Boundary Adaptive Local Polynomial Conditional Density Estimators [*]

Matias D. Cattaneo[†] Rajita Chandak[†] Michael Jansson[‡]
Xinwei Ma[§]

November 12, 2022

Abstract

We begin by introducing a class of conditional density estimators based on local polynomial techniques. The estimators are boundary adaptive and easy to implement. We then study the (pointwise and) uniform statistical properties of the estimators, offering characterizations of both probability concentration and distributional approximation. In particular, we establish precise optimal uniform convergence rates in probability and valid Gaussian distributional approximations for the t-statistic process indexed over the data support. We also discuss implementation issues such as consistent estimation of the covariance function of the Gaussian approximation, optimal integrated mean squared error bandwidth selection, and valid robust bias-corrected inference. We illustrate the applicability of our results by constructing valid confidence bands and hypothesis tests for both parametric specification and shape constraints, explicitly characterizing their approximation errors. A companion R software package implementing our main results is provided.

Keywords: Conditional distribution estimation, local polynomial methods, strong approximations, uniform inference, confidence bands, specification testing.

[*] We thank Jianqing Fan, Jason Klusowski and Will Underwood for comments. Cattaneo gratefully acknowledges financial support from the National Science Foundation through grant SES-1947805 and from the National Institute of Health (R01 GM072611-16), and Jansson gratefully acknowledges financial support from the National Science Foundation through grant SES-1947662 and the research support of CREATES. A companion R package is available at https://nppackages.github.io/lpcde/.

[†] Department of Operations Research and Financial Engineering, Princeton University.

[‡] Department of Economics, UC Berkeley and *CREATES*.

[§] Department of Economics, UC San Diego.

1 Introduction

Suppose that $(y_1, \mathbf{x}_1^{\mathrm{T}}), (y_2, \mathbf{x}_2^{\mathrm{T}}), \ldots, (y_n, \mathbf{x}_n^{\mathrm{T}})$ is a random sample from a distribution supported on $\mathcal{Y} \times \mathcal{X}$, where $\mathcal{Y} \subset \mathbb{R}$ and $\mathcal{X} \subset \mathbb{R}^d$ are compact. Letting $F(y|\mathbf{x})$ be the conditional cumulative distribution function (CDF) of y_i given $\mathbf{x}_i$, important parameters of interest in statistics, econometrics, and many other data science disciplines, are the conditional probability density function (PDF) and derivatives thereof:

$$f^{(\vartheta)}(y|\mathbf{x}) = \frac{\partial^{1+\vartheta}}{\partial y^{1+\vartheta}} F(y|\mathbf{x}), \qquad \vartheta \in \mathbb{N}_0 = \{0, 1, 2, \ldots\},$$

where, in particular, $f(y|\mathbf{x}) = f^{(0)}(y|\mathbf{x})$ is the conditional Lebesgue density of y_i given $\mathbf{x}_i$.

Estimation and inference methodology for (conditional) PDFs has a long tradition in statistics (e.g., Wand and Jones, 1995; Wasserman, 2006; Simonoff, 2012; Scott, 2015, and references therein). Unfortunately, without specific modifications, smoothing methods employing kernel, series, or other local approximation techniques are invalid at or near boundary points of $\mathcal{Y} \times \mathcal{X}$. To address this challenge, we introduce a boundary adaptive and closed-form nonparametric estimator of $f^{(\vartheta)}(y|\mathbf{x})$ based on local polynomial techniques (Fan and Gijbels, 1996) and provide an array of distributional approximation results that are valid (pointwise and) uniformly over $\mathcal{Y} \times \mathcal{X}$. In particular, we obtain a uniformly valid stochastic linear representation for the estimator and develop uniform inference methods based on strong approximation techniques leading to, for example, asymptotically valid confidence bands and specification testing methods for $f^{(\vartheta)}(y|\mathbf{x})$ with careful characterization of their associated approximation errors.

To motivate our proposed estimation approach, suppose first that $\mathbf{x} \in \mathbb{R}^d$ is an evaluation point at which an estimator $\widehat{F}(\cdot|\mathbf{x})$ of $F(\cdot|\mathbf{x})$ is available. Then, for $y \in \mathbb{R}$, a natural estimator of $f^{(\vartheta)}(y|\mathbf{x})$ is the local polynomial estimator

$$\widehat{f}^{(\vartheta)}(y|\mathbf{x}) = \mathbf{e}_{1+\vartheta}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}(y|\mathbf{x}), \qquad \widehat{\boldsymbol{\beta}}(y|\mathbf{x}) = \operatorname*{argmin}_{\mathbf{u} \in \mathbb{R}^{\mathfrak{p}+1}} \sum_{i=1}^{n} \Big(\widehat{F}(y_i|\mathbf{x}) - \mathbf{p}(y_i - y)^{\mathrm{T}} \mathbf{u} \Big)^2 K_h(y_i; y), \quad (1)$$

where $\mathfrak{p} \geqslant 1 + \vartheta$ is the order of the polynomial basis $\mathbf{p}(u) = (1, u/1!, u^2/2!, \ldots, u^{\mathfrak{p}}/\mathfrak{p}!)^{\mathrm{T}}$, $\mathbf{e}_l$ is the conformable $(1 + l)$-th unit vector, and $K_h(u; y) = K((u - y)/h)/h$ for some kernel function K and some positive bandwidth h. In this paper, we employ the following $\mathfrak{q}$-th order local polynomial regression estimator of $F(y|\mathbf{x})$:

$$\widehat{F}(y|\mathbf{x}) = \mathbf{e}_{\mathbf{0}}^{\mathrm{T}} \widehat{\boldsymbol{\gamma}}(y|\mathbf{x}), \qquad \widehat{\boldsymbol{\gamma}}(y|\mathbf{x}) = \operatorname*{argmin}_{\mathbf{u} \in \mathbb{R}^{\mathfrak{q}_d+1}} \sum_{i=1}^{n} \big(\mathbb{1}(y_i \leqslant y) - \mathbf{q}(\mathbf{x}_i - \mathbf{x})^{\mathrm{T}} \mathbf{u} \big)^2 L_b(\mathbf{x}_i; \mathbf{x}),$$

where, using standard multi-index notation, $\mathbf{q}(\mathbf{u})$ denotes the $\mathfrak{q}_d$-dimensional vector collecting the ordered elements $\mathbf{u}^{\boldsymbol{\nu}}/\boldsymbol{\nu}!$ for $0 \leqslant |\boldsymbol{\nu}| \leqslant \mathfrak{q}$, where $\mathbf{u}^{\boldsymbol{\nu}} = u_1^{\nu_1} u_2^{\nu_2} \cdots u_d^{\nu_d}$, $|\boldsymbol{\nu}| = \nu_1 + \nu_2 + \cdots + \nu_d$ for $\mathbf{u} = (u_1, u_2, \ldots, u_d)^{\mathrm{T}}$, $\boldsymbol{\nu} = (\nu_1, \nu_2, \ldots, \nu_d)^{\mathrm{T}}$, $\mathfrak{q}_d = (d + \mathfrak{q})!/(\mathfrak{q}!d!) - 1$, and $L_b(\mathbf{u}; \mathbf{x}) = L((\mathbf{u} - \mathbf{x})/b)/b^d$ for some (multivariate) kernel function L and positive bandwidth b.

By virtue of being based on a local polynomial smoothing approach, the estimator $\widehat{f}^{(\vartheta)}(y|\mathbf{x})$ is not only intuitive, but also boundary adaptive. Furthermore, $\widehat{f}^{(\vartheta)}(y|\mathbf{x})$ admits a simple closed-form representation as shown in Equation (2) below, making it easy to implement. These features follow directly from its construction: unlike classical kernel-based conditional density (derivative) estimators, which seek to approximate the conditional PDF indirectly (e.g., by constructing a ratio of two unconditional kernel-based density estimators), our proposed estimator applies local polynomial techniques directly to the conditional CDF estimator $\widehat{F}(y|\mathbf{x})$, which itself has automatic boundary carpentry. In addition, our approach offers an easy way to construct higher-order kernels to reduce misspecification (or smoothing) bias via the choice of polynomial orders $\mathfrak{p}$ and $\mathfrak{q}$, while still retaining all its other appealing features. We discuss related literature further below.

