\infty}\mathscr{L}_s[\theta](\mathbf{x}) &= \sum_{s=s_n}^{\infty}\sum_{\boldsymbol{\alpha}}\sum_{\mathbf{l}}\langle\theta(\check{\mathbf{x}}), \bar{\psi}_{s\mathbf{l}\boldsymbol{\alpha}}(\check{\mathbf{x}})\rangle\bar{\psi}_{s\mathbf{l}\boldsymbol{\alpha}}(\mathbf{x}) \\ &= \sum_{s=s_n}^{\infty}\sum_{\boldsymbol{\alpha}}\sum_{\mathbf{l}}\left\langle \sum_{[\boldsymbol{\varsigma}]\leq m}\partial^{\boldsymbol{\varsigma}}\theta(\mathbf{x})\frac{(\check{\mathbf{x}}-\mathbf{x})^{\boldsymbol{\varsigma}}}{\boldsymbol{\varsigma}!} + \vartheta_n(\check{\mathbf{x}}, \mathbf{x}), \ \bar{\psi}_{s\mathbf{l}\boldsymbol{\alpha}}(\check{\mathbf{x}})\right\rangle\bar{\psi}_{s\mathbf{l}\boldsymbol{\alpha}}(\mathbf{x})\end{aligned}$$

where $\vartheta_n(\check{\mathbf{x}}, \mathbf{x}) \lesssim \|\check{\mathbf{x}} - \mathbf{x}\|^{m+\varrho}$, and the inner product in the above equations are taken with respect to $\check{\mathbf{x}}$ in terms of Lebesgue measure. It follows from Assumption II.1 and Assumption II.3 that

$$\begin{aligned}&\sup_{\mathbf{x}\in\mathcal{X}}\left|\sum_{s=s_n}^{\infty}\sum_{\boldsymbol{\alpha}}\sum_{\mathbf{l}}\left\langle\vartheta_n(\check{\mathbf{x}}, \mathbf{x}), \ \bar{\psi}_{s\mathbf{l}\boldsymbol{\alpha}}(\check{\mathbf{x}})\right\rangle\bar{\psi}_{s\mathbf{l}\boldsymbol{\alpha}}(\mathbf{x})\right| \\ =&\sup_{\mathbf{x}\in\mathcal{X}}\left|\sum_{s=s_n}^{\infty}\left(\frac{b}{2^{s-s_n}}\right)^{m+\varrho}\sum_{\boldsymbol{\alpha}}\sum_{\mathbf{l}}\left\langle\vartheta_n(\check{\mathbf{x}}, \mathbf{x})2^{s(m+\varrho)}, \ \bar{\psi}_{s\mathbf{l}\boldsymbol{\alpha}}(\check{\mathbf{x}})\right\rangle\bar{\psi}_{s\mathbf{l}\boldsymbol{\alpha}}(\mathbf{x})\right| \lesssim b^{m+\varrho}.\end{aligned}$$

Recall that $b = 2^{-s_n}$.

Regarding the leading terms

$$\sum_{s=s_n}^{\infty}\sum_{\boldsymbol{\alpha}}\sum_{\mathbf{l}}\left\langle\sum_{[\boldsymbol{\varsigma}]\leq m}\partial^{\boldsymbol{\varsigma}}\theta(\mathbf{x})\frac{(\check{\mathbf{x}}-\mathbf{x})^{\boldsymbol{\varsigma}}}{\boldsymbol{\varsigma}!}, \ \bar{\psi}_{s\mathbf{l}\boldsymbol{\alpha}}(\check{\mathbf{x}})\right\rangle\bar{\psi}_{s\mathbf{l}\boldsymbol{\alpha}}(\mathbf{x}),$$

it is clear that the coefficients of the wavelet basis can be viewed as a linear combination of the inner products of monomials and the mother wavelets themselves, and thus by Equation (C.2) the leading error is of order b^m and can be characterized as

$$\mathfrak{B}_{m,\mathbf{0}}(\mathbf{x}) = -\sum_{\boldsymbol{u}\in\Lambda_m} \frac{b^m}{\boldsymbol{u}!}\partial^{\boldsymbol{u}}\theta(\mathbf{x})B^{\mathtt{W}}_{\boldsymbol{u},\mathbf{0}}(\mathbf{x}/b).$$

$B^{\mathtt{W}}_{\boldsymbol{u},\mathbf{0}}$ is referred to as "monowavelet" in Sweldens and Piessens (1994). Here we extend it to the multidimensional case. Specifically, define a mapping

$$\begin{aligned}\varphi :\ \Lambda_m &\to \{1,\ldots,d\}\\ \boldsymbol{u} &\mapsto \ell\end{aligned}$$

such that $\varphi(\boldsymbol{u})$th element of $\boldsymbol{u}$ is nonzero. We denote $\mathbf{l}_{-\ell} := (l_1,\ldots,l_{\ell-1},l_{\ell+1},\ldots,l_d)$ and $\mathcal{L}_s^{-\ell} := \left\{\mathbf{l}_{-\ell} : l_{\ell'} \in \mathcal{L}_s, j' = \{1,\cdots,d\}\setminus\{\ell\}\right\}$. Then define

$$\varpi_{\boldsymbol{u},s}(\mathbf{x}) = \sum_{l_{\varphi(\boldsymbol{u})}\in\mathcal{L}_s}\ \sum_{\mathbf{l}_{-\varphi(\boldsymbol{u})}\in\mathcal{L}_s^{-\varphi(\boldsymbol{u})}} c_m\psi(2^s x_{\varphi(\boldsymbol{u})} - l_{\varphi(\boldsymbol{u})}) \prod_{\substack{\ell=1,\ldots d\\ \ell\neq\varphi(\boldsymbol{u})}} \phi(2^s x_\ell - l_\ell),$$

where $c_m := \int_0^1 x^m\psi(x)\,dx$. Then $B^{\mathtt{W}}_{\boldsymbol{u},\mathbf{0}}(\cdot)$ can be expressed as

$$B^{\mathtt{W}}_{\boldsymbol{u},\mathbf{0}}(\mathbf{x}) = \sum_{s=0}^{\infty} 2^{-sm}\varpi_{\boldsymbol{u},s}(\mathbf{x}) =: \sum_{s=0}^{\infty}\xi_{\boldsymbol{u},s}(\mathbf{x}). \tag{C.3}$$

