Journal of Econometrics 240 (2024) 105074

Contents lists available at ScienceDirect

Journal of Econometrics

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

Local regression distribution estimators☆

Matias D. Cattaneo [a,*], Michael Jansson [b,c], Xinwei Ma [d]

[a] Department of Operations Research and Financial Engineering, Princeton University, United States of America
[b] Department of Economics, UC Berkeley, United States of America
[c] CREATES, Denmark
[d] Department of Economics, UC San Diego, United States of America

ARTICLE INFO

Article history:
Received 5 December 2019
Received in revised form 30 September 2020
Accepted 14 January 2021
Available online 6 March 2021

Keywords:
Distribution and density estimation
Local polynomial methods
Uniform approximation
Efficiency
Optimal kernel
Program evaluation

ABSTRACT

This paper investigates the large sample properties of local regression distribution estimators, which include a class of boundary adaptive density estimators as a prime example. First, we establish a pointwise Gaussian large sample distributional approximation in a unified way, allowing for both boundary and interior evaluation points simultaneously. Using this result, we study the asymptotic efficiency of the estimators, and show that a carefully crafted minimum distance implementation based on "redundant" regressors can lead to efficiency gains. Second, we establish uniform linearizations and strong approximations for the estimators, and employ these results to construct valid confidence bands. Third, we develop extensions to weighted distributions with estimated weights and to local L^2 estimation. Finally, we illustrate our methods with two applications in program evaluation: counterfactual density testing, and IV specification and heterogeneity density analysis. Companion software packages in `Stata` and `R` are available.

1. Introduction

Kernel-based nonparametric estimation of distribution and density functions, as well as higher-order derivatives thereof, play an important role in econometrics. These nonparametric estimators often feature both as the main object of interest and as preliminary ingredients in multi-step semiparametric procedures (Newey and McFadden, 1994; Ichimura and Todd, 2007). Whitney Newey's path-breaking contributions to non/semiparametric econometrics employing kernel smoothing are numerous.[1] This paper hopes to honor his influential work in this area by studying the main large sample properties of a new class of *local regression distribution estimators*, which can be used for non/semiparametric estimation and inference.

The class of local regression distribution estimators is constructed using a local least squares approximation to the empirical distribution function of a random variable $x \in \mathcal{X} \subseteq \mathbb{R}$, where the localization at the evaluation point $\mathrm{x} \in \mathcal{X}$ is

☆ Prepared for "Celebrating Whitney Newey's Contributions to Econometrics" Conference at MIT, May 17–18, 2019. We thank the conference participants for comments, and Guido Imbens and Yingjie Feng for very useful discussions. We are also thankful to the handling co-Editor, Xiaohong Chen, an Associate Editor and two reviewers for their input. Cattaneo gratefully acknowledges financial support from the National Science Foundation, United States of America through grant SES-1947805, and Jansson gratefully acknowledges financial support from the National Science Foundation, United States of America through grant SES-1947662 and the research support of CREATES, Denmark.

* Corresponding author.
E-mail address: cattaneo@princeton.edu (M.D. Cattaneo).

[1] See, for example, Newey and Stoker (1993), Newey (1994a,b), Hausman and Newey (1995), Robins et al. (1995), Newey et al. (2004), Newey and Ruud (2005), Ichimura and Newey (2020), and Chernozhukov et al. (2020).

https://doi.org/10.1016/j.jeconom.2021.01.006
0304-4076/

implemented via a kernel function and a bandwidth parameter. The local functional form approximation is done using a finite-dimension basis function. When the basis function contains polynomials up to order $p \in \mathbb{N}$, the associated least squares coefficients give estimators of the distribution function, density function, and higher-order derivatives (up to order $p-1$), all evaluated at $x \in \mathcal{X}$. If only a polynomial basis is used, then the estimator reduces to the one recently proposed in Cattaneo et al. (2020b).

We present two main large sample distributional results for the local regression distribution estimators. First, in Section 3, we establish a pointwise (in $x \in \mathcal{X}$) Gaussian distributional approximation with consistent standard errors. Because these estimators have a U-statistic structure with an n-varying kernel, where n denotes the sample size, we construct a fully automatic Studentization given a choice of basis, kernel, and bandwidth. Furthermore, we show that when the basis function includes polynomials, the associated density and its higher-order derivatives estimators are boundary adaptive without further modifications. This result generalizes (Cattaneo et al., 2020b) by allowing for arbitrary local basis functions, which is particularly useful for efficiency considerations.

To be more precise, for the special case of local polynomial density estimation, Cattaneo et al. (2020b) showed that the asymptotic variance of the estimator is of the "sandwich" form, which does not reduce to a single matrix (up to a proportional factor) by a choice of kernel function. This finding indicates that more efficient estimators can be constructed via a minimum distance approach based on "redundant" regressors, following well-known results in econometrics (Newey and McFadden, 1994). In Section 3.3, we present a novel minimum distance construction for estimation of the density and its derivatives, and obtain an efficiency bound for the new minimum distance density estimator. Furthermore, we show that the efficiency bound coincides with the well-known asymptotic variance lower bound for kernel-based density estimation (Granovsky and Müller, 1991; Cheng, Fan, and Marron, 1997). We also show that this efficiency bound is tight: we construct a feasible minimum distance procedure exploiting carefully chosen redundant regressors, which leads to an estimator with asymptotic variance arbitrarily close to the theoretical efficiency bound. These results offer not only a novel theoretical perspective on efficiency of classical nonparametric kernel-based density estimation, but also a new class of more efficient boundary adaptive density estimators for practice. We also discuss how these results generalize to other local regression distribution estimators in the supplemental appendix.

Our second main large sample distributional result, in Section 4, concerns uniform estimation and inference over a region $\mathcal{I} \subseteq \mathcal{X}$, based on either the basic local regression distribution estimators or the associated more efficient estimators obtained via our proposed minimum distance procedure. More precisely, we establish a strong approximation to the boundary adaptive Studentized statistic, uniformly over $x \in \mathcal{I}$, relying on a "coupling" result in Giné et al. (2004); see also Rio (1994) and Giné and Nickl (2010) for closely related results, and Zaitsev (2013) for a review on strong approximation methods. This approach allows us to deduce a distributional approximation for many functionals of the Studentized statistic, including its supremum, following ideas in Chernozhukov et al. (2014b). For further discussion and references on strong approximations and their applications to non/semiparametric econometrics see Chernozhukov et al. (2014a), Belloni et al. (2015, 2019), and Cattaneo et al. (2020a, 2021a), and references therein.