We present two main uniform results for our proposed estimator. First, we provide precise uniform probability concentration bounds associated with a stochastic linear representation of $\widehat{f}^{(\vartheta)}(y|\mathbf{x})$. In addition to being useful for the purposes of characterizing the distributional properties of the conditional density estimator itself, the first main result can be used to analyze multi-step estimation and inference procedures whenever $\widehat{f}^{(\vartheta)}(y|\mathbf{x})$ enters as a preliminary step. As a by-product of the development of the first main result, we obtain a related class of conditional density estimators based on local smoothing, which may be of independent interest. For details, see the supplemental appendix.

Our second main result employs the stochastic linear representation of $\widehat{f}^{(\vartheta)}(y|\mathbf{x})$ to establish a valid strong approximation for the standardized t-statistic stochastic process based on $\widehat{f}^{(\vartheta)}(y|\mathbf{x})$ and indexed over $\mathcal{Y} \times \mathcal{X}$. This result is established using a powerful result due to Rio (1994), which in turn builds on the celebrated Hungarian construction (Komlós *et al.*, 1975). As is well known, t-statistic stochastic processes based on kernel-based nonparametric estimators are not asymptotically tight and, as a consequence, do not converge weakly as a process indexed over $\mathcal{Y} \times \mathcal{X}$ (van der Vaart and Wellner, 1996; Giné and Nickl, 2016). Nevertheless, using strong approximations to such processes, it is possible to deduce distributional approximations for functionals thereof employing anti-concentration (Chernozhukov *et al.*, 2014a). For example, combining these ideas, we obtain valid distributional approximations for the suprema of the standardized t-statistic stochastic process based on $\widehat{f}^{(\vartheta)}(y|\mathbf{x})$ with approximation rates that are faster than those currently available in the literature for the case of $d = 1$ (e.g., Remark 3.1(ii) in Chernozhukov *et al.*

2014b).

In addition to our two main uniform estimation and distributional results, we discuss several implementation results that are useful for practice. First, we present a covariance function estimator for the Gaussian approximation and prove its uniform consistency. This result enables us to estimate the statistical uncertainty underlying the Gaussian approximation for a feasible version of the t-statistic process. Second, we discuss optimal bandwidth selection based on an asymptotic approximation to the integrated mean squared error (IMSE) of the estimator $\widehat{f}^{(\vartheta)}(y|\mathbf{x})$. This result allows us to implement our proposed estimator using point estimation optimal data-driven bandwidth selection rules. Finally, we employ robust bias correction (Calonico *et al.*, 2018, 2022) to develop valid inference methods based on the Gaussian approximation when using the estimated covariance function and IMSE-optimal bandwidth rule.

We illustrate our theoretical and methodological results with three substantive applications: we construct valid confidence bands for the unknown conditional density function and derivatives thereof, and develop valid hypothesis testing procedures for parametric specification and shape constraints of $f^{(\vartheta)}(y|\mathbf{x})$. All these methods are data-driven and, in some cases, optimal in terms of probability and/or distributional concentration, possibly up to $\log(n)$ factors. Furthermore, thanks to the precise probability approximation errors we obtain via strong approximation and other exponential concentration methods, we are able to characterize precise coverage error and rejection probability error rates for all the feasible inference procedures considered. We also present a small simulation study supporting our theoretical work. All proofs are given in the supplemental appendix, which considers a more general setup and also offers additional technical and methodological results of potential independent interest. Last but not least, we provide a general purpose `R` software package implementing the main results in this paper.

1.1 Related Literature

Our paper contributes to the literature on kernel-based conditional density estimation and inference. See Hall *et al.* (1999), De Gooijer and Zerom (2003) and Hall *et al.* (2004) for earlier reviews, and Wand and Jones (1995), Wasserman (2006), Simonoff (2012) and Scott (2015) for textbook introductions.

Traditional methods for conditional density estimation typically employ ratios of unconditional kernel density estimators, non-linear kernel-based derivative of distribution function estimators, or local polynomial estimators based on some preliminary density-like approximation. In particular, in the leading special case of $\vartheta = 0$, the closest antecedent to our proposed conditional density estimator is the local polynomial conditional density estimator introduced by Fan *et al.* (1996). Unlike their proposal, our estimator is boundary adaptive without requiring knowledge of the support $\mathcal{Y}$. To further highlight the connections between the estimators, notice that (in our notation) their estimator takes the form

$$\widehat{f}_{\texttt{FYT}}(y|\mathbf{x}) = \mathbf{e}_{\mathbf{0}}^{\mathrm{T}} \operatorname*{argmin}_{\mathbf{u}\in\mathbb{R}^{\mathfrak{q}_d+1}} \sum_{i=1}^{n} \left(K_h(y_i;y) - \mathbf{q}(\mathbf{x}_i-\mathbf{x})^{\mathrm{T}}\mathbf{u}\right)^2 L_b(\mathbf{x}_i;\mathbf{x}),$$

where, by the way of motivation, Fan *et al.* (1996) note that if y belongs to the interior of $\mathcal{Y}$, then

$$\lim_{h\downarrow 0} \mathbb{E}[K_h(y_i;y)|\mathbf{x}_i=\mathbf{x}] = \lim_{h\downarrow 0} \int_{\mathcal{Y}} K_h(u;y) f(u|\mathbf{x})\mathrm{d}u = f(y|\mathbf{x}).$$

The displayed equality does not hold when y is a boundary point of $\mathcal{Y}$, and for this reason their estimator has poor bias properties when y is on (or near) the boundary of $\mathcal{Y}$.

Our estimator of $f(y|x)$ is similar to their estimator insofar as it can be interpreted as

$$\widehat{f}(y|\mathbf{x}) = \mathbf{e}_{\mathbf{0}}^{\mathrm{T}} \operatorname*{argmin}_{\mathbf{u}\in\mathbb{R}^{\mathfrak{q}_d+1}} \sum_{i=1}^{n} \left(\widehat{K}_h(y_i;y) - \mathbf{q}(\mathbf{x}_i-\mathbf{x})^{\mathrm{T}}\mathbf{u}\right)^2 L_b(\mathbf{x}_i;\mathbf{x}),$$

where

$$\widehat{K}_h(u;y) = \mathbf{e}_1^{\mathrm{T}} \operatorname*{argmin}_{\mathbf{u}\in\mathbb{R}^{\mathfrak{p}+1}} \sum_{j=1}^{n} \left(\mathbb{1}(u \leqslant y_j) - \mathbf{p}(y_j-y)^{\mathrm{T}}\mathbf{u}\right)^2 K_h(y_j,y).$$

In other words, our conditional density estimator $\widehat{f}(y|\mathbf{x})$ can be interpreted as the estimator proposed by Fan *et al.* (1996) but with a different (data-driven) kernel function, $\widehat{K}_h(y_i;y)$, smoothing out the variable y. The implied kernel $\widehat{K}_h(u;y)$ satisfies

$$\int_{\mathcal{Y}} \widehat{K}_h(u;y) f(u|\mathbf{x})\mathrm{d}u = \mathbf{e}_1^{\mathrm{T}} \operatorname*{argmin}_{\mathbf{u}\in\mathbb{R}^{\mathfrak{p}+1}} \sum_{j=1}^{n} \left(F(y_j|\mathbf{x}) - \mathbf{p}(y_j-y)^{\mathrm{T}}\mathbf{u}\right)^2 K_h(y_j,y).$$

Standard local polynomial reasoning therefore suggests that our estimator should enjoy good bias properties even when y is on (or near) the boundary of $\mathcal{Y}$. Indeed, our estimator offers automatic boundary carpentry, higher-order derivative estimation, and automatic higher-order kernel constructions, among other features.

More generally, classical methods for conditional density estimation are not boundary adaptive without specific modifications, and in some cases do not have a closed-form representation. Boundary carpentry could be achieved by employing boundary-corrected kernels in some cases, but such conditional density estimation methods do not appear to have been considered in the literature before. Therefore, our first contribution is to introduce a novel automatic boundary adaptive, closed-form conditional density (derivative) estimator. Our proposed construction does not rely on boundary-corrected kernels explicitly nor does it exploit knowledge of the support of the data in its construction, but it rather builds on the idea that automatic boundary-adaptive density estimators can be constructed using local polynomial methods to smooth out the (discontinuous) distribution function (Cattaneo *et al.*, 2020).

1.2 Notation and Assumptions

To simplify the presentation, in the remainder of this paper we set L to be the product kernel based on K; that is, $L(\mathbf{u}) = K(u_1)K(u_2)\cdots K(u_d)$. We also employ the same bandwidth, $b = h$, in the construction of our proposed estimator, and assume $\mathfrak{q} = \mathfrak{p}-\vartheta-1 \geqslant 0$ throughout. General results are available in the supplemental appendix.