Moreover, since the series in Equation (C.3) converges uniformly and for $s \geq s_n$, $\varpi^*_{\boldsymbol{u},s}(\mathbf{x})$ is orthogonal to the tensor-product wavelet basis $\mathbf{p}$ with respect to the Lebesgue measure, it follows from Dominated Convergence Theorem that the approximate orthogonality condition holds.

For other $\boldsymbol{\varsigma}$, let

$$\mathscr{B}_{m,\boldsymbol{\varsigma}}(\mathbf{x}) = -\sum_{\boldsymbol{u}\in\Lambda_m}\frac{b^{m-[\boldsymbol{\varsigma}]}}{\boldsymbol{u}!}\partial^{\boldsymbol{u}}\theta(\mathbf{x})B^{\mathtt{W}}_{\boldsymbol{u},\boldsymbol{\varsigma}}(\mathbf{x}/b)$$

where $B^{\mathtt{W}}_{\boldsymbol{u},\boldsymbol{\varsigma}}(\mathbf{x}) = \partial^{\boldsymbol{\varsigma}}B^{\mathtt{W}}_{\boldsymbol{u},\mathbf{0}}(\mathbf{x})$. By assumption that for $\boldsymbol{\varsigma}$ such that $[\boldsymbol{\varsigma}] \leq \varsigma$, $\partial^{\boldsymbol{\varsigma}}\phi$ and $\partial^{\boldsymbol{\varsigma}}\psi$ are continuously differentiable, we have $\sum_{s=0}^{\infty} 2^{-sm}\partial^{\boldsymbol{\varsigma}}\varpi_{\boldsymbol{u},s}(\mathbf{x})$ converge uniformly, and hence we can interchange the differentiation and infinite summation. Therefore, $B^{\mathtt{W}}_{\boldsymbol{u},\boldsymbol{\varsigma}}(\cdot)$ is well defined and continuously differentiable. Then the lipschitz condition on $B^{\mathtt{W}}_{\boldsymbol{u},\boldsymbol{\varsigma}}(\cdot)$ in Assumption II.4 holds.

Let s^* be the orthogonal projection of θ onto $\mathcal{V}_{s_n}$. To complete the proof of part

(2), it suffices to show $\|\partial^{\boldsymbol{\varsigma}}\theta - \partial^{\boldsymbol{\varsigma}} s^* + \mathscr{B}_{m,\boldsymbol{\varsigma}}\|_{L_\infty(\mathcal{X})} \lesssim b^{m+\varrho-[\boldsymbol{\varsigma}]}$. For a given s_n,

$$\begin{aligned}
&\sum_{s=s_n}^{\infty}\sum_{\boldsymbol{\alpha}}\sum_{\mathbf{l}}\langle\theta(\check{\mathbf{x}}), \bar{\psi}_{s\mathbf{l}\boldsymbol{\alpha}}(\check{\mathbf{x}})\rangle\partial^{\boldsymbol{\varsigma}}\bar{\psi}_{s\mathbf{l}\boldsymbol{\alpha}}(\mathbf{x}) \\
={}&\sum_{s=s_n}^{\infty}\sum_{\boldsymbol{\alpha}}\sum_{\mathbf{l}}\Big\langle \sum_{[\boldsymbol{u}]\le m}\partial^{\boldsymbol{u}}\theta(\mathbf{x})\frac{(\check{\mathbf{x}}-\mathbf{x})^{\boldsymbol{u}}}{\boldsymbol{u}!} + \vartheta_n(\check{\mathbf{x}},\mathbf{x}),\ \bar{\psi}_{s\mathbf{l}\boldsymbol{\alpha}}(\check{\mathbf{x}})\Big\rangle\partial^{\boldsymbol{\varsigma}}\bar{\psi}_{s\mathbf{l}\boldsymbol{\alpha}}(\mathbf{x}) \\
={}&b^{m-[\boldsymbol{\varsigma}]}\sum_{s=s_n}^{\infty}\frac{2^{[\boldsymbol{\varsigma}](s_n-s)}}{2^{m(s-s_n)}}\sum_{\boldsymbol{\alpha}}\sum_{\mathbf{l}}\frac{2^{sd}}{2^{-sm}}\Big\langle\sum_{[\boldsymbol{u}]\le m}\partial^{\boldsymbol{u}}\theta(\mathbf{x})\frac{(\check{\mathbf{x}}-\mathbf{x})^{\boldsymbol{u}}}{\boldsymbol{u}!}+\vartheta_n(\check{\mathbf{x}},\mathbf{x}), \\
&\qquad\qquad 2^{-sd/2}\bar{\psi}_{s\mathbf{l}\boldsymbol{\alpha}}(\check{\mathbf{x}})\Big\rangle\ \partial^{\boldsymbol{\varsigma}}\Big(2^{-sd/2}\bar{\psi}_{s\mathbf{l}\boldsymbol{\alpha}}(\mathbf{x})\Big)
\end{aligned}$$

By changing variables, the vanishing moments of the wavelet function and the fact that geometric series converges, the last line uniformly converges to the $\boldsymbol{\varsigma}$th derivative of the approximation error of $\mathcal{V}_{s_n}$, $\mathscr{B}_{m,\boldsymbol{\varsigma}}(\cdot)$ is the leading error and the remainder behaves like $O(b^{m+\varrho-[\boldsymbol{\varsigma}]})$.