We employ our strong approximation results for local regression distribution estimators to construct asymptotically valid confidence bands for the density function and derivatives thereof. Other applications of our results, not discussed here to conserve space, include specification and shape restriction testing. As a by-product, we also establish a linear approximation to the boundary adaptive Studentized statistic, uniformly over $x \in \mathcal{I}$, which gives uniform convergence rates and can be used for further theoretical developments. See the supplemental appendix for more details.

In addition to our main large sample results for local regression distribution and related estimators, we briefly discuss several extensions in Section 5. First, we allow for a weighted empirical distribution function entering our estimators, where the weights themselves may be estimated. Our results continue to hold in this more general case, which is practically relevant as illustrated in our empirical applications. Second, we present and study an alternative class of estimators that employ a non-random L^2 smoothing, instead of the more standard least squares approximation underlying our local regression distribution estimators. These alternative estimators enjoy certain theoretical advantages, but require ex-ante knowledge of the boundary location of $\mathcal{X}$. In particular, we show in the supplemental appendix how these alternative estimators can be implemented to achieve maximum asymptotic efficiency in estimating the density function and its derivatives. Third, we also discuss incorporating shape restrictions using the general local basis function entering the local regression distribution estimators.

Finally, in Section 6, we illustrate our methods with two applications in program evaluation (for a review see Abadie and Cattaneo, 2018). First, we discuss counterfactual density analysis following DiNardo et al. (1996); see also Chernozhukov et al. (2013) for related discussion based on distribution functions. Second, we discuss specification testing and heterogeneity analysis in the context of instrumental variables following Kitagawa (2015) and Abadie (2003), respectively; see also Imbens and Rubin (2015) for background and other applications of nonparametric density estimation to causal inference and program evaluation. In all these applications, we develop formal estimation and inference methods based on nonparametric density estimation using local regression distribution estimators implemented with weighted distribution functions. We showcase our new methods using a subsample of the data in Abadie et al. (2002), corresponding to the Job Training Partnership Act (JTPA).

From both methodological and technical perspectives, our proposed class of local regression distribution estimators is different from, and exhibits demonstrable advantages over, other related estimators available in the literature. For the special case of density estimation (i.e., when the basis function is taken to be polynomial), our resulting kernel-based density estimator enjoys boundary carpentry over the possibly unknown boundary of $\mathcal{X}$, does not require

preliminary smoothing of the data and hence avoids preliminary tuning parameter choices, and is easy to implement and interpret. Cattaneo et al. (2020b) gave a detailed discussion of that density estimator and related approaches in the literature, which include the influential local polynomial estimator of Cheng et al. (1997) and related estimators (Zhang and Karunamuni, 1998; Karunamuni and Zhang, 2008, and references therein). The class of estimators we consider here can be more efficient by employing minimum distance estimation ideas (Section 3), easily delivers intuitive estimators of density-weighted averages (Section 5.1), and allows for incorporating shape and other restrictions (Section 5.3), among other features that we discuss below. Last but not least, some of the technical results presented herein for the general class of estimators, such as asymptotic efficiency (Section 3.3) and uniform inference (Section 4) are new, even for the special case of density estimation in Cattaneo et al. (2020b).

The rest of the paper proceeds as follows. Section 2 introduces the class of local regression distribution estimators. Section 3 establishes a pointwise distributional approximation, along with a consistent standard error estimator, and discusses efficiency focusing in particular on the leading special case of density estimation. Section 4 establishes uniform results, including valid linearizations and strong approximations, which are then used to construct confidence bands. Section 5 discusses extensions of our methodology, while Section 6 illustrates our new methods with two distinct program evaluation applications. Section 7 concludes the paper. The supplemental appendix (SA) includes all proofs of our theoretical results as well as other technical, methodological and numerical results that may be of independent interest. Software packages for `Stata` and `R` implementing the main results in this paper are discussed in Cattaneo et al. (2021b).

2. Setup

Suppose $x_1, x_2, \ldots, x_n$ is a random sample from a univariate random variable x with absolute continuous cumulative distribution function $F(\cdot)$, and associated Lebesgue density $f(\cdot)$, over its support $\mathcal{X} \subseteq \mathbb{R}$, which may be compact and not necessarily known. We propose, and study the large sample properties of a new class of nonparametric estimators of $F(\cdot)$, $f(\cdot)$, and derivatives thereof, both pointwise at $\mathsf{x} \in \mathcal{X}$ and uniformly over some region $\mathcal{I} \subseteq \mathcal{X}$.

Our proposed estimators are applicable whenever $F(\cdot)$ is suitably smooth near x and admits a sufficiently accurate linear-in-parameters local approximation of the form:

$$\varrho(h, \mathsf{x}) = \sup_{|x-\mathsf{x}|\leq h} \left| F(x) - R(x-\mathsf{x})'\theta(\mathsf{x}) \right| \qquad \text{is small for } h \text{ small}, \tag{1}$$

where $R(\cdot)$ is a known local basis function and $\theta(\mathsf{x})$ is a parameter vector to be estimated. As an estimator of $\theta(\mathsf{x})$ in (1), we consider the local regression estimator

$$\hat{\theta}(\mathsf{x}) = \operatorname*{argmin}_{\theta} \sum_{i=1}^{n} W_i \left(\hat{F}_i - R_i'\theta \right)^2, \tag{2}$$

where $W_i = K((x_i - \mathsf{x})/h)/h$ for some kernel $K(\cdot)$ and some bandwidth h, $R_i = R(x_i - \mathsf{x})$, and

$$\hat{F}_i = \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}(x_j \leq x_i) \tag{3}$$

is the empirical distribution function evaluated at x_i, with $\mathbb{1}(\cdot)$ denoting the indicator function.