Limits are taken with respect to the sample size tending to infinity (i.e., $n \to \infty$). For two non-negative sequences a_n and b_n, $a_n \lesssim b_n$ means that a_n/b_n is bounded and $a_n \lesssim_{\mathbb{P}} b_n$ means that a_n/b_n is bounded in probability. Constants that do not depend on the sample size or the bandwidth will be denoted by $\mathfrak{c}$, $\mathfrak{c}_1$, $\mathfrak{c}_2$, etc.

We also introduce the notation $\lesssim_{\mathbb{TC}}$, which not only provides an asymptotic order but also controls the tail probability. To be precise, $a_n \lesssim_{\mathbb{TC}} b_n$ implies that for any $\mathfrak{c}_1 > 0$, there exists some $\mathfrak{c}_2$ such that

$$\limsup_{n\to\infty}\ n^{\mathfrak{c}_1}\ \mathbb{P}\big[a_n \geqslant \mathfrak{c}_2 b_n\big] < \infty.$$

Finally, let $\mathbf{X} = (\mathbf{x}_1^{\mathrm{T}}, \ldots, \mathbf{x}_n^{\mathrm{T}})^{\mathrm{T}}$ and $\mathbf{Y} = (y_1, \ldots, y_n)^{\mathrm{T}}$ be the data matrices. We make the following assumptions on the joint distribution and the kernel function.

Assumption 1 (DGP)
(i) $(y_1, \mathbf{x}_1^{\mathrm{T}})^{\mathrm{T}}, \ldots, (y_n, \mathbf{x}_n^{\mathrm{T}})^{\mathrm{T}}$ *is a random sample from a distribution supported on* $\mathcal{Y}\times\mathcal{X} = [0,1]^{1+d}$*, and the joint Lebesgue density,* $f(y,\mathbf{x})$*, is continuous and bounded away from zero on* $\mathcal{Y}\times\mathcal{X}$*. (ii)* $f^{(\mathfrak{p})}(y|\mathbf{x})$ *exists and is continuous. (iii)* $\partial^{\boldsymbol{\nu}} f^{(\vartheta)}(y|\mathbf{x})/\partial\mathbf{x}^{\boldsymbol{\nu}}$ *exists and is continuous for all* $|\boldsymbol{\nu}| = \mathfrak{p}-\vartheta$.

Assumption 2 (Kernel)
K *is a symmetric, Lipschitz continuous PDF supported on* $[-1,1]$.

1.3 Paper Organization

Section 2 first presents a stochastic linear representation for $\widehat{f}^{(\vartheta)}(y|\mathbf{x})$ uniformly over $\mathcal{Y}\times\mathcal{X}$. We then discuss the main theoretical properties of our proposed estimator, offering precise concentration characterizations in probability and in distribution uniformly over $\mathcal{Y} \times \mathcal{X}$. Section 3 deploys our theoretical results to three applications: construction of confidence bands, parametric specification hypothesis testing, and shape constrained hypothesis testing for $f^{(\vartheta)}(y|\mathbf{x})$. Section 4 reports a small simulation study employing our companion `R` package (Cattaneo *et al.*, 2022). Section 5 concludes. The supplemental appendix contains additional results not included here to simplify the presentation: (i) boundary adaptive estimators for the CDF and its derivatives with respect to $\mathbf{x}$, (ii) a new class of estimators based on non-random local smoothing that is less sensitive to "low" density regions, (iii) complete proofs, (iv) details on bandwidth selection, (v) alternative covariance function estimators, and (vi) other technical lemmas that may be of independent interest. Leveraging the uniform stochastic linear representation, we also discuss in the supplemental appendix how our estimator can be easily adjusted to satisfy additional properties, such as nonnegativity and integrating to 1. Interestingly, the latter requires introducing a normalization factor which affects the strong approximation in nontrivial ways, leading in particular to a different Gaussian process distributional approximation.

2 Main Results

This section presents four main theoretical results. First, we provide a stochastic linearization of our estimator. Based on this representation, we obtain a uniform probability concentration result for $\widehat{f}^{(\vartheta)}(y|\mathbf{x})$. Next, we obtain valid strong approximation results for the standardized t-process based on $\widehat{f}^{(\vartheta)}(y|\mathbf{x})$. Finally, we develop a feasible distributional approximation for the suprema of the t-process. To accomplish the latter, we obtain a uniform consistency result for an estimator of the covariance function. The supplemental appendix discusses pointwise in $(y, \mathbf{x})$ results under slightly weaker conditions: because our uniform results are sharp, the only substantive difference is that in the pointwise results the $\log(n)$ terms can be dropped.

2.1 Stochastic Linearization

Our proposed estimator can be written in closed-form as

$$\widehat{f}^{(\vartheta)}(y|\mathbf{x}) = \mathbf{e}_{1+\vartheta}^{\mathrm{T}} \widehat{\mathbf{S}}_y^{-1} \widehat{\mathbf{R}}_{y,\mathbf{x}} \widehat{\mathbf{S}}_{\mathbf{x}}^{-1} \mathbf{e}_{\mathbf{0}}, \tag{2}$$

where

$$\widehat{\mathbf{S}}_y = \frac{1}{n}\sum_{i=1}^{n}\mathbf{p}\Big(\frac{y_i - y}{h}\Big)\frac{1}{h}\mathbf{P}\Big(\frac{y_i - y}{h}\Big)^{\mathrm{T}}, \qquad \widehat{\mathbf{S}}_{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{q}\Big(\frac{\mathbf{x}_i - \mathbf{x}}{h}\Big)\frac{1}{h^d}\mathbf{Q}\Big(\frac{\mathbf{x}_i - \mathbf{x}}{h}\Big)^{\mathrm{T}},$$

$$\widehat{\mathbf{R}}_{y,\mathbf{x}} = \frac{1}{n^2 h^{1+\vartheta}}\sum_{j=1}^{n}\sum_{i=1}^{n}\frac{1}{h}\mathbf{P}\Big(\frac{y_j - y}{h}\Big)\frac{1}{h^d}\mathbf{Q}\Big(\frac{\mathbf{x}_i - \mathbf{x}}{h}\Big)^{\mathrm{T}}\mathbb{1}(y_i \leqslant y_j),$$

with the definitions $\mathbf{P}(u) = \mathbf{p}(u)K(u)$ and $\mathbf{Q}(\mathbf{u}) = \mathbf{q}(\mathbf{u})L(\mathbf{u})$, which absorb the kernel function into the basis. The matrices $\widehat{\mathbf{S}}_{\mathrm{y}}$ and $\widehat{\mathbf{S}}_{\mathbf{x}}$ are well approximated by $\mathbf{S}_y$ and $\mathbf{S}_{\mathbf{x}}$, respectively, where

$$\mathbf{S}_y = \int_{\mathcal{Y}}\mathbf{p}\Big(\frac{u - y}{h}\Big)\frac{1}{h}\mathbf{P}\Big(\frac{u - y}{h}\Big)^{\mathrm{T}}\mathrm{d}F_y(u), \qquad \mathbf{S}_{\mathbf{x}} = \int_{\mathcal{X}}\mathbf{q}\Big(\frac{\mathbf{u} - \mathbf{x}}{h}\Big)\frac{1}{h^d}\mathbf{Q}\Big(\frac{\mathbf{u} - \mathbf{x}}{h}\Big)^{\mathrm{T}}\mathrm{d}F_{\mathbf{x}}(\mathbf{u}),$$

with F_y and $F_{\mathbf{x}}$ denoting the CDFs of y_i and $\mathbf{x}_i$, respectively. Obtaining and characterizing a simple approximation to the matrix $\widehat{\mathbf{R}}_{y,\mathbf{x}}$ requires a little more care, but the end result can be combined with the results for $\widehat{\mathbf{S}}_{\mathrm{y}}$ and $\widehat{\mathbf{S}}_{\mathbf{x}}$ to obtain the following uniform stochastic linear representation for $\widehat{f}^{(\vartheta)}(y|\mathbf{x})$:

Lemma 1 (Stochastic Linearization)