(3) By construction of $\tilde{\mathbf{p}}$, $\rho = 1$. It follows from the same argument as that for part (1) and (2) that $\tilde{\mathbf{p}}$ satisfies Assumption II.3 and II.4. Finally, both $\mathbf{p}$ and $\tilde{\mathbf{p}}$ reproduce polynomials of degree no greater than $m-1$. Thus Assumption II.5 holds. The proof is complete. □

Proof of Lemma IV.3.

(1) By construction, each basis function $p_k(\mathbf{x})$ is supported on one subrectangle only, and there are only a fixed number of $p_k(\mathbf{x})$'s which are not identically zero on each subrectangle. Thus Assumption II.3(1) is satisfied. In addition, given one particular subrectangle $\delta_{l_1\ldots l_d}$, store all basis functions supported on $\delta_{l_1\ldots l_d}$ in a vector $\mathbf{p}_{l_1\ldots l_d}$. By Cattaneo and Farrell (2013, Lemma A.3),

$$\mathbf{Q}_{l_1\ldots l_d} := \mathbb{E}[\mathbf{p}_{l_1\ldots l_d}(\mathbf{x}_i)\mathbf{p}_{l_1\ldots l_d}(\mathbf{x}_i)'] \asymp \mathbf{I}_{\dim(\mathbf{R}(\cdot))}$$

where $\mathbf{I}_{\dim(\mathbf{R}(\cdot))}$ is an identity matrix of size $\dim(\mathbf{R}(\cdot))$. $\int_{\delta_{l_1\ldots l_d}}\mathbf{p}_{l_1\ldots l_d}(\mathbf{x})\mathbf{p}_{l_1\ldots l_d}(\mathbf{x})'\,d\mathbf{x}$ is a finite-dimensional matrix with the minimum eigenvalue bounded from below by Ch^d for some $C>0$. Hence for any $\mathbf{a}\in\mathbb{R}^{\dim(\mathbf{R}(\cdot))}$,

$$\mathbf{a}'\int_{\delta_{l_1\ldots l_d}}\mathbf{p}_{l_1\ldots l_d}(\mathbf{x})\mathbf{p}_{l_1\ldots l_d}(\mathbf{x})'\,d\mathbf{x}\,\mathbf{a} \ge Ch^d\mathbf{a}'\mathbf{a}$$

which suffices to show Assumption II.3(2).

To show Assumption II.3(3), simply notice that given any $\mathbf{x}\in\mathcal{X}$, there are only a fixed number of nonzero elements in $\partial^{\mathbf{q}}\mathbf{p}(\mathbf{x})$, and for any $k = 1, \ldots, K$,

$$\sup_{\delta\in\Delta}\ \sup_{\mathbf{x}\in\mathrm{clo}(\delta)} |\partial^{\boldsymbol{\varsigma}} p_k(\mathbf{x})| \lesssim h^{-[\boldsymbol{\varsigma}]} \max_{[\boldsymbol{\alpha}]=m-1} \frac{\boldsymbol{\alpha}!}{(\boldsymbol{\alpha}-\boldsymbol{\varsigma})!}.$$

Moreover, for any $\mathbf{x} \in \mathcal{X}$, there exists some p_k in $\mathbf{p}$ such that for $[\boldsymbol{\varsigma}] \leq m-1$, $|\partial^{\boldsymbol{\varsigma}} p_k(\mathbf{x})| \gtrsim h^{-[\boldsymbol{\varsigma}]}$.

(2) The result directly follows from the proofs of Lemma A.2 and Theorem 3 in Cattaneo and Farrell (2013). The only difference here is that we use shifted Legendre polynomials to re-express the approximating function $s^*(\mathbf{x}) = \mathbf{p}(\mathbf{x})'\boldsymbol{\beta}^*$ and the leading error. Clearly, $\boldsymbol{\beta}^*$ is just a linear combination of coefficients of power series basis defined in their paper. The orthogonality between approximating basis and leading error directly follows from the property of Legendre polynomials and the fact that every basis function is locally supported on only one cell.

(3) By construction of $\tilde{\mathbf{p}}$, $\rho = 1$. It follows from the same argument as that for part (1) and (2) that $\tilde{\mathbf{p}}$ satisfies Assumption II.3 and II.4. Finally, when the degree of piecewise polynomials is increased, $\tilde{\mathbf{p}}$ spans a larger space containing the span of $\mathbf{p}$, and both bases reproduce polynomials of degree no greater than $m-1$. Thus Assumption II.5 holds. □

Proof of Theorem IV.1.