The generic formulation (1) is motivated in part by the important special case where $F(\cdot)$ is sufficiently smooth, in which case

$$F(x) \approx F(\mathsf{x}) + f(\mathsf{x})(x-\mathsf{x}) + \cdots + f^{(p-1)}(\mathsf{x})\frac{1}{p!}(x-\mathsf{x})^p \qquad \text{for } x \approx \mathsf{x}, \tag{4}$$

and $f^{(s)}(\mathsf{x}) = \mathrm{d}^s f(x)/\mathrm{d}x^s|_{x=\mathsf{x}}$ are higher-order density derivatives. Of course, the approximation (4) is of the form (1) with $R(u) = (1, u, \ldots, u^p/p!)'$, and hence $\theta(\mathsf{x}) = (F(\mathsf{x}), f(\mathsf{x}), \ldots, f^{(p-1)}(\mathsf{x}))'$. In such special case, the estimator $\hat{\theta}(\mathsf{x})$ corresponds to one of the estimators introduced in Cattaneo et al. (2020b). But, as further discussed below, other choices of $R(\cdot)$ and/or $\theta(\cdot)$ can be attractive, and as a consequence we take (1) as the starting point for our analysis. Section 5 discusses other extensions and generalization of the basic local regression distribution estimator $\hat{\theta}(\mathsf{x})$ in (2).

The class of estimators defined in (2) is motivated by standard local polynomial regression methods (Fan and Gijbels, 1996). However, well-known results for local polynomial regression are not applicable to the local regression distribution estimator, $\hat{\theta}(\mathsf{x})$, because the empirical distribution function estimator, $\hat{F}_i$, which plays the role of the "dependent" variable in the construction, depends not only on x_i but also on all of the "independent" observations $x_1, x_2, \ldots, x_n$. This implies that, unlike the case of standard local polynomial regression, $\hat{\theta}(\mathsf{x})$ cannot be studied by conditioning on the "covariates" $x_1, x_2, \ldots, x_n$. Instead, we employ U-statistic methods for analyzing the statistical properties of $\hat{\theta}(\mathsf{x})$. This observation explains the quite different asymptotic variance of our estimator: see Section 3.3 for details. Furthermore, as discussed in Section 5.1, when a weighted distribution function is used in place of $\hat{F}_i$ in (2), the resulting (weighted) local regression distribution estimators are consistent for a density-weighted regression function, as opposed to being consistent for the regression function itself (as it is the case for standard local polynomial regression methods). Finally, the SA highlights other technical differences between the two types of local regression estimators.

3. Pointwise distribution theory

This section discusses the large sample properties of the estimator $\hat{\theta}(\mathsf{x})$, pointwise in $\mathsf{x} \in \mathcal{X}$. We first establish asymptotic normality, and then discuss asymptotic efficiency. Other results are reported in the SA to conserve space. We drop the dependence on the evaluation point x whenever possible.

3.1. Assumptions

We impose the following assumption throughout this section. We do not restrict the support of $\mathcal{X}$, which can be a compact set or unbounded, because our estimator automatically adapts to boundary evaluation points.

Assumption 1. $x_1, \ldots, x_n$ is a random sample from a distribution $F(\cdot)$ supported on $\mathcal{X} \subseteq \mathbb{R}$, and $\mathsf{x} \in \mathcal{X}$.

(i) For some $\delta > 0$, $F(\cdot)$ is absolutely continuous on $[\mathsf{x}-\delta, \mathsf{x}+\delta]$ with a density $f(\cdot)$ admitting constants $f(\mathsf{x}-)$, $\dot{f}(\mathsf{x}-)$, $f(\mathsf{x}+)$, and $\dot{f}(\mathsf{x}+)$ such that

$$\sup_{u\in[-\delta,0)} \frac{|f(\mathsf{x}+u)-f(\mathsf{x}-)-\dot{f}(\mathsf{x}-)u|}{|u|^2} + \sup_{u\in(0,\delta]} \frac{|f(\mathsf{x}+u)-f(\mathsf{x}+)-\dot{f}(\mathsf{x}+)u|}{|u|^2} < \infty.$$

(ii) $K(\cdot)$ is nonnegative, symmetric, and continuous on its support $[-1, 1]$, and integrates to 1.

(iii) $R(\cdot)$ is locally bounded, and there exists a positive-definite diagonal matrix Υ_h for each $h > 0$, such that $\Upsilon_h R(u) = R(u/h)$.

(iv) Let $\mathcal{X}_{h,\mathsf{x}} = \frac{\mathcal{X}-\mathsf{x}}{h}$. For all h sufficiently small, the minimum eigenvalues of $\Gamma_{h,\mathsf{x}}$ and $h^{-1}\Sigma_{h,\mathsf{x}}$ are bounded away from zero, where

$$\Gamma_{h,\mathsf{x}} = \int_{\mathcal{X}_{h,\mathsf{x}}} R(u)R(u)'K(u)f(\mathsf{x}+hu)\mathrm{d}u,$$

$$\Sigma_{h,\mathsf{x}} = \int_{\mathcal{X}_{h,\mathsf{x}}}\int_{\mathcal{X}_{h,\mathsf{x}}} R(u)R(v)'\big[F(\mathsf{x}+h\min\{u,v\}) - F(\mathsf{x}+hu)F(\mathsf{x}+hv)\big]K(u)K(v)f(\mathsf{x}+hu)f(\mathsf{x}+hv)\mathrm{d}u\mathrm{d}v.$$

Part (i) imposes smoothness conditions on the distribution function $F(\cdot)$, separately for the two regions on the left and on the right of the evaluation point x. In most applications, the distribution function will also be smooth at the evaluation point, in which case $f(\mathsf{x}-) = f(\mathsf{x}+)$ and $\dot{f}(\mathsf{x}-) = \dot{f}(\mathsf{x}+)$. However, there are important situations where $F(\cdot)$ only has one-sided derivatives, such as at boundary or kink evaluation points. Part (ii) imposes standard restrictions on the kernel function, which allows for all commonly used (compactly supported) second-order kernel functions. Part (iii) requires that the local basis $R(\cdot)$ can be stabilized by a suitable normalization. Parts (iv) give assumptions on two (non-random) matrices which will feature in the asymptotic distribution.

The error of the approximation in (1) depends on the choice of $R(\cdot)$ and θ, and is quantified by $\varrho(h)$, where we suppress the dependence on the evaluation point x to save notation. The approximation error will be required to be "small" in the sense that $n\varrho(h)^2/h \to 0$. In the cases of main interest (i.e., when $R(\cdot)$ is polynomial), we have either $\varrho(h) = O(h^{p+1})$ or $\varrho(h) = o(h^p)$ for some p. The condition can therefore be stated as $nh^{2p+1} \to 0$ and $nh^{2p-1} = O(1)$, respectively, in those cases.