Suppose Assumptions 1 and 2 hold. If $h \to 0$ and if $nh^{1+d}/\log(n) \to \infty$, then

$$\sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}}\Big|\widehat{f}^{(\vartheta)}(y|\mathbf{x}) - f^{(\vartheta)}(y|\mathbf{x}) - \bar{f}^{(\vartheta)}(y|\mathbf{x})\Big| \lesssim_{\mathbb{TC}} \mathtt{r}_{\mathtt{SL}},$$

where $\bar{f}^{(\vartheta)}(y|\mathbf{x}) = n^{-1}\sum_{i=1}^{n}\mathscr{K}^{\circ}_{\vartheta,h}(y_i, \mathbf{x}_i; y, \mathbf{x})$,

$$\mathscr{K}^{\circ}_{\vartheta,h}(a, \mathbf{b}; y, \mathbf{x}) = \frac{1}{h^{1+\vartheta}}\mathbf{e}_{1+\vartheta}^{\mathrm{T}}\mathbf{S}_y^{-1}\int_{\mathcal{Y}}\Big(\mathbb{1}(a \leqslant u) - F(u|\mathbf{b})\Big)\frac{1}{h}\mathbf{P}\Big(\frac{u - y}{h}\Big)\mathrm{d}F_y(u)\frac{1}{h^d}\mathbf{Q}\left(\frac{\mathbf{b} - \mathbf{x}}{h}\right)^{\mathrm{T}}\mathbf{S}_{\mathbf{x}}^{-1}\mathbf{e}_{\mathbf{0}},$$

and

$$\mathtt{r}_{\mathtt{SL}} = h^{\mathfrak{p}-\vartheta} + \frac{\log(n)}{\sqrt{n^2 h^{1+2\vartheta+d+(2\vee d)}}}.$$

The properties of $\widehat{f}^{(\vartheta)}(y|\mathbf{x})$ are thus governed by the properties of the stochastic linear representation. In the supplemental appendix, we demonstrate important features of $\mathscr{K}^{\circ}_{\vartheta,h}$, such as boundedness and Lipschitz continuity, which will play a crucial role in our strong approximation results. We also bound the uniform covering number for the class of functions formed by varying the evaluation point. This uniform covering number result takes into account the fact that the shape of $\mathscr{K}^{\circ}_{\vartheta,h}$ changes across different evaluation points and bandwidths, and is established using a generic result, which may be of independent interest.

Remark 1 (Imposing additional constraints) Specific applications may require additional constraints on the estimates. For example, setting $\vartheta = 0$ (probability density function), it may be desirable to require that the estimator is nonnegative and integrates to 1. While nonnegativity can be directly imposed on the local regression step, the latter requires normalizing the estimator globally. With a slight abuse of notation, we can define

$$\widehat{f}(y|\mathbf{x}) = \max\left\{\mathbf{e}_1^{\mathrm{T}}\widehat{\mathbf{S}}_y^{-1}\widehat{\mathbf{R}}_{y,\mathbf{x}}\widehat{\mathbf{S}}_{\mathbf{x}}^{-1}\mathbf{e}_{\mathbf{0}}\ ,\ 0\right\}, \qquad \breve{f}(y|\mathbf{x}) = \frac{\widehat{f}(y|\mathbf{x})}{\int_{\mathcal{Y}}\widehat{f}(u|\mathbf{x})\mathrm{d}u}.$$

The normalized estimator, $\breve{f}(y|\mathbf{x})$, admits a different stochastic linear representation:

$$\sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}}\left|\breve{f}(y|\mathbf{x}) - f(y|\mathbf{x}) - \left(\bar{f}(y|\mathbf{x}) - f(y|\mathbf{x})\int_{\mathcal{Y}}\bar{f}(u|\mathbf{x})\mathrm{d}u\right)\right| \lesssim_{\mathbb{TC}} \mathtt{r}_{\mathtt{SL}},$$

where $\bar{f}$ and $\mathtt{r}_{\mathtt{SL}}$ are defined in Lemma 1 above. See the supplemental appendix for additional details, including the uniform Gaussian approximation result for this normalized estimator.

Remark 2 (Local smoothing based estimator) In the supplemental appendix, we also study an intermediate estimator, which replaces the local regression in Equation (1) by local smoothing. This intermediate estimator has some distinctive features that may be of independent interest in some settings: due to the non-random weighting employed, it is less sensitive to "low" density regions, but it requires ex-ante knowledge of the support $\mathcal{Y}\times\mathcal{X}$.

In the remainder of the paper, we use the representation established by Lemma 1 to study the properties of $\widehat{f}^{(\vartheta)}(y|\mathbf{x})$. The lemma is also useful when studying the properties of multi-step nonparametric and semiparametric procedures employing conditional density estimators as preliminary estimators, but to conserve space we do not discuss those applications here.

2.2 Uniform Probability Concentration

The following theorem gives a uniform probability concentration result for our conditional density and derivative estimator.

Theorem 1 (Probability Concentration)
Suppose Assumptions 1 and 2 hold. If $h \to 0$ and if $nh^{1+d}/\log(n) \to \infty$, then

$$\sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}}\left|\widehat{f}^{(\vartheta)}(y|\mathbf{x}) - f^{(\vartheta)}(y|\mathbf{x})\right| \lesssim_{\mathbb{TC}} h^{\mathfrak{p}-\vartheta} + \sqrt{\frac{\log(n)}{nh^{1+d+2\vartheta}}}.$$

In the theorem, $h^{\mathfrak{p}-\vartheta}$ stems from a bias term whose magnitude coincides with that of the pointwise bias at interior evaluation points. As a consequence, the theorem implies that the estimator is boundary adaptive. The other term represents "noise", whose magnitude is larger than its counterpart in Lemma 1. As a consequence, the theorem implies that the estimation error $\widehat{f}^{(\vartheta)}(y|\mathbf{x}) - f^{(\vartheta)}(y|\mathbf{x})$ is asymptotically equivalent to $\bar{f}^{(\vartheta)}(y|\mathbf{x})$ whenever the bias is asymptotically negligible. By setting $h = (\log(n)/n)^{\frac{1}{1+d+2\mathfrak{p}}}$, it follows from the theorem that the estimator achieves the minimax optimal uniform convergence rate established by Khas'minskii (1979): $(\log(n)/n)^{\frac{\mathfrak{p}-\vartheta}{1+d+2\mathfrak{p}}}$.

Section 3 characterizes the leading bias and variance constants and then uses these to obtain (approximate) IMSE-optimal bandwidths. When doing so, we follow the local polynomial regression literature (Fan and Gijbels, 1996) and require $\mathfrak{p} - \vartheta$ to be even so that the leading bias term is easily characterized, but this condition is not required in Theorem 1; see the supplemental appendix for more general results.

2.3 Strong Approximation

Next, we study the distributional properties of the process $(\widehat{\mathbb{S}}_\vartheta(y, \mathbf{x}) : y \in \mathcal{Y}, \mathbf{x} \in \mathcal{X})$, where

$$\widehat{\mathbb{S}}_\vartheta(y, \mathbf{x}) = \frac{\widehat{f}^{(\vartheta)}(y|\mathbf{x}) - f^{(\vartheta)}(y|\mathbf{x})}{\sqrt{\mathsf{V}_\vartheta(y, \mathbf{x})}}, \tag{3}$$

with

$$\mathsf{V}_\vartheta(y, \mathbf{x}) = \frac{1}{n}\mathbb{V}\left[\mathscr{K}^\circ_{\vartheta,h}(y_i, \mathbf{x}_i; y, \mathbf{x})\right].$$

Using elementary tools, Theorem SA-2.1 in the supplemental appendix obtains a pointwise Gaussian approximation to $\widehat{\mathbb{S}}_\vartheta(y, \mathbf{x})$. As is well-known, however, the process $\widehat{\mathbb{S}}_\vartheta$ is not asymptotically tight and therefore does not converge weakly to a Gaussian process in $\ell^\infty(\mathcal{Y} \times \mathcal{X})$, the set of uniformly bounded real-valued functions on $\mathcal{Y} \times \mathcal{X}$ equipped with the uniform norm (van der Vaart and Wellner, 1996; Giné and Nickl, 2016). To obtain a uniform distributional approximation, we use the result of Rio (1994) and establish a strong approximation result for $(\widehat{\mathbb{S}}_\vartheta(y, \mathbf{x}) : y \in \mathcal{Y}, \mathbf{x} \in \mathcal{X})$. To state the result, define the covariance function

$$\rho_\vartheta(y, \mathbf{x}, y', \mathbf{x}') = \frac{\mathsf{C}_\vartheta(y, \mathbf{x}, y', \mathbf{x}')}{\sqrt{\mathsf{V}_\vartheta(y, \mathbf{x})}\sqrt{\mathsf{V}_\vartheta(y', \mathbf{x}')}},$$

where

$$\mathsf{C}_\vartheta(y, \mathbf{x}, y', \mathbf{x}') = \frac{1}{n}\mathbb{E}\left[\mathscr{K}^\circ_{\vartheta,h}(y_i, \mathbf{x}_i; y, \mathbf{x})\mathscr{K}^\circ_{\vartheta,h}(y_i, \mathbf{x}_i; y', \mathbf{x}')\right].$$