For the integrated variance, define an operator $\boldsymbol{\mathscr{M}}(\cdot)$: $\boldsymbol{\mathscr{M}}(\phi) := \int_{\mathcal{X}} \mathbf{p}(\mathbf{x})\mathbf{p}(\mathbf{x})'\phi(\mathbf{x})d\mathbf{x}$. Then,

$$\int_{\mathcal{X}} \mathbb{V}[\widehat{\theta}_0(\mathbf{x})|\mathbf{X}]w(\mathbf{x})d\mathbf{x} = \frac{1}{n}\,\mathrm{trace}\left[\boldsymbol{\mathscr{M}}(f)^{-1}\boldsymbol{\mathscr{M}}(\sigma^2 f)\boldsymbol{\mathscr{M}}(f)^{-1}\boldsymbol{\mathscr{M}}(w)\right] + o_{\mathbb{P}}\left(\frac{1}{nh^{d+2[\mathbf{q}]}}\right).$$

Let $\boldsymbol{\tau}_k$ be an arbitrary point in $\mathrm{supp}(p_k)$, for $k = 1, \ldots, K$. Define another operator generating $K \times K$ diagonal matrix: $\boldsymbol{\mathscr{D}}(\phi) := \mathrm{diag}\{\phi(\tau_1), \phi(\tau_2), \cdots, \phi(\tau_K)\}$. Then we can write

$$\boldsymbol{\mathscr{M}}(\phi) = \boldsymbol{\mathscr{M}}(1)\boldsymbol{\mathscr{D}}(\phi) - \boldsymbol{\mathscr{E}}(\phi) \tag{C.4}$$

where $\boldsymbol{\mathscr{E}}(\phi)$ can be viewed as errors defined by Eq. (C.4). Then it directly follows that

$$\boldsymbol{\mathscr{M}}(f)^{-1}\boldsymbol{\mathscr{M}}(\sigma^2 f) = [\mathbf{I} - \boldsymbol{\mathscr{U}}(f)]^{-1}[\boldsymbol{\mathscr{D}}(f)^{-1}\boldsymbol{\mathscr{D}}(\sigma^2 f) - \boldsymbol{\mathscr{L}}(f, \sigma^2 f)] \quad \text{and}$$
$$\boldsymbol{\mathscr{M}}(f)^{-1}\boldsymbol{\mathscr{M}}(w) = [\mathbf{I} - \boldsymbol{\mathscr{U}}(f)]^{-1}[\boldsymbol{\mathscr{D}}(f)^{-1}\boldsymbol{\mathscr{M}}(1)^{-1}\boldsymbol{\mathscr{M}}(1)\boldsymbol{\mathscr{D}}(w) - \boldsymbol{\mathscr{L}}(f, w)]$$

where

$$\begin{aligned}\boldsymbol{\mathscr{U}}(\phi) &:= \boldsymbol{\mathscr{D}}(\phi)^{-1}\boldsymbol{\mathscr{M}}(1)^{-1}\boldsymbol{\mathscr{E}}(\phi),\\ \boldsymbol{\mathscr{L}}(\phi,\varphi) &:= \boldsymbol{\mathscr{D}}(\phi)^{-1}\boldsymbol{\mathscr{M}}(1)^{-1}\boldsymbol{\mathscr{E}}(\varphi)\end{aligned}$$

The number of nonzeros on any row or any column of $\boldsymbol{\mathscr{E}}(\phi)$ is bounded by some constant. It may take a multi-layer banded structure when we rectangularize the partition and arrange the ordering of basis functions properly. If $\mathrm{supp}(p_k)\cap\mathrm{supp}(p_l)\neq\varnothing$, then by Assumption II.3 and the continuity of f, the (k,l)th element of $\boldsymbol{\mathscr{M}}(f)$ can be approximated as follows:

$$\int_{\mathcal{X}} p_k(\mathbf{x})p_l(\mathbf{x})f(\mathbf{x})d\mathbf{x} = f(\tau_k)\int_{\mathcal{X}} p_k(\mathbf{x})p_l(\mathbf{x})d\mathbf{x} + o(h^d) \tag{C.5}$$

Moreover, since $\mathcal{X}$ is compact, f is uniformly continuous. Thus we have $\|\boldsymbol{\mathscr{E}}(f)\|_1 = o(h^d)$, $\|\boldsymbol{\mathscr{E}}(f)\|_\infty = o(h^d)$, and then $\|\boldsymbol{\mathscr{E}}(f)\| = o(h^d)$. Since $\|\boldsymbol{\mathscr{D}}(f)^{-1}\| \lesssim 1$ and $\|\boldsymbol{\mathscr{M}}(1)^{-1}\| \lesssim h^{-d}$, we conclude $\|\boldsymbol{\mathscr{U}}(f)\| = o(1)$. For K large enough, we can make $\|\boldsymbol{\mathscr{U}}(f)\| < 1$, and thus $[\mathbf{I}-\boldsymbol{\mathscr{U}}(f)]^{-1} = \mathbf{I}+\boldsymbol{\mathscr{U}}(f)+\boldsymbol{\mathscr{U}}(f)^2+\cdots = \mathbf{I}+\boldsymbol{\mathscr{W}}(f)$ where $\boldsymbol{\mathscr{W}}(f) := \sum_{l=1}^{\infty}\boldsymbol{\mathscr{U}}(f)^l$. Now we can write