We do not discuss how to choose the bandwidth h, or the order p if $R(\cdot)$ contains polynomials, as both choices can be developed following standard ideas in the local polynomial literature. We focus instead on distributional approximation (Section 3.2) and asymptotic variance minimization (Section 3.3), given a choice of bandwidth sequence and polynomial order. Bandwidth selection can be developed by extending the results in Cattaneo et al. (2020b) and polynomial order selection can be developed following Fan and Gijbels (1996, Section 3.3). In particular, a larger p can lead to more bias reduction whenever the target population function is smooth enough at the expense of a larger asymptotic variance. We discuss this trade-off explicitly in our efficiency calculations (Section 3.3).

3.2. Asymptotic normality

We show that, under regularity conditions and if h vanishes at a suitable rate as $n \to \infty$, then

$$\hat{\Omega}^{-1/2}(\hat{\theta}-\theta) \rightsquigarrow \mathcal{N}(0, I), \qquad \hat{\Omega} = \hat{\Gamma}^{-1}\hat{\Sigma}\hat{\Gamma}^{-1}, \tag{5}$$

where

$$\hat{\Gamma} = \frac{1}{n}\sum_{i=1}^{n} W_i R_i R_i', \qquad \hat{\Sigma} = \frac{1}{n^2}\sum_{i=1}^{n} \hat{\psi}_i\hat{\psi}_i', \qquad \hat{\psi}_i = \frac{1}{n}\sum_{j=1}^{n} W_j R_j(\mathbb{1}(x_i \le x_j) - \hat{F}_j).$$

It follows from this result that inference on θ can be based on $\hat{\theta}$ by employing the (pointwise) distributional approximation $\hat{\theta} \overset{a}{\sim} \mathcal{N}(\theta, \hat{\Omega})$. The three matrices, $\hat{\Gamma}$, $\hat{\Sigma}$ and $\hat{\Omega}$, depend on the evaluation point x, but such dependence is again suppressed for simplicity. This distributional result will rely on the "small" bias condition $n\varrho(h)^2/h \to 0$ mentioned above, which

makes the asymptotic approximation (or smoothing) bias of $\hat{\theta}$ negligible relative to the standard error. From an inference perspective, such bias condition can be achieved by employing undersmoothing or robust bias correction: see Calonico et al. (2018, 2020) for discussion and background references. The SA includes more details on the bias of the estimator.

To provide some insight into the distributional approximation (5), and to see why it cannot be established using standard results for local polynomial regression, first observe that

$$\hat{\theta} - \theta = \hat{\Gamma}^{-1} S, \qquad S = \frac{1}{n} \sum_{i=1}^{n} W_i R_i (\hat{F}_i - R_i' \theta),$$

assuming $\hat{\Gamma}$ is invertible with probability approaching one. The statistic S can be written as

$$S = U + B, \qquad U = \frac{1}{n(n-1)} \sum_{i,j=1, i \neq j}^{n} W_j R_j \Big(\mathbb{1}(x_i \leq x_j) - F(x_j) \Big), \tag{6}$$

where B consists of a leave-in bias term and a smoothing bias term. Since S is approximately a second-order U-statistic, result (5) should follow from a central limit theorem for (n-varying) U-statistics under suitable regularity conditions, including conditions ensuring that the approximation errors are negligible. More specifically, result (5) follows if U is asymptotically mean-zero Gaussian $\mathbb{V}[U]^{-1/2} U \rightsquigarrow \mathcal{N}(0, I)$, where $\mathbb{V}[U]$ denotes the variance of U, $\mathbb{V}[U]^{-1/2} B \to_{\mathbb{P}} 0$, and if the variance estimator $\hat{\Sigma}$ is consistent in the sense that $\mathbb{V}[U]^{-1}(\hat{\Sigma} - \mathbb{V}[U]) \to_{\mathbb{P}} 0$. Moreover, the projection theorem for U-statistics implies that, under appropriate regularity conditions,

$$\mathbb{V}[U] \approx \frac{1}{n} \mathbb{E}[\psi_i \psi_i'], \qquad \psi_i = \mathbb{E}[W_j R_j \mathbb{1}(x_i \leq x_j) - F(x_j) | x_i],$$

which motivates the functional form of the variance estimator $\hat{\Sigma}$ used to form $\hat{\Omega}$.

The following theorem formalizes the above intuition with precise sufficient conditions.

Theorem 1 (*Pointwise Asymptotic Normality*)**.** *Suppose Assumption 1 holds. If* $n\varrho(h)^2/h \to 0$ *and* $nh^2 \to \infty$, *then (5) holds.*

This theorem establishes a (pointwise) Gaussian distributional approximation for the Studentized statistic $\hat{\Omega}^{-1/2}(\hat{\theta} - \theta)$, which is valid for each evaluation point $\mathsf{x} \in \mathcal{X}$. For example, letting c be a vector of conformable dimension and $\alpha \in (0, 1)$, this result justifies the standard $100(1-\alpha)\%$ confidence interval

$$\mathrm{CI}_\alpha(\mathsf{x}) = \left[c'\hat{\theta}(\mathsf{x}) - \mathsf{q}_{1-\alpha/2} \sqrt{c'\hat{\Omega}(\mathsf{x})c} \ , \ c'\hat{\theta}(\mathsf{x}) - \mathsf{q}_{\alpha/2} \sqrt{c'\hat{\Omega}(\mathsf{x})c} \right],$$

where $\mathsf{q}_a = \inf\{u \in \mathbb{R} : \mathbb{P}[\mathcal{N}(0,1) \leq u] \geq a\}$. The above confidence interval is asymptotically valid for each evaluation point x, which is reflected by the notation $\mathrm{CI}_\alpha(\mathsf{x})$. That is,

$$\lim_{n \to \infty} \mathbb{P}\Big[c'\theta(\mathsf{x}) \in \mathrm{CI}_\alpha(\mathsf{x}) \Big] = 1 - \alpha, \qquad \text{for all } \mathsf{x} \in \mathcal{X}.$$

Section 4 develops asymptotically valid confidence bands, which will be denoted by $\mathrm{CI}_\alpha(\mathcal{I})$ for some region $\mathcal{I} \subseteq \mathcal{X}$.