Theorem 2 (Strong Approximation)

Suppose Assumptions 1 and 2 hold. If $nh^{1+d+2\mathfrak{p}} \to 0$ and if $nh^{1+d}/\log(n) \to \infty$, then there exist two stochastic processes, $\widehat{\mathbb{S}}'_\vartheta$ and $\mathbb{G}_\vartheta$, in a possibly enlarged probability space, such that:

(i) $\widehat{\mathbb{S}}_\vartheta$ *and* $\widehat{\mathbb{S}}'_\vartheta$ *have the same distribution,*

(ii) $\mathbb{G}_\vartheta$ *is a centered Gaussian process with covariance function* ρ_ϑ*, and*

(iii)

$$\sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}} \left|\widehat{\mathbb{S}}'_\vartheta(y,\mathbf{x}) - \mathbb{G}_\vartheta(y,\mathbf{x})\right| \precsim_{\mathbb{TC}} \mathtt{r}_{\mathtt{SA}}, \qquad \mathtt{r}_{\mathtt{SA}} = \sqrt{nh^{1+d+2\mathfrak{p}}} + \left(\frac{\log^{1+d}(n)}{nh^{1+d}}\right)^{\frac{1}{2+2d}}.$$

The theorem provides a Gaussian approximation for the entire stochastic process $\widehat{\mathbb{S}}_\vartheta$ rather than for a particular functional thereof. Later we will employ this result to approximate the distribution of the suprema of the two processes, from which uniform confidence bands can be constructed.

2.4 Covariance Estimation

Because both the process $\widehat{\mathbb{S}}_\vartheta$ and the covariance function ρ_ϑ depend on unknown features of the underlying data generating process (namely, the covariance function C_ϑ), Theorem 2 in isolation cannot be used for inference. Equipped with a suitably accurate estimator of C_ϑ, on the other hand, Theorem 2 becomes immediately useful for inference. The purpose of this subsection is to propose and study an estimator of C_ϑ.

The covariance function C_ϑ can be expressed as a functional of two unknowns, namely conditional CDF of y_i given $\mathbf{x}_i$ and the marginal CDF of y_i. Replacing $F(y|\mathbf{x})$ and $F_y(y)$ with $\widehat{F}(y|\mathbf{x})$ and $\widehat{F}_y(y) = n^{-1}\sum_{i=1}^n \mathbb{1}(y_i \leqslant y)$, respectively, we obtain the following natural plug-in covariance function estimator:

$$\widehat{\mathsf{C}}_\vartheta(y,\mathbf{x},y',\mathbf{x}') = \frac{1}{n^2}\sum_{i=1}^n \widehat{\mathscr{K}^\circ_{\vartheta,h}}\Big(y_i,\mathbf{x}_i;y,\mathbf{x}\Big)\widehat{\mathscr{K}^\circ_{\vartheta,h}}\Big(y_i,\mathbf{x}_i;y',\mathbf{x}'\Big),$$

where

$$\widehat{\mathscr{K}^\circ_{\vartheta,h}}\Big(a,\mathbf{b};y,\mathbf{x}\Big) = \frac{1}{h^{1+\vartheta}}\mathbf{e}_{1+\vartheta}^{\mathrm{T}}\widehat{\mathbf{S}}_y^{-1}\left[\frac{1}{n}\sum_{j=1}^n\Big(\mathbb{1}(a\leqslant y_j) - \widehat{F}(y_j|\mathbf{b})\Big)\frac{1}{h}\mathbf{P}\Big(\frac{y_j-y}{h}\Big)\right]\frac{1}{h^d}\mathbf{Q}\left(\frac{\mathbf{b}-\mathbf{x}}{h}\right)^{\mathrm{T}}\widehat{\mathbf{S}}_{\mathbf{x}}^{-1}\mathbf{e}_{\mathbf{0}}.$$

The corresponding estimators of V_ϑ and ρ_ϑ are given by $\widehat{\mathsf{V}}_\vartheta(y,\mathbf{x}) = \widehat{\mathsf{C}}_\vartheta(y,\mathbf{x},y,\mathbf{x})$ and

$$\widehat{\rho}_\vartheta(y,\mathbf{x},y',\mathbf{x}') = \frac{\widehat{\mathsf{C}}_\vartheta(y,\mathbf{x},y',\mathbf{x}')}{\sqrt{\widehat{\mathsf{V}}_\vartheta(y,\mathbf{x})}\sqrt{\widehat{\mathsf{V}}_\vartheta(y',\mathbf{x}')}},$$

respectively. The next lemma establishes a uniform probability concentration result for $\widehat{\mathsf{C}}_\vartheta$.

Lemma 2 (Covariance Estimation)
Suppose Assumptions 1 and 2 hold. If $h \to 0$ and if $nh^{1+d}/\log(n) \to \infty$, then

$$\sup_{y,y' \in \mathcal{Y}, \mathbf{x},\mathbf{x}' \in \mathcal{X}} \left| \widehat{\mathsf{C}}_\vartheta(y,\mathbf{x},y',\mathbf{x}') - \mathsf{C}_\vartheta(y,\mathbf{x},y',\mathbf{x}') \right| \lesssim_{\mathbb{TC}} h^{\mathfrak{p}-\vartheta-\frac{1}{2}} + \sqrt{\frac{\log(n)}{nh^{1+d}}}.$$

Now it is possible to simulate a Gaussian process $\widehat{\mathbb{G}}_\vartheta$, which, conditional on the data, is mean zero and has the covariance $\widehat{\rho}_\vartheta$.

2.5 Suprema Approximation

Replacing $\mathsf{V}_\vartheta(y,\mathbf{x})$ with $\widehat{\mathsf{V}}_\vartheta(y,\mathbf{x})$ in (3), we obtain

$$\widehat{\mathbb{T}}_\vartheta(y,\mathbf{x}) = \frac{\widehat{f}^{(\vartheta)}(y|\mathbf{x}) - f^{(\vartheta)}(y|\mathbf{x})}{\sqrt{\widehat{\mathsf{V}}_\vartheta(y,\mathbf{x})}}.$$

By Theorem 2 and Lemma 2, the law of $(\widehat{\mathbb{T}}_\vartheta(y,\mathbf{x}) : y \in \mathcal{Y}, \mathbf{x} \in \mathcal{X})$ can be approximated by that of a centered Gaussian process with covariance function ρ_ϑ, where the latter is well approximated by $\widehat{\rho}_\vartheta$. As a consequence, functionals of $\widehat{\mathbb{T}}_\vartheta$ admit feasible distributional approximations. To illustrate this general phenomenon, the following theorem gives a result for the supremum of $|\widehat{\mathbb{T}}_\vartheta|$. Recall that $\widehat{\mathbb{G}}_\vartheta$ represents a process whose law, conditionally on the data, is centered Gaussian with covariance function $\widehat{\rho}_\vartheta$.

Theorem 3 (Kolmogorov-Smirnov Distance: Suprema)
Suppose Assumptions 1 and 2 hold. If $n\log(n)h^{1+d+2\mathfrak{p}} \to 0$ and if $nh^{1+d}/\log(n) \to \infty$, then

$$\sup_{u \in \mathbb{R}} \left| \mathbb{P}\Big[\sup_{y \in \mathcal{Y}, \mathbf{x} \in \mathcal{X}} \big|\widehat{\mathbb{T}}_\vartheta(y,\mathbf{x})\big| \leqslant u \Big] - \mathbb{P}\Big[\sup_{y \in \mathcal{Y}, \mathbf{x} \in \mathcal{X}} \big|\widehat{\mathbb{G}}_\vartheta(y,\mathbf{x})\big| \leqslant u \Big| \mathbf{X}, \mathbf{Y} \Big] \right| \lesssim_{\mathbb{P}} \mathtt{r}_{\mathtt{KS}}$$

$$\textit{where } \mathtt{r}_{\mathtt{KS}} = \sqrt{n\log(n)h^{1+d+2\mathfrak{p}}} + \left(\frac{\log^{2+2d}(n)}{nh^{1+d}} \right)^{\frac{1}{2+2d}} + \left(\frac{\log^5(n)}{nh^{1+d}} \right)^{\frac{1}{4}}.$$

To compare the rate of distributional approximation with existing results, we follow the literature and ignore the first (smoothing bias) term. Then, the resulting rate takes the form

$$\left(\frac{\log^{2+2d}(n)}{nh^{1+d}} \right)^{\frac{1}{2+2d}} + \left(\frac{\log^5(n)}{nh^{1+d}} \right)^{\frac{1}{4}}.$$

This rate matches what Chernozhukov *et al.* (2014b) obtained when $d = 2$ (see their Remark 3.1(ii)), but it is strictly faster when $d = 1$.