$$\begin{aligned}
&\mathrm{trace}\left[\boldsymbol{\mathscr{M}}(f)^{-1}\boldsymbol{\mathscr{M}}(\sigma^2 f)\boldsymbol{\mathscr{M}}(f)^{-1}\boldsymbol{\mathscr{M}}(w)\right]\\
=&\,\mathrm{trace}\left[\Big(\mathbf{I}+\boldsymbol{\mathscr{W}}(f)\Big)\Big(\boldsymbol{\mathscr{D}}(f)^{-1}\boldsymbol{\mathscr{D}}(\sigma^2 f)-\boldsymbol{\mathscr{L}}(f,\sigma^2 f)\Big)\Big(\mathbf{I}+\boldsymbol{\mathscr{W}}(f)\Big)\right.\\
&\qquad\left.\times\Big(\boldsymbol{\mathscr{D}}(f)^{-1}\boldsymbol{\mathscr{M}}(1)^{-1}\boldsymbol{\mathscr{M}}(1)\boldsymbol{\mathscr{D}}(w)-\boldsymbol{\mathscr{L}}(f,w)\Big)\right]\\
=&\,\mathrm{trace}\left[\Big(\boldsymbol{\mathscr{D}}(f)^{-1}\boldsymbol{\mathscr{D}}(\sigma^2 f)+\mathbf{E}_1\Big)\Big(\boldsymbol{\mathscr{D}}(f)^{-1}\boldsymbol{\mathscr{D}}(w)+\mathbf{E}_2\Big)\right]\\
=&\,\mathrm{trace}\left[\boldsymbol{\mathscr{D}}(f)^{-1}\boldsymbol{\mathscr{D}}(\sigma^2 f)\boldsymbol{\mathscr{D}}(f)^{-1}\boldsymbol{\mathscr{D}}(w)+\mathbf{E}_1\boldsymbol{\mathscr{D}}(f)^{-1}\boldsymbol{\mathscr{D}}(w)+\boldsymbol{\mathscr{D}}(f)^{-1}\boldsymbol{\mathscr{D}}(\sigma^2 f)\mathbf{E}_2+\mathbf{E}_1\mathbf{E}_2\right]
\end{aligned}$$

where $\mathbf{E}_1 = -\boldsymbol{\mathscr{L}}(f,\sigma^2 f)+\boldsymbol{\mathscr{W}}(f)\boldsymbol{\mathscr{D}}(f)^{-1}\boldsymbol{\mathscr{D}}(\sigma^2 f)-\boldsymbol{\mathscr{W}}(f)\boldsymbol{\mathscr{L}}(f,\sigma^2 f)$ and $\mathbf{E}_2 = -\boldsymbol{\mathscr{L}}(f,w)+\boldsymbol{\mathscr{W}}(f)\boldsymbol{\mathscr{D}}(f)^{-1}\boldsymbol{\mathscr{D}}(w)-\boldsymbol{\mathscr{W}}(f)\boldsymbol{\mathscr{L}}(f,w)$. By assumptions in the theorem, $\mathrm{vol}(\delta_{\mathbf{x}}) = \prod_{\ell=1}^{d} b_{\mathbf{x},\ell} = \prod_{\ell=1}^{d}\kappa_\ell^{-1}g_\ell(\mathbf{x})^{-1}+o(h^d)$. Hence

$$\begin{aligned}
&\mathrm{trace}\left[\boldsymbol{\mathscr{D}}(f)^{-1}\boldsymbol{\mathscr{D}}(\sigma^2 f)\boldsymbol{\mathscr{D}}(f)^{-1}\boldsymbol{\mathscr{D}}(w)\right]\\
=&\prod_{\ell=1}^{d}\kappa_\ell\,\mathrm{trace}\left[\boldsymbol{\mathscr{D}}(f)^{-1}\boldsymbol{\mathscr{D}}(\sigma^2 f)\boldsymbol{\mathscr{D}}(f)^{-1}\boldsymbol{\mathscr{D}}(w)\boldsymbol{\mathscr{D}}\Big(\prod_{\ell=1}^{d}g_\ell\Big)\boldsymbol{\mathscr{D}}\Big(\prod_{\ell=1}^{d}g_\ell\Big)^{-1}\prod_{\ell=1}^{d}\kappa_\ell^{-1}\right]\\
=&\prod_{\ell=1}^{d}\kappa_\ell\left(\sum_{k=1}^{K}\left[\frac{\sigma^2(\boldsymbol{\tau}_k)w(\boldsymbol{\tau}_k)}{f(\boldsymbol{\tau}_k)}\prod_{\ell=1}^{d}g_\ell(\boldsymbol{\tau}_k)\right]\mathrm{vol}(\delta_{\boldsymbol{\tau}_k})\right)+o(\prod_{\ell=1}^{d}\kappa_\ell)\\
=&\prod_{\ell=1}^{d}\kappa_\ell\times J\int_{\mathcal{X}}\frac{\sigma^2(\mathbf{x})w(\mathbf{x})}{f(\mathbf{x})}\prod_{\ell=1}^{d}g_\ell(\mathbf{x})\,d\mathbf{x}+o(\prod_{\ell=1}^{d}\kappa_\ell)
\end{aligned}$$

It directly follows from the same argument as that in the proof of Agarwal and Studden (1980, Theorem 6.1) that the trace of the remaining terms is $o(\boldsymbol{\kappa}^{\mathbf{1}})$.

For the integrated squared bias, consider the three leading terms B_1, B_2 and B_3 as defined in Equation (A.1). Since the approximate orthogonality condition holds, both B_2 and B_3 are of smaller order and the leading term in the integrated squared bias reduces to B_1 only. For B_1, notice that by assumption of the theorem,

$$\mathscr{B}_{m,\mathbf{q}}(\mathbf{x}) = -\sum_{\boldsymbol{u}\in\Lambda_m} \partial^{\boldsymbol{u}}\theta(\mathbf{x})\Big(\prod_{\ell=1}^{d} \kappa_\ell^{-u_\ell+q_\ell} g_\ell(\mathbf{x})^{-u_\ell+q_\ell}\Big)\mathbf{b}_{\mathbf{x}}^{-\boldsymbol{u}+\mathbf{q}} h_{\mathbf{x}}^{m-[\mathbf{q}]} B_{m,\mathbf{q}}(\mathbf{x}) + o(h^{m-[\mathbf{q}]}).$$