3.3. Efficiency

As it is well known in the literature (Fan and Gijbels, 1996), the standard local polynomial regression estimator of $\mathbb{E}[y|x = \mathsf{x}]$, for dependent variable y and independent variable x, has a limiting asymptotic variance of the "sandwich form" $e_0' \Gamma^{-1} A \Gamma^{-1} e_0$, where e_ℓ denotes the $(\ell+1)$th standard basis vector, and

$$\Gamma = f(\mathsf{x}) \int_{-1}^{1} R(u) R(u)' K(u) \mathrm{d}u, \quad A = \mathbb{V}[y|x = \mathsf{x}] f(\mathsf{x}) \int_{-1}^{1} R(u) R(u)' K(u)^2 \mathrm{d}u.$$

This variance structure implies that setting $K(\cdot)$ to be the uniform kernel makes Γ proportional to A (i.e., $K(u) = K(u)^2$ whenever $K(u) = \mathbb{1}(|u| \leq 1)$), and hence minimizes the above asymptotic variance, at least in the sense that $\Gamma^{-1} A \Gamma^{-1} \geq A^{-1}$. See also Granovsky and Müller (1991) for a more general discussion on the optimality of the uniform kernel for kernel-based estimation.

Unlike the case of the asymptotic variance of local polynomial regression, however, our local regression distribution estimators exhibit a more complex and uneven asymptotic variance formula due to their construction. As a result, employing the uniform kernel may not exhaust the potential efficiency gains. For example, in the case of local polynomial density estimation (Cattaneo, Jansson, and Ma, 2020b), $R(u)$ is polynomial of order $p \geq 1$ and the asymptotic variance of the density estimator $\hat{f}(\mathsf{x}) = e_1' \hat{\theta}(\mathsf{x})$ takes the form $e_1' \Gamma^{-1} \Sigma \Gamma^{-1} e_1$ with

$$\Sigma = f(\mathsf{x})^3 \int_{-1}^{1} \int_{-1}^{1} \min\{u, v\} R(u) R(v)' K(u) K(v) \mathrm{d}u \mathrm{d}v,$$

which implies that Γ is no longer proportional to Σ even when the kernel function is uniform. (To show this result, one first recognizes that the asymptotic variance of $\hat{f}(\mathsf{x})$ is $h^{-1}e_1'\Gamma_h^{-1}\Sigma_h\Gamma_h^{-1}e_1$, where the matrices are defined in Assumption 1. Then the expression reduces to $e_1'\Gamma^{-1}\Sigma\Gamma^{-1}e_1$ after taking the limit $h \to 0$, provided that x is an interior evaluation point. See the SA for omitted details.) This observation applies to the general case where the local basis function $R(\cdot)$ need not be of polynomial form, or when higher-order derivatives are of interest. See the SA for further discussion and detailed formulas.

In this section we employ a minimum distance approach to develop a lower bound on the asymptotic variance of the local regression distribution estimators, and also propose more efficient estimators based on the observation that their asymptotic variance is of the sandwich form $\Gamma^{-1}\Sigma\Gamma^{-1}$ but with Γ not proportional to Σ even when the uniform kernel is used.

To motivate our approach, notice that in many cases it is possible to specify $R(\cdot)$ in such a way that θ can be partitioned as $\theta = (\theta_1', \theta_2')'$, where $\theta_2 = 0$. In such cases several distinct estimators of θ_1 are available. To describe some leading candidates and their salient properties, partition $\hat{\theta}$, $\hat{\Gamma}$, $\hat{\Sigma}$, and $\hat{\Omega}$ conformable with θ as $\hat{\theta} = (\hat{\theta}_1', \hat{\theta}_2')'$ and

$$\hat{\Gamma} = \begin{pmatrix} \hat{\Gamma}_{11} & \hat{\Gamma}_{12} \\ \hat{\Gamma}_{21} & \hat{\Gamma}_{22} \end{pmatrix}, \qquad \hat{\Sigma} = \begin{pmatrix} \hat{\Sigma}_{11} & \hat{\Sigma}_{12} \\ \hat{\Sigma}_{21} & \hat{\Sigma}_{22} \end{pmatrix}, \qquad \hat{\Omega} = \begin{pmatrix} \hat{\Omega}_{11} & \hat{\Omega}_{12} \\ \hat{\Omega}_{21} & \hat{\Omega}_{22} \end{pmatrix}.$$

The "short" regression counterpart of $\hat{\theta}_1$ obtained by dropping $R_2(\cdot)$ from $R(\cdot) = (R_1(\cdot)', R_2(\cdot)')'$ is given by

$$\hat{\theta}_{\mathrm{R},1} = \hat{\theta}_1 + \hat{\Gamma}_{11}^{-1}\hat{\Gamma}_{12}\hat{\theta}_2,$$

while an optimal minimum distance estimator of θ_1 is given by

$$\hat{\theta}_{\mathrm{MD},1} = \operatorname*{argmin}_{\theta_1} \begin{pmatrix} \hat{\theta}_1 - \theta_1 \\ \hat{\theta}_2 \end{pmatrix}' \hat{\Omega}^{-1} \begin{pmatrix} \hat{\theta}_1 - \theta_1 \\ \hat{\theta}_2 \end{pmatrix} = \hat{\theta}_1 - \hat{\Omega}_{12}\hat{\Omega}_{22}^{-1}\hat{\theta}_2. \tag{7}$$

As a by-product of results obtained when establishing (5), it follows that

$$\begin{aligned} &\hat{\Omega}_{11}^{-1/2}(\hat{\theta}_1 - \theta_1) \rightsquigarrow \mathcal{N}(0, I), \\ &\hat{\Omega}_{\mathrm{R},11}^{-1/2}(\hat{\theta}_{\mathrm{R},1} - \theta_1) \rightsquigarrow \mathcal{N}(0, I), \qquad \hat{\Omega}_{\mathrm{R},11} = \hat{\Gamma}_{11}^{-1}\hat{\Sigma}_{11}\hat{\Gamma}_{11}^{-1}, \\ \text{and} \quad &\hat{\Omega}_{\mathrm{MD},11}^{-1/2}(\hat{\theta}_{\mathrm{MD},1} - \theta_1) \rightsquigarrow \mathcal{N}(0, I), \qquad \hat{\Omega}_{\mathrm{MD},11} = \hat{\Omega}_{11} - \hat{\Omega}_{12}\hat{\Omega}_{22}^{-1}\hat{\Omega}_{21}, \end{aligned}$$