3 Applications

This section illustrates our theoretical and methodological results by means of three applications. Before turning to these applications, we discuss bandwidth selection, a necessary step for implementation. It is customary to select the bandwidth by minimizing an approximation to the IMSE of $\widehat{f}^{(\vartheta)}(y|\mathbf{x})$. Employing Lemma 1 and assuming that $\mathfrak{p} - \vartheta$ is even, we propose to select the bandwidth by minimizing (a feasible analogue of)

$$\iint\limits_{\mathcal{Y}\times\mathcal{X}} \left(h^{2\mathfrak{p}-2\vartheta} B_\vartheta(y,\mathbf{x})^2 + \frac{1}{nh^{1+2\vartheta+d}} V_\vartheta(y,\mathbf{x}) \right) \mathrm{d}y \mathrm{d}\mathbf{x},$$

where $B_\vartheta(y|\mathbf{x})$ and $V_\vartheta(y|\mathbf{x})$ are the constants in the leading bias and variance, respectively, defined as

$$\begin{aligned} B_\vartheta(y,\mathbf{x}) &= f^{(\mathfrak{p})}(y|\mathbf{x})\mathbf{e}_{1+\vartheta}^{\mathrm{T}}\mathbf{S}_y^{-1}\mathbf{c}_{y,\mathfrak{p}+1} + \sum_{|\boldsymbol{\nu}|=\mathfrak{p}-\vartheta} \frac{\partial^{\boldsymbol{\nu}}}{\partial \mathbf{x}^{\boldsymbol{\nu}}} f^{(\vartheta)}(y|\mathbf{x})\mathbf{e}_{\mathbf{0}}^{\mathrm{T}}\mathbf{S}_{\mathbf{x}}^{-1}\mathbf{c}_{\mathbf{x},\boldsymbol{\nu}}, \\ V_\vartheta(y,\mathbf{x}) &= f(y|\mathbf{x})\left(\mathbf{e}_{1+\vartheta}^{\mathrm{T}}\mathbf{S}_y^{-1}\mathbf{T}_y\mathbf{S}_y^{-1}\mathbf{e}_{1+\vartheta}\right)\left(\mathbf{e}_{\mathbf{0}}^{\mathrm{T}}\mathbf{S}_{\mathbf{x}}^{-1}\mathbf{T}_{\mathbf{x}}\mathbf{S}_{\mathbf{x}}^{-1}\mathbf{e}_{\mathbf{0}}\right), \end{aligned}$$

with

$$\begin{aligned} \mathbf{c}_{y,\mathfrak{p}+1} &= \int_{\mathcal{Y}} \frac{1}{(\mathfrak{p}+1)!}\left(\frac{u-y}{h}\right)^{\mathfrak{p}+1}\frac{1}{h}\mathbf{P}\left(\frac{u-y}{h}\right)\mathrm{d}F_y(u), \\ \mathbf{c}_{\mathbf{x},\boldsymbol{\nu}} &= \int_{\mathcal{X}} \frac{1}{\boldsymbol{\nu}!}\left(\frac{\mathbf{u}-\mathbf{x}}{h}\right)^{\boldsymbol{\nu}}\frac{1}{h^d}\mathbf{Q}\left(\frac{\mathbf{u}-\mathbf{x}}{h}\right)\mathrm{d}F_{\mathbf{x}}(\mathbf{u}), \\ \mathbf{T}_y &= \iint\limits_{\mathcal{Y}\times\mathcal{Y}} \frac{\min(u_1,u_2)-y}{h}\frac{1}{h^2}\mathbf{P}\left(\frac{u_1-y}{h}\right)\mathbf{P}\left(\frac{u_2-y}{h}\right)^{\mathrm{T}}\mathrm{d}F_y(u_1)\mathrm{d}F_y(u_2), \\ \mathbf{T}_{\mathbf{x}} &= \int_{\mathcal{X}} \frac{1}{h^d}\mathbf{Q}\left(\frac{\mathbf{u}-\mathbf{x}}{h}\right)\mathbf{Q}\left(\frac{\mathbf{u}-\mathbf{x}}{h}\right)^{\mathrm{T}}\mathrm{d}F_{\mathbf{x}}(\mathbf{u}). \end{aligned}$$

(The supplemental appendix also discusses the case where $\mathfrak{p} - \vartheta$ is odd and provides more general results.)

The bandwidth that minimizes the approximate IMSE, $h_{\mathfrak{p}}^{\star}$, is proportional to $n^{-\frac{1}{1+d+2\mathfrak{p}}}$. Although this bandwidth delivers estimates that are approximately IMSE-optimal, a non-vanishing bias will be present in their asymptotic distribution, complicating statistical inference. To address this well-known problem, our construction of confidence bands and test statistics for parametric or shape restrictions employs robust bias correction (Calonico *et al.*, 2018, 2022): first we construct an IMSE-optimal point estimator, and then we bias correct the estimator and adjust the covariance function estimator accordingly to obtain a valid and improved distributional approximation.

To make the robust bias-correction procedure precise, we augment the notation so that it reflects the local polynomial order (and possibly also the bandwidth) used. For example, the conditional density estimator using polynomial order $\mathfrak{p}$ (and $\mathfrak{q} = \mathfrak{p} - \vartheta - 1$) and employing the bandwidth h (and $b = h$) is written as $\widehat{f}^{(\vartheta)}_{\mathfrak{p}}(y|\mathbf{x}; h)$.

Application 1: Confidence Bands

Confidence bands can be constructed using the process $(\widehat{\mathbb{T}}^{\mathtt{CB}}_{\vartheta,\mathfrak{p}+1}(y, \mathbf{x}) : y \in \mathcal{Y}, \mathbf{x} \in \mathcal{X})$, where

$$\widehat{\mathbb{T}}^{\mathtt{CB}}_{\vartheta,\mathfrak{p}+1}(y, \mathbf{x}) = \frac{\widehat{f}^{(\vartheta)}_{\mathfrak{p}+1}(y|\mathbf{x}; h^{\star}_{\mathfrak{p}}) - f^{(\vartheta)}(y|\mathbf{x})}{\sqrt{\widehat{\mathbb{V}}_{\vartheta,\mathfrak{p}+1}(y, \mathbf{x}; h^{\star}_{\mathfrak{p}})}},$$

By Theorem 3, the distribution of $\widehat{\mathbb{T}}^{\mathtt{CB}}_{\vartheta,\mathfrak{p}+1}$ is well-approximated by the conditional (on the data) distribution of $\widehat{\mathbb{G}}_{\vartheta,\mathfrak{p}+1}$, the latter being a centered Gaussian process whose law, conditionally on the data, is Gaussian with covariance function $\widehat{\rho}_{\vartheta,\mathfrak{p}+1}(y, \mathbf{x}; h^{\star}_{\mathfrak{p}})$. Accordingly, let

$$\mathrm{CB}_{\vartheta,\mathfrak{p}+1}(1-\alpha) = \Big[\ \widehat{f}^{(\vartheta)}_{\mathfrak{p}+1}(y|\mathbf{x}; h^{\star}_{\mathfrak{p}}) \pm \mathtt{cv}^{\mathtt{CB}}_{\vartheta,\mathfrak{p}+1}(\alpha)\sqrt{\widehat{\mathbb{V}}_{\vartheta,\mathfrak{p}+1}(y, \mathbf{x}; h^{\star}_{\mathfrak{p}})}\ :\ y \in \mathcal{Y}, \mathbf{x} \in \mathcal{X}\Big],$$

where

$$\mathtt{cv}^{\mathtt{CB}}_{\vartheta,\mathfrak{p}+1}(\alpha) = \inf\Big\{u \in \mathbb{R}_+ : \mathbb{P}\Big[\sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}} \big|\widehat{\mathbb{G}}_{\vartheta,\mathfrak{p}+1}(y, \mathbf{x})\big| \leqslant u \ \Big|\ \mathbf{X}, \mathbf{Y}\Big] \geqslant 1-\alpha\Big\}.$$

As the notation suggests, $\mathrm{CB}_{\vartheta,\mathfrak{p}+1}(1-\alpha)$ is a $100(1-\alpha)\%$ confidence band. To be specific, we have:

Theorem 4 (Confidence Bands)
Suppose Assumptions 1 and 2 hold, $f^{(\mathfrak{p}+1)}(y|\mathbf{x})$ exists and is continuous, and $\partial^{\boldsymbol{\nu}} f^{(\vartheta)}(y|\mathbf{x})/\partial\mathbf{x}^{\boldsymbol{\nu}}$ exists and is continuous for all $|\boldsymbol{\nu}| = \mathfrak{p} + 1 - \vartheta$. Then

$$\Big|\mathbb{P}\left[f^{(\vartheta)} \in \mathrm{CB}_{\vartheta,\mathfrak{p}+1}(1-\alpha)\right] - (1-\alpha)\Big| \lesssim \log^{\frac{5}{4}}(n)\mathtt{r}_{\mathtt{CB}},$$

where $\mathtt{r}_{\mathtt{CB}} = n^{-\frac{1}{1+d+2\mathfrak{p}}} + n^{-\frac{2\mathfrak{p}-2\vartheta+1}{4(1+d+2\mathfrak{p})}} + n^{-\frac{\mathfrak{p}}{(1+d+2\mathfrak{p})(1+d)}}$.