Recall that $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_d)$ and define $\mathbf{g}(\mathbf{x}) := (g_1(\mathbf{x}), \ldots, g_d(\mathbf{x}))$. Given the above fact and using the same notation as in the proof of Theorem II.1, we have

$$\begin{aligned}
\mathrm{B}_1 &= \sum_{\boldsymbol{u}_1,\boldsymbol{u}_2\in\Lambda_m} \int_{\mathcal{X}} \left[\frac{\partial^{\boldsymbol{u}_1}\theta(\mathbf{x})\partial^{\boldsymbol{u}_2}\theta(\mathbf{x}) h_{\mathbf{x}}^{2m-2[\mathbf{q}]} B_{\boldsymbol{u}_1,\mathbf{q}}(\mathbf{x}) B_{\boldsymbol{u}_2,\mathbf{q}}(\mathbf{x})}{\boldsymbol{\kappa}^{\boldsymbol{u}_1+\boldsymbol{u}_2-2\mathbf{q}}\mathbf{g}(\mathbf{x})^{\boldsymbol{u}_1+\boldsymbol{u}_2-2\mathbf{q}}\mathbf{b}_{\mathbf{x}}^{\boldsymbol{u}_1+\boldsymbol{u}_2-2\mathbf{q}}}\right] w(\mathbf{x})d\mathbf{x} + o(h^{2m-2[\mathbf{q}]}) \\
&= \sum_{\boldsymbol{u}_1,\boldsymbol{u}_2\in\Lambda_m} \boldsymbol{\kappa}^{-(\boldsymbol{u}_1+\boldsymbol{u}_2-2\mathbf{q})} \Bigg(\sum_{\delta\in\Delta}\left[\frac{\partial^{\boldsymbol{u}_1} g(t_\delta^*)\partial^{\boldsymbol{u}_2} g(t_\delta^*) w(t_\delta^*)}{\mathbf{g}(t_\delta^*)^{\boldsymbol{u}_1+\boldsymbol{u}_2-2\mathbf{q}}}\right] \\
&\qquad\qquad \times \left[\frac{h_{\mathbf{x}}^{2m-2[\mathbf{q}]}}{\mathbf{b}_{\mathbf{x}}^{\boldsymbol{u}_1+\boldsymbol{u}_2-2\mathbf{q}}}\int_{\delta} B_{\boldsymbol{u}_1,\mathbf{q}}(\mathbf{x}) B_{\boldsymbol{u}_2,\mathbf{q}}(\mathbf{x}) d\mathbf{x}\right]\Bigg) + o(h^{2m-2[\mathbf{q}]}) \\
&= \sum_{\boldsymbol{u}_1,\boldsymbol{u}_2,\in\Lambda_m} \boldsymbol{\kappa}^{-(\boldsymbol{u}_1+\boldsymbol{u}_2-2\mathbf{q})} \eta_{\boldsymbol{u}_1,\boldsymbol{u}_2,\mathbf{q}} \int_{\mathcal{X}} \frac{\partial^{\boldsymbol{u}_1}\theta(\mathbf{x})\partial^{\boldsymbol{u}_2}\theta(\mathbf{x}) w(\mathbf{x})}{\mathbf{g}(\mathbf{x})^{\boldsymbol{u}_1+\boldsymbol{u}_2-2\mathbf{q}}} d\mathbf{x} + o(h^{2m-2[\mathbf{q}]})
\end{aligned}$$

where the last line follows from the integrability of $\partial^{\boldsymbol{u}_1}\theta(\mathbf{x})\partial^{\boldsymbol{u}_2}\theta(\mathbf{x})w(\mathbf{x})/\mathbf{g}(\mathbf{x})^{\boldsymbol{u}_1+\boldsymbol{u}_2-2\mathbf{q}}$ over $\mathcal{X}$. □

Bibliography

Bibliography

Agarwal, G. G., and W. Studden (1980): "Asymptotic Integrated Mean Square Error Using Least Squares and Bias Minimizing Splines," *Annals of Statistics*, 8(6), 1307–1325.

Angrist, J. D., and J. S. Pischke (2008): *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Barrow, D. L., and P. W. Smith (1978): "Asymptotic Properties of Best $L_2[0,1]$ Approximation by Splines with Variable Knots," *Quarterly of Applied Mathematics*, 36(3), 293–304.

Belloni, A., V. Chernozhukov, D. Chetverikov, and I. Fernandez-Val (2018): "Conditional Quantile Processes based on Series or Many Regressors," *Journal of Econometrics*, forthcoming.

——— (2019): "Conditional Quantile Processes based on Series or Many Regressors," *Journal of Econometrics*, forthcoming.

Belloni, A., V. Chernozhukov, D. Chetverikov, and K. Kato (2015): "Some New Asymptotic Theory for Least Squares Series: Pointwise and Uniform Results," *Journal of Econometrics*, 186(2), 345–366.

Bhatia, R. (2013): *Matrix Analysis*. Springer.

Breiman, L., J. Friedman, C. J. Stone, and R. A. Olshen (1984): *Classification and regression trees*. CRC press.

Calonico, S., M. D. Cattaneo, and M. H. Farrell (2018a): "Coverage Error Optimal Confidence Intervals," working paper, University of Michigan.

——— (2018b): "On the Effect of Bias Estimation on Coverage Accuracy in Nonparametric Inference," *Journal of the American Statistical Association*, 113(522), 767–779.

Calonico, S., M. D. Cattaneo, and R. Titiunik (2014): "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs," *Econometrica*, 82(6), 2295–2326.