under regularity conditions. Since $\hat{\Omega}$ is of "sandwich" form, the estimators $\hat{\theta}_1$ and $\hat{\theta}_{\mathrm{R},1}$ cannot be ranked in terms of (asymptotic) efficiency in general. On the other hand, $\hat{\theta}_{\mathrm{MD},1}$ will always be (weakly) superior to both $\hat{\theta}_1$ and $\hat{\theta}_{\mathrm{R},1}$. In fact, because

$$\hat{\theta}_1 = \operatorname*{argmin}_{\theta_1} \begin{pmatrix} \hat{\theta}_1 - \theta_1 \\ \hat{\theta}_2 \end{pmatrix}' \begin{pmatrix} \hat{\Omega}_{11}^{-1} & 0 \\ 0 & \hat{\Omega}_{22}^{-1} \end{pmatrix} \begin{pmatrix} \hat{\theta}_1 - \theta_1 \\ \hat{\theta}_2 \end{pmatrix},$$

and

$$\hat{\theta}_{\mathrm{R},1} = \operatorname*{argmin}_{\theta_1} \begin{pmatrix} \hat{\theta}_1 - \theta_1 \\ \hat{\theta}_2 \end{pmatrix}' \hat{\Gamma} \begin{pmatrix} \hat{\theta}_1 - \theta_1 \\ \hat{\theta}_2 \end{pmatrix},$$

each estimator admits a minimum distance interpretation, but only $\hat{\theta}_{\mathrm{MD},1}$ can be interpreted as an optimal minimum distance estimator based on $\hat{\theta}$. See Newey and McFadden (1994) for more discussion on minimum distance estimation.

As a consequence, we investigate whether an appropriately implemented $\hat{\theta}_{\mathrm{MD},1}$ can lead to asymptotic efficiency gains relative to $\hat{\theta}_1$ and $\hat{\theta}_{\mathrm{R},1}$. More generally, as a by-product, we obtain an efficiency bound among minimum distance estimators and show that this bound coincides with those known in the literature for kernel-based density estimation at interior points (Granovsky and Müller, 1991; Cheng, Fan, and Marron, 1997).

In the remaining of this section we focus on the case of local polynomial density estimation at an interior point for concreteness, but the SA presents more general results. Consequently, we assume that $F(\cdot)$ is p-times continuously differentiable in a neighborhood of x. Then, (4) is satisfied and a natural choice of $R(\cdot)$ is

$$R(u) = \left(R_1(u)', R_2(u)'\right)' = \left(1, P(u)', Q(u)'\right)', \tag{8}$$

where $P(u) = (u, u^2/2, \ldots, u^p/p!)'$ is a polynomial basis, and $Q(\cdot)$ represents redundant regressors. Therefore, in our minimum distance construction, the parameters are

$$\theta = \Big(\underbrace{F(\mathsf{x})}_{\text{intercept}}, \underbrace{f(\mathsf{x}), \ \ldots, \ f^{(p-1)}(\mathsf{x})}_{\text{slope, } P(\cdot)}, \underbrace{0, \ \ldots, \ 0}_{\text{redundant, } Q(\cdot)} \Big)', \tag{9}$$

with smoothing error of order $\varrho(h) = o(h^p)$.

With (8) and (9), we define the minimum distance density estimator as $\hat{f}_{\mathrm{MD}}(\mathsf{x}) = e_1'\hat{\theta}_{\mathrm{MD},1}$. Similarly, we have $\hat{f}(\mathsf{x}) = e_1'\hat{\theta}_1$ and $\hat{f}_{\mathrm{R}}(\mathsf{x}) = e_1'\hat{\theta}_{\mathrm{R},1}$. Of course, if it is known a priori that the distribution function is $p+q$ times continuously differentiable, then one can specify $Q(\cdot)$ to include higher order polynomials: $Q(u) = (u^{p+1}/(p+1)!, \ldots, u^{p+q}/(p+q)!)'$. By redefining the parameters as $\theta = (F(\mathsf{x}),\ f(\mathsf{x}),\ \ldots,\ f^{(p+q-1)}(\mathsf{x}))'$, the smoothing error will be of order $\varrho(h) = o(h^{p+q})$. Notice that, in this case, $\hat{f}(\mathsf{x})$ and $\hat{f}_{\mathrm{R}}(\mathsf{x})$ correspond to the density estimator introduced in Cattaneo et al. (2020b) implemented with $R(u) = (1, u, \ldots, u^{p+q}/(p+q)!)'$ and $R(u) = (1, u, \ldots, u^{p}/p!)'$, respectively. Since the purpose of this section is to investigate the efficiency gains of incorporating additional redundant regressors, we do not exploit the extra smoothness condition, and we will treat $Q(\cdot)$ as redundant regressors even if $Q(\cdot)$ contains higher order polynomials.

As both $\hat{f}(\mathsf{x})$ and $\hat{f}_{\mathrm{R}}(\mathsf{x})$ are (weakly) asymptotically inefficient relative to $\hat{f}_{\mathrm{MD}}(\mathsf{x})$ for any choice of $Q(\cdot)$, we consider the asymptotic variance of the minimum distance estimator, which can be obtained by establishing asymptotic counterparts of $\hat{\Gamma}$ and $\hat{\Sigma}$ after suitable scaling. Under regularity conditions (e.g., lack of perfect collinearity between P and Q), the asymptotic variance of the minimum distance ℓth derivative density estimator, $\hat{f}_{\mathrm{MD}}^{(\ell)}(\mathsf{x}) = e_{\ell+1}'\hat{\theta}_{\mathrm{MD},1}$ with $0 \le \ell \le p-1$, is

$$\mathsf{AsyVar}[\hat{f}_{\mathrm{MD}}^{(\ell)}(\mathsf{x})] = e_\ell' \left[\Omega_{PP} - \Omega_{PQ}\Omega_{QQ}^{-1}\Omega_{QP}\right] e_\ell,$$

where

$$\begin{pmatrix} \Omega_{11} & \Omega_{1P} & \Omega_{1Q} \\ \Omega_{P1} & \Omega_{PP} & \Omega_{PQ} \\ \Omega_{Q1} & \Omega_{QP} & \Omega_{QQ} \end{pmatrix} = \Gamma^{-1}\Sigma\Gamma^{-1}.$$