The confidence band $\mathrm{CB}_{\vartheta,\mathfrak{p}+1}(1-\alpha)$ is easy to construct because, by discretizing the index set of the Gaussian process, the critical value $\mathtt{cv}_{\vartheta,\mathfrak{p}+1}(1-\alpha)$ can be computed by simulation from a conditionally (on the data) multivariate Gaussian distribution. We illustrate the performance of our proposed confidence bands using simulated data in Section 4.

Theorem 4 provides a formal, theoretical justification for employing strong approximation methods to construct confidence bands instead of relying on extreme value theory for approximating the distribution of the suprema of the process $\widehat{\mathbb{T}}^{\mathtt{CB}}_{\vartheta,\mathfrak{p}+1}$. More specifically, the coverage error rate $\mathtt{r}_{\mathtt{CB}}$ is polynomial in n for the former inference approach, while the latter inference approach would enjoy a logarithmic in n convergence rate (see, e.g., Hall, 1979, 1993, and references therein). The same remark applies to Theorems 5 and 6, which characterize the error in rejection probability of two different classes of hypothesis testing procedures.

Application 2: Parametric Specification Testing

Suppose the researcher postulates that the conditional density (derivative) belongs to the parametric class $\{f^{(\vartheta)}(y|\mathbf{x};\boldsymbol{\gamma}) : \boldsymbol{\gamma} \in \Gamma_\vartheta\}$, where Γ_ϑ is some parameter space. Abstracting away from the specifics of the estimation technique, we assume that the researcher also picks some estimator $\widehat{\boldsymbol{\gamma}}$ (e.g., maximum likelihood or minimum distance), which is assumed to converge in probability to some $\bar{\boldsymbol{\gamma}} \in \Gamma_\vartheta$. A natural statistic for the problem of testing

$$\begin{aligned}
&\mathsf{H}_0^{\mathtt{PS}}: \ f^{(\vartheta)}(y|\mathbf{x};\bar{\boldsymbol{\gamma}}) = f^{(\vartheta)}(y|\mathbf{x}) \qquad \text{for all } (y,\mathbf{x}) \in \mathcal{Y}\times\mathcal{X}\\
&\qquad \text{vs.}\\
&\mathsf{H}_1^{\mathtt{PS}}: \ f^{(\vartheta)}(y|\mathbf{x};\bar{\boldsymbol{\gamma}}) \neq f^{(\vartheta)}(y|\mathbf{x}) \qquad \text{for some } (y,\mathbf{x}) \in \mathcal{Y}\times\mathcal{X},
\end{aligned}$$

is

$$\sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}} \left|\widehat{\mathsf{T}}^{\mathtt{PS}}_{\vartheta,\mathfrak{p}+1}(y,\mathbf{x})\right|, \qquad \widehat{\mathsf{T}}^{\mathtt{PS}}_{\vartheta,\mathfrak{p}+1}(y,\mathbf{x}) = \frac{\widehat{f}^{(\vartheta)}_{\mathfrak{p}+1}(y|\mathbf{x};h^\star_{\mathfrak{p}}) - f^{(\vartheta)}(y|\mathbf{x};\widehat{\boldsymbol{\gamma}})}{\sqrt{\widehat{\mathsf{V}}_{\vartheta,\mathfrak{p}+1}(y,\mathbf{x};h^\star_{\mathfrak{p}})}}.$$

Assuming the estimation error of $\widehat{\boldsymbol{\gamma}}$ is asymptotically negligible, a valid $100\alpha\%$ critical value is given by $\mathtt{cv}^{\mathtt{CB}}_{\vartheta,\mathfrak{p}+1}(\alpha)$. To be specific, we have:

Theorem 5 (Parametric Specification Testing)
Suppose Assumptions 1 and 2 hold, $f^{(\mathfrak{p}+1)}(y|\mathbf{x})$ exists and is continuous, and $\partial^{\boldsymbol{\nu}} f^{(\vartheta)}(y|\mathbf{x})/\partial\mathbf{x}^{\boldsymbol{\nu}}$ exists and is continuous for all $|\boldsymbol{\nu}| = \mathfrak{p}+1-\vartheta$. If

$$n^{\frac{\mathfrak{p}-\vartheta}{1+d+2\mathfrak{p}}} \sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}} \left|f^{(\vartheta)}(y|\mathbf{x};\widehat{\boldsymbol{\gamma}}) - f^{(\vartheta)}(y|\mathbf{x};\bar{\boldsymbol{\gamma}})\right| \precsim_{\mathbb{TC}} \mathtt{r}_{\mathtt{CB}},$$

then, under $\mathsf{H}_0^{\texttt{PS}}$,

$$\left|\mathbb{P}\left[\sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}}|\widehat{\mathbb{T}}^{\texttt{PS}}_{\vartheta,\mathfrak{p}+1}(y,\mathbf{x})|>\texttt{cv}^{\texttt{CB}}_{\vartheta,\mathfrak{p}+1}(\alpha)\right]-\alpha\right|\lesssim\log^{\frac{5}{4}}(n)\mathtt{r}_{\texttt{CB}}.$$

Application 3: Testing Shape Restrictions

As a third application, suppose the researcher wants to test shape restrictions on $f^{(\vartheta)}$. Letting c_ϑ be a pre-specified function, consider the problem of testing

$$\begin{aligned}&\mathsf{H}_0^{\texttt{SR}}: f^{(\vartheta)}(y|\mathbf{x})\leqslant c_\vartheta(y|\mathbf{x}) \qquad \text{for all } (y,\mathbf{x})\in\mathcal{Y}\times\mathcal{X}\\ &\qquad\text{vs.}\\ &\mathsf{H}_1^{\texttt{SR}}: f^{(\vartheta)}(y|\mathbf{x})> c_\vartheta(y|\mathbf{x}) \qquad \text{for some } (y,\mathbf{x})\in\mathcal{Y}\times\mathcal{X}.\end{aligned}$$

For example, if $\vartheta=0$ and if $c_\vartheta(y|\mathbf{x})$ is some (positive) constant value c, the testing problem refers to whether the conditional density exceeds c somewhere on its support. As another example, if $\vartheta=1$ and if $c_\vartheta(y|\mathbf{x})=0$, then the testing problem refers to whether the conditional density is non-increasing in y for all values of $\mathbf{x}$. More generally, the testing problem above can be used to test for monotonicity, convexity, and other shape features of the conditional density, possibly relative to the function $c_\vartheta(y|\mathbf{x})$.