——— (2015): "Optimal Data-Driven Regression Discontinuity Plots," *Journal of the American Statistical Association*, 110(512), 1753–1769.

Cattaneo, M. D., R. K. Crump, M. H. Farrell, and Y. Feng (2019a): "Binscatter Regressions," arXiv:1902.09615.

——— (2019b): "On Binscatter," arXiv:1902.09608.

Cattaneo, M. D., R. K. Crump, M. H. Farrell, and E. Schaumburg (2019a): "Characteristic-Sorted Portfolios: Estimation and Inference," arXiv:1809.03584.

——— (2019b): "Characteristic-Sorted Portfolios: Estimation and Inference," *Review of Economics and Statistics*, forthcoming.

Cattaneo, M. D., and M. H. Farrell (2011a): "Efficient Estimation of the Dose-Response Function under Ignorability using Subclassification on the Covariates," in *Missing-Data Methods: Cross-sectional Methods and Applications (Advances in Econometrics, vol. 27)*, ed. by D. Drukker, pp. 93–127. Emerald Group Publishing.

——— (2011b): "Efficient Estimation of the Dose Response Function under Ignorability using Subclassification on the Covariates," in *Advances in Econometrics: Missing Data Methods*, ed. by D. Drukker, vol. 27A, pp. 93–127. Emerald Group Publishing Limited.

——— (2013): "Optimal Convergence Rates, Bahadur Representation, and Asymptotic Normality of Partitioning Estimators," *Journal of Econometrics*, 174(2), 127–143.

Cattaneo, M. D., M. H. Farrell, and Y. Feng (2018a): "Large Sample Properties of Partitioning-Based Estimators," arXiv:1804.04916.

——— (2018b): "`lspartition`: Partitioning-Based Least Squares Regression," working paper, University of Michigan.

Cattaneo, M. D., M. Jansson, and W. K. Newey (2018a): "Alternative Asymptotics and the Partially Linear Model with Many Regressors," *Econometric Theory*, 34(2), 277–301.

——— (2018b): "Inference in Linear Regression Models with Many Covariates and Heteroscedasticity," *Journal of the American Statistical Association*, 113(523), 1350–1361.

Chen, X. (2007): "Large Sample Sieve Estimation of Semi-Nonparametric Models," in *Handbook of Econometrics, Volume VI*, ed. by J. J. Heckman, and E. Leamer, pp. 5549–5632. Elsevier Science B.V., New York.

Chen, X., and T. M. Christensen (2015): "Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions," *Journal of Econometrics*, 188(2), 447–465.

Chernozhukov, V., D. Chetverikov, and K. Kato (2014a): "Gaussian Approximation of Suprema of Empirical Processes," *Annals of Statistics*, 42(4), 1564–1597.

——— (2014b): "Anti-Concentration and Honest Adaptive Confidence Bands," *Annals of Statistics*, 42(5), 1787–1818.

——— (2015): "Comparison and Anti-concentration Bounds for Maxima of Gaussian Random Vectors," *Probability Theory and Related Fields*, 162(1), 47–70.

——— (2016): "Empirical and Multiplier Bootstraps for Suprema of Empirical Processes of Increasing Complexity, and Related Gaussian Couplings," *Stochastic Processes and their Applications*, 126(12), 3632–3651.

Chernozhukov, V., S. Lee, and A. M. Rosen (2013): "Intersection bounds: estimation and inference," *Econometrica*, 81(2), 667–737.

Chetty, R., J. N. Friedman, N. Hilger, E. Saez, D. W. Schanzenbach, and D. Yagan (2011): "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR," *Quarterly Journal of Economics*, 126(4), 1593–1660.

Chetty, R., J. N. Friedman, T. Olsen, and L. Pistaferri (2011): "Adjustment Costs, Firm Responses, and Micro vs. Macro Labor Supply Elasticities: Evidence from Danish Tax Records," *Quarterly Journal of Economics*, 126(2), 749–804.

Chetty, R., J. N. Friedman, and J. Rockoff (2014): "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood," *American Economic Review*, 104(9), 2633–2679.

Chetty, R., A. Looney, and K. Kroft (2009): "Salience and Taxation: Theory and Evidence," *American Economic Review*, 99(4), 1145–1177.

Chetty, R., and A. Szeidl (2006): "Marriage, Housing, and Portfolio Choice: A Test of Grossman-Laroque," Working Paper, UC-Berkeley.

Chui, C. K. (2016): *An introduction to wavelets*. Elsevier.

Cochran, W. G. (1968): "The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies," *Biometrics*, 24(2), 295–313.

Cohen, A., I. Daubechies, and P. Vial (1993): "Wavelets on the interval and fast wavelet transforms," *Applied and Computational Harmonic Analysis*, 1(1), 54–81.

Davydov, O. (2001): "Stable Local Bases for Multivariate Spline Spaces," *Journal of Approximation Theory*, 111(2), 267–297.

Demko, S. (1977): "Inverses of Band Matrices and Local Convergence of Spline Projections," *SIAM Journal on Numerical Analysis*, 14(4), 616–619.

Eggermont, P. P. B., and V. N. LaRiccia (2009): *Maximum Penalized Likelihood Estimation: Regression*. Springer, New York, NY.

Fama, E. F. (1976): *Foundations of Finance: Portfolio Decisions and Securities Prices*. Basic Books, New York, NY.

Fan, J., and I. Gijbels (1996): *Local Polynomial Modelling and Its Applications.* Chapman & Hall/CRC, New York.

Friedman, J. H. (1977): “A Recursive Partitioning Decision Rule for Nonparametric Classification,” *IEEE Transactions on Computers*, C-26(4), 404–408.