Therefore, the objective is to find a function $Q(\cdot)$ that minimizes the asymptotic variance $\mathsf{AsyVar}[\hat{f}_{\mathrm{MD}}^{(\ell)}(\mathsf{x})]$. Taking $Q(\cdot)$ scalar and properly orthogonalized, without loss of generality, we have $\int_{-1}^{1} P(u)K(u)\mathrm{d}u = 0$ and $\int_{-1}^{1}(1, P(u)')'Q(u)K(u)\mathrm{d}u = 0$. It follows that the problem of selecting an optimal $Q(\cdot)$ to minimize $\mathsf{AsyVar}[\hat{f}_{\mathrm{MD}}^{(\ell)}(\mathsf{x})]$ is equivalent to the following variational problem:

$$\sup_{Q\in\mathcal{Q}} \frac{\left[\int_{-1}^{1}\int_{-1}^{1} P_\ell(u)Q(v)\min\{u,v\}K(u)K(v)\mathrm{d}u\mathrm{d}v\right]^2}{\int_{-1}^{1}\int_{-1}^{1} Q(u)Q(v)\min\{u,v\}K(u)K(v)\mathrm{d}u\mathrm{d}v} \tag{10}$$

where

$$\mathcal{Q} = \left\{Q(\cdot) : \int_{-1}^{1} Q(u)K(u)\mathrm{d}u = 0, \quad \int_{-1}^{1} P(u)Q(u)K(u)\mathrm{d}u = 0\right\},$$

with $P_\ell(u) = e_\ell'\left(\int_{-1}^{1} P(u)P(u)'K(u)\mathrm{d}u\right)^{-1}P(u)$ and $\ell = 1, 2, \ldots, p-1$. The objective function is obtained from the fact that, after proper orthogonalization, the matrix Γ becomes block diagonal. See the SA for all other omitted details.

The following theorem characterizes a lower bound for the asymptotic variance of the minimum distance density estimator among all possible choices of redundant regressors.

Theorem 2 (*Efficiency: Local Polynomial Density Estimator at Interior Points*)**.** *Suppose the conditions of Theorem 1 hold. If* $\mathsf{x} \in \mathcal{X}$ *is an interior point, then*

$$\inf_{Q\in\mathcal{Q}} \mathsf{AsyVar}[\hat{f}_{\mathrm{MD}}^{(\ell)}(\mathsf{x})] \ge \nu_\ell, \qquad \nu_\ell = f(\mathsf{x})e_\ell'\left(\int_{-1}^{1}\dot{P}(u)\dot{P}(u)'\mathrm{d}u\right)^{-1}e_\ell, \qquad 0 \le \ell \le p-1,$$

where $\dot{P}(u) = (1, u, \ldots, u^{p-1}/(p-1)!)'$ *is the derivative of* $P(u)$*.*

This theorem establishes a lower bound among minimum distance estimators. Importantly, it is shown in the SA that this bound coincides with the variance bound of all kernel-type density (and derivatives thereof) estimators employing the same order of the (induced) kernel function (Granovsky and Müller, 1991). Therefore, our minimum distance approach sheds new light on minimum variance results for nonparametric kernel-based estimators of the density function and its derivatives.

This lower bound can be (approximately) achieved by setting the redundant regressor $Q(\cdot)$ to include a certain higher order polynomial function. By direct calculation for each $p = 1, 2, \ldots, 10$, it is also shown in the SA that $\lim_{j\to\infty} \mathsf{AsyVar}[\hat{f}_{\mathrm{MD},j}^{(\ell)}(\mathsf{x})] = \nu_\ell$, where the minimum distance estimator $\hat{f}_{\mathrm{MD},j}^{(\ell)}(\mathsf{x}) = e_\ell'\hat{\theta}_{\mathrm{MD},j}$ is constructed with

$$Q(u) = u^{2j+1} - P(u)'\left(\int_{-1}^{1} P(u)P(u)'\mathrm{d}u\right)^{-1}\int_{-1}^{1} P(u)u^{2j+1}\mathrm{d}u, \quad \text{for } \ell = 0,\ 2,\ 4,\ \ldots,$$

or

$$Q(u) = u^{2j+2} - P(u)'\left(\int_{-1}^{1} P(u)P(u)'\mathrm{d}u\right)^{-1}\int_{-1}^{1} P(u)u^{2j+2}\mathrm{d}u, \quad \text{for } \ell = 1,\ 3,\ 5,\ \ldots,$$

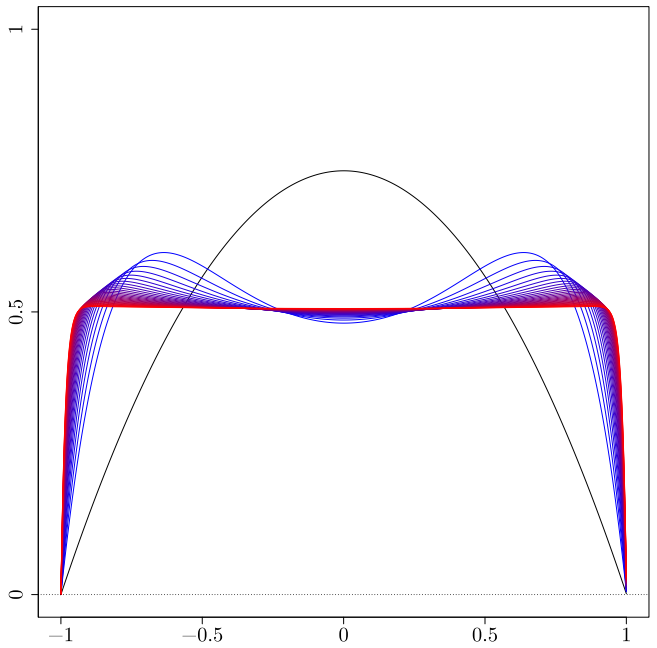


Fig. 1. Equivalent kernels of the minimum distance density estimators. *Notes.* We set $P(u) = u$ or $P(u) = (u, u^2/2)'$, and K uniform. The redundant regressor is $Q(u) = u^{2j+1}$ for $j = 1, 2, \ldots, 30$. The initial equivalent kernel is quadratic (black solid line), and the minimum variance kernel is uniform (red solid line). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and $K(\cdot)$ being the uniform kernel. While we found that other kernel shapes can also be used, we chose the uniform kernel in this construction for three reasons. First, this choice is intuitive and coincides with the optimal choice in standard local polynomial regression settings. Second, when $p \geq 3$ the other allowed kernel shapes overweight observations near the boundary of the kernel's support. Third, the uniform kernel makes the asymptotic variance calculation more tractable. See the SA for further details.