A natural testing procedure rejects $\mathsf{H}_0^{\texttt{SR}}$ whenever the test statistic

$$\sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}}\mathbb{T}^{\texttt{SR}}_{\vartheta,\mathfrak{p}+1}(y,\mathbf{x}),\qquad \mathbb{T}^{\texttt{SR}}_{\vartheta,\mathfrak{p}+1}(y,\mathbf{x})=\frac{\widehat{f}^{(\vartheta)}_{\mathfrak{p}+1}(y|\mathbf{x};h^\star_{\mathfrak{p}})-c_\vartheta(y|\mathbf{x})}{\sqrt{\widehat{\mathbb{V}}_{\vartheta,\mathfrak{p}+1}(y,\mathbf{x};h^\star_{\mathfrak{p}})}}$$

exceeds a critical value of the form

$$\texttt{cv}^{\texttt{SR}}_{\vartheta,\mathfrak{p}+1}(\alpha)=\inf\left\{u\in\mathbb{R}_+:\mathbb{P}\Big[\sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}}\widehat{\mathbb{G}}_{\vartheta,\mathfrak{p}+1}(y,\mathbf{x})\leqslant u\;\Big|\;\mathbf{X},\mathbf{Y}\Big]\geqslant 1-\alpha\right\}.$$

Theorem 6 (Testing Shape Restriction)
Suppose Assumptions 1 and 2 hold, $f^{(\mathfrak{p}+1)}(y|\mathbf{x})$ *exists and is continuous, and* $\partial^{\boldsymbol{\nu}}f^{(\vartheta)}(y|\mathbf{x})/\partial\mathbf{x}^{\boldsymbol{\nu}}$ *exists and is continuous for all* $|\boldsymbol{\nu}|=\mathfrak{p}+1-\vartheta$. *Then, under* $\mathsf{H}_0^{\texttt{SR}}$,

$$\left|\mathbb{P}\left[\sup_{y\in\mathcal{Y},\mathbf{x}\in\mathcal{X}}\widehat{\mathbb{T}}^{\texttt{SR}}_{\vartheta,\mathfrak{p}+1}(y,\mathbf{x})>\texttt{cv}^{\texttt{SR}}_{\vartheta,\mathfrak{p}+1}(\alpha)\right]-\alpha\right|\lesssim\log^{\frac{5}{4}}(n)\mathtt{r}_{\texttt{CB}}.$$

4 Simulations

We illustrate the effectiveness of our proposed methods with a Monte Carlo experiment. Replication files, additional simulation results, and details of the companion `R` package, `lpcde`, can be found at `https://nppackages.github.io/lpcde/` and in our companion software article (Cattaneo *et al.*, 2022).

For the sake of simplicity, we set $d = 1$ and assume that $\mathbf{x}$ and y are simulated by a joint normal distribution with variance 2 and covariance -0.1, truncated on $[-1, 1]^2$. We simulate 1,000 data sets of sample size $n = 5,000$. Table 1 presents the simulation results for the conditional PDF at three different conditioning values: (a) interior ($\mathbf{x} = 0$), (b) near-boundary ($\mathbf{x} = 0.8$), and (c) at-boundary ($\mathbf{x} = 1$). Point estimates are generated on 20 equally spaced points for y on $[0, 1]$. We report average bandwidth in column "$\widehat{h}$". We consider bands formed by pointwise confidence intervals (columns "pointwise"), which are not uniformly valid and hence should exhibit considerable under coverage, as well as the uniform confidence bands discussed in Section 3 (columns "uniform"). We report their empirical uniform coverage probabilities (column "Coverage") and the average width (column "Width"). Without bias correction (rows "NBC"), the polynomial orders for bandwidth selection, point estimation and statistical inference are $\mathfrak{p} = 2$ and $\mathfrak{q} = 1$, while those for robust bias-corrected statistical inference (rows "RBC") are $\mathfrak{p} = 3$ and $\mathfrak{q} = 2$.

			Coverage		Width	
		$\widehat{h}$	pointwise	uniform	pointwise	uniform
$\mathbf{x} = 0$	NBC	0.32	62.6	74.8	0.01	0.02
	RBC	0.32	83.4	93.9	0.05	0.05
$\mathbf{x} = 0.8$	NBC	0.30	72.8	89.4	0.02	0.03
	RBC	0.30	86.9	94.3	0.13	0.19
$\mathbf{x} = 1.0$	NBC	0.32	74.9	91.3	0.02	0.05
	RBC	0.32	88.1	93.2	0.11	0.23

Table 1. Empirical uniform coverage probabilities.

The simulation results in Table 1 support our main theoretical findings. First, robust bias-correction leads to better performance of the inference procedures, both pointwise and uniformly over $\mathcal{Y}$. Second, our uniform distributional approximation leads to feasible confidence bands with good finite sample performance, when coupled with robust bias correction methods.

For example, for $\mathbf{x} = 0$, the averaged (across simulations) estimated approximate IMSE-optimal bandwidth choice is $\widehat{h} = 0.32$, with $\mathfrak{p} = 2$ and $\mathfrak{q} = \mathfrak{p} - 1$. Bands constructed with pointwise confidence intervals have empirical uniform coverage of 62.6% without bias correction, and 83.4% with robust bias correction, both are substantially below the 95% nominal level because they are not uniformly valid over the range of y. The feasible confidence bands are designed to address that issue: our proposed confidence bands have empirical coverage of 93.9% when robust bias correction is employed. It also highlights the importance of addressing the misspecification (smoothing) bias for statistical inference: without bias correction, the confidence bands only covers the true conditional PDF with a probability 74.8%.

5 Conclusion

We introduced a new boundary adaptive estimator of the conditional density and derivatives thereof. This estimator is conceptually distinct from prior proposals in the literature, as it relies on two (nested) local polynomial estimators. Our proposed estimation approach has several appealing features, most notably automatic boundary carpentry. We provided an array of uniform estimation and distributional results, including a valid uniform equivalent kernel representation and uniform distributional approximations. Our methods are applicable in data science settings either where the conditional density or its derivatives are the main object of interest, or where they are preliminary estimands entering a multi-step statistical procedure. The supplemental appendix contains several other technical and methodological results not included here to streamline the presentation.

References

Calonico, S., Cattaneo, M. D., and Farrell, M. H. (2018). "On the Effect of Bias Estimation on Coverage Accuracy in Nonparametric Inference," *Journal of the American Statistical Association*, *113*(522), 767–779.

Calonico, S., Cattaneo, M. D., and Farrell, M. H. (2022). "Coverage Error Optimal Confidence Intervals for Local Polynomial Regression," *Bernoulli*, *28*(4), 2998–3022.

Cattaneo, M. D., Chandak, R., Jansson, M., and Ma, X. (2022). "`lpcde`: Local Polynomial Conditional Density Estimation and Inference," *working paper*.

Cattaneo, M. D., Jansson, M., and Ma, X. (2020). "Simple Local Polynomial Density Estimators," *Journal of the American Statistical Association*, *115*(531), 1449–1455.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2014a). "Anti-Concentration and Honest, Adaptive Confidence Bands," *Annals of Statistics*, *42*(5), 1787–1818.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2014b). "Gaussian Approximation of Suprema of Empirical Processes," *Annals of Statistics*, *42*(4), 1564–1597.

De Gooijer, J. G. and Zerom, D. (2003). "On Conditional Density Estimation," *Statistica Neerlandica*, *57*(2), 159–176.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*, New York: Chapman & Hall/CRC.

Fan, J., Yao, Q., and Tong, H. (1996). "Estimation of Conditional Densities and Sensitivity Measures in Nonlinear Dynamical Systems," *Biometrika*, *83*(1), 189–206.

Giné, E. and Nickl, R. (2016). *Mathematical Foundations of Infinite-dimensional Statistical Models*, New York: Cambridge University Press.

Hall, P. (1979). "On the rate of convergence of normal extremes," *Journal of Applied Probability*, *16*(2), 433–439.

Hall, P. (1993). "On Edgeworth expansion and bootstrap confidence bands in nonparametric curve estimation," *Journal of the Royal Statistical Society: Series B (Methodological)*, *55*(1), 291–304.

Hall, P., Racine, J., and Li, Q. (2004). "Cross-Validation and the Estimation of Conditional Probability Densities," *Journal of the American Statistical Association*, *99*(468), 1015–1026.

Hall, P., Wolff, R. C., and Yao, Q. (1999). "Methods for Estimating a Conditional Distribution Function," *Journal of the American Statistical association*, *94*(445), 154–163.

Khas'minskii, R. Z. (1979). "A Lower Bound on the Risks of Non-parametric Estimates of Densities in the Uniform Metric," *Theory of Probability & Its Applications*, *23*(4), 794–798.

Komlós, J., Major, P., and Tusnády, G. (1975). "An Approximation of Partial Sums of Independent RV'-s, and the sample DF. I," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, *32*(1), 111–131.

Rio, E. (1994). "Local Invariance Principles and Their Application to Density Estimation," *Probability Theory and Related Fields*, *98*(1), 21–45.

Scott, D. W. (2015). *Multivariate Density Estimation: Theory, Practice, and Visualization*: John Wiley & Sons.

Simonoff, J. S. (2012). *Smoothing Methods in Statistics*: Springer.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*: Springer.

Wand, M. and Jones, M. (1995). *Kernel Smoothing*: Chapman & Hall/CRC.

Wasserman, L. (2006). *All of Nonparametric Statistics*: Springer.