Genovese, C., and L. Wasserman (2008): “Adaptive confidence bands,” *Annals of Statistics*, 36(2), 875–905.

Genovese, C. R., and L. Wasserman (2005): “Confidence Sets for Nonparametric Wavelet Regression,” *Annals of statistics*, 33(2), 698–729.

Giné, E., and R. Nickl (2016): *Mathematical Foundations of Infinite-Dimensional Statistical Models*, vol. 40. Cambridge University Press.

Györfi, L., M. Kohler, A. Krzyżak, and H. Walk (2002): *A Distribution-Free Theory of Nonparametric Regression.* Springer-Verlag.

Hall, P., and J. Horowitz (2013): “A simple bootstrap method for constructing nonparametric confidence bands for functions,” *The Annals of Statistics*, 41(4), 1892–1921.

Härdle, W., G. Kerkyacharian, D. Picard, and A. Tsybakov (2012): *Wavelets, Approximation, and Statistical Applications.* Springer.

Hastie, T., R. Tibshirani, and J. Friedman (2009): *The elements of statistical learning*, Springer Series in Statistics. Springer-Verlag, New York.

Horowitz, J. L. (2009): *Semiparametric and Nonparametric Methods in Econometrics.* Springer.

Huang, J. Z. (1998): “Projection Estimation in Multiple Regression with Application to Functional ANOVA Models,” *Annals of Statistics*, 26(1), 242–272.

——— (2003): “Local Asymptotics for Polynomial Spline Regression,” *Annals of Statistics*, 31(5), 1600–1635.

Kleven, H. J. (2016): “Bunching,” *Annual Review of Economics*, 8, 435–464.

Komlós, J., P. Major, and G. Tusnády (1975): “An approximation of partial sums of independent RV’-s, and the sample DF. I,” *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 32(1-2), 111–131.

——— (1976): “An approximation of partial sums of independent RV’s, and the sample DF. II,” *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 34(1), 33–58.

Meyer, Y. (1995): *Wavelets and Operators.* Cambridge university press.

Newey, W. K. (1997): "Convergence Rates and Asymptotic Normality for Series Estimators," *Journal of Econometrics*, 79(1), 147–168.

Nobel, A. (1996): "Histogram Regression Estimation Using Data-Dependent Partitions," *Annals of Statistics*, 24(3), 1084–1105.

Ruppert, D., M. P. Wand, and R. Carroll (2009): *Semiparametric Regression.* Cambridge University Press, New York.

Sakhanenko, A. (1985): "Convergence Rate in the Invariance Principle for Non-identically Distributed Variables with Exponential Moments," *Advances in Probability Theory: Limit Theorems for Sums of Random Variables*, pp. 2–73.

——— (1991): "On the Accuracy of Normal Approximation in the Invariance Principle," *Siberian Advances in Mathematics*, 1, 58–91.

Schumaker, L. (2007): *Spline Functions: Basic Theory.* Cambridge University Press.

Starr, E., and B. Goldfarb (2018): "A Binned Scatterplot is Worth a Hundred Regressions: Diffusing a Simple Tool to Make Empirical Research Easier and Better," SSRN Working paper No. 3257345.

Stepner, M. (2014): "Binned Scatterplots: Introducing -binscatter- and Exploring its Applications," 2014 Stata Conference 4, Stata Users Group.

Stone, C. J. (1982): "Optimal Global Rates of Convergence for Nonparametric Regression," *Annals of Statistics*, 10(4), 1040–1053.

Sweldens, W., and R. Piessens (1994): "Asymptotic error expansion of wavelet approximations of smooth functions II," *Numerische Mathematik*, 68(3), 377–401.

Tibshirani, R. J. (2014): "Adaptive Piecewise Polynomial Estimation via Trend Filtering," *The Annals of Statistics*, 42(1), 285–323.

Tropp, J. A. (2012): "User-friendly tail bounds for sums of random matrices," *Foundations of computational mathematics*, 12(4), 389–434.

Tukey, J. W. (1961a): "Curves As Parameters, and Touch Estimation," in *Fourth Berkeley Symposium on Mathematical Statistics and Probability*, ed. by J. Neyman, vol. 1, pp. 681–694.

——— (1961b): "Curves As Parameters, and Touch Estimation," in *Fourth Berkeley Symposium on Mathematical Statistics and Probability*, ed. by J. Neyman, vol. 1, pp. 681–694.

van der vaart, A., and J. Wellner (1996): *Weak Convergence and Empirical Processes: With Application to Statistics.* Springer.

Wasserman, L. (2006): *All of Nonparametric Statistics*. Springer Science & Business Media.

Yurinskii, V. V. (1978): "On the error of the Gaussian approximation for convolutions," *Theory of Probability & Its Applications*, 22(2), 236–247.

Zaitsev, A. Y. (2013): "The Accuracy of Strong Gaussian Approximation for Sums of Independent Random Vectors," *Russian Mathematical Surveys*, 68(4), 721–761.

Zhai, A. (2018): "A High-Dimensional CLT in W2 Distance with Near Optimal Convergence Rate," *Theoretical Probability and Related Fields*, forthcoming.

Zhang, H., and B. H. Singer (2010): *Recursive Partitioning and Applications*. Springer.

Zhou, S., X. Shen, and D. Wolfe (1998): "Local Asymptotics for Regression Splines and Confidence Regions," *Annals of Statistics*, 26(5), 1760–1782.

Zhou, S., and D. A. Wolfe (2000): "On Derivative Estimation in Spline Regression," *Statistica Sinica*, 10(1), 93–108.