The resulting recipe for implementation is simple: it proposes a specific choice of $Q(\cdot)$ so that the corresponding minimum distance estimator approximately achieves the variance bound for j large enough. Interestingly, $Q(\cdot)$ is scalar and known, but the larger j the closer the asymptotic variance of the minimum distance density estimator will be to the efficiency bound. We assume $Q(\cdot)$ is orthogonal to $P(\cdot)$ for theoretical convenience. To implement this estimator, one only needs to run a local polynomial regression of the empirical distribution function on a constant, the polynomial basis $P(\cdot)$, and one additional regressor, either u^{2j+1} or u^{2j+2} (depending on the choice of ℓ), and then apply (7) with the corresponding estimated variance–covariance matrix.

In Fig. 1, we consider the local linear/quadratic density estimator ($\ell = 0$) with the redundant regressor being a higher order polynomial (i.e., $P(u) = u$ or $P(u) = (u, u^2/2)'$, and $Q(u) = u^{2j+1}$), and plot the corresponding equivalent kernel of our minimum distance density estimator for $j = 1, 2, \ldots, 30$. As j increases, the equivalent kernel converges to the uniform kernel, which is well-known to minimize the (asymptotic) variance among all density estimators employing second order kernels (Granovsky and Müller, 1991). The asymptotic variance of our proposed minimum distance density estimator converges to the optimal asymptotic variance as $j \to \infty$.

Finally, in this paper we focus on minimizing the asymptotic variance of the estimator $\hat{\theta}$ and its variants because our main goal is inference. However, our results could be modified and extended to optimize the asymptotic mean square error (MSE). We do not pursue point estimation optimality further for brevity, but we do note that in the case of local polynomial density estimation (Cattaneo, Jansson, and Ma, 2020b), the resulting estimator is automatically MSE-optimal at interior points when $p \leq 2$, because the induced equivalent kernel coincides with the Epanechnikov kernel (Granovsky and Müller, 1991; Cheng, Fan, and Marron, 1997).

4. Uniform distribution theory

The distributional result presented in Theorem 1 is valid pointwise for $\mathsf{x} \in \mathcal{X}$. We now develop a uniform distributional approximation for the Studentized process

$$\left\{ T(\mathsf{x}) = \frac{c'\hat{\theta}(\mathsf{x}) - c'\theta(\mathsf{x})}{\sqrt{c'\hat{\Omega}(\mathsf{x})c}} \ : \ \mathsf{x} \in \mathcal{I} \right\},$$

using the notation in (5), where c is a conformable vector and $\mathcal{I} \subseteq \mathcal{X}$ is some prespecified region. This stochastic process is not asymptotically tight, and hence does not converge in distribution. Our approximation proceeds in two steps. First, for a positive (vanishing) sequence, $r_{\mathrm{L},n}$, we establish a uniform "linearization" of the process $T(\cdot)$ of the form:

$$\sup_{\mathsf{x}\in\mathcal{I}} |T(\mathsf{x}) - \mathfrak{T}(\mathsf{x})| = O_{\mathbb{P}}(r_{\mathrm{L},n}), \tag{11}$$

where

$$\left\{ \mathfrak{T}(\mathsf{x}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathcal{K}_{h,\mathsf{x}}(X_i) \ : \ \mathsf{x} \in \mathcal{I} \right\}$$

with

$$\mathcal{K}_{h,\mathsf{x}}(X_i) = \frac{c'\Upsilon_h \Gamma_{h,\mathsf{x}}^{-1} \int_{\mathcal{X}_h} R(u)\Big[\mathbb{1}(X_i \le \mathsf{x} + hu) - F(\mathsf{x} + hu)\Big] K(u) f(\mathsf{x} + hu)\mathrm{d}u}{\sqrt{c'\Upsilon_h \Omega_{h,\mathsf{x}} \Upsilon_h c}},$$

and $\Omega_{h,\mathsf{x}} = \Gamma_{h,\mathsf{x}}^{-1} \Sigma_{h,\mathsf{x}} \Gamma_{h,\mathsf{x}}^{-1}$. In words, we show that the Studentized process $T(\cdot)$, which involves various pre-asymptotic estimated quantities, is uniformly close to the linearized process $\mathfrak{T}(\cdot)$, which is a sample average of independent observations. To obtain (11), we develop new uniform approximations with precise convergence rates, which may be of independent interest in semiparametric estimation and inference settings. See the SA for more details.

Second, in a possibly enlarged probability space, we show that there exist a copy of $\mathfrak{T}(\cdot)$, denoted by $\tilde{\mathfrak{T}}(\cdot)$, and a centered Gaussian process $\{\mathfrak{B}(\mathsf{x}) : \mathsf{x} \in \mathcal{I}\}$, with a suitable variance–covariance structure, such that

$$\sup_{\mathsf{x}\in\mathcal{I}} \left|\tilde{\mathfrak{T}}(\mathsf{x}) - \mathfrak{B}(\mathsf{x})\right| = O_{\mathbb{P}}(r_{\mathrm{G},n}), \tag{12}$$

where $r_{\mathrm{G},n}$ is another positive (vanishing) sequence. This type of strong approximation result, when established with suitably fast rate $r_{\mathrm{G},n} \to 0$, can be used to deduce distributional approximations for statistics such as $\sup_{\mathsf{x}\in\mathcal{I}} |T(\mathsf{x})|$, which are useful for constructing confidence bands or for conducting hypothesis tests about shape or other restrictions on the function of interest. To obtain (12), we employ a result established by Rio (1994), and later extended in Giné et al. (2004); see also Giné and Nickl (2010, Proof of their Proposition 5).

In this section we consider a fixed linear combination c for ease of exposition, but in the SA we discuss the more general case where c can depend on both the evaluation point x and the tuning parameter h, which is necessary to establish uniform distribution approximations for the minimum distance estimator introduced in Section 3.3. All the results reported in this section apply to the latter class of estimators as well.

4.1. Assumptions

In addition to Assumption 1, we impose the following conditions on the data generating process. In the sequel, continuity and differentiability conditions at boundary points should be interpreted as one-sided statements (i.e., as in part (i) of Assumption 1).

Assumption 2. Let $\mathcal{I} \subseteq \mathcal{X}$ be a compact interval.

(i) The density function $f(\mathsf{x})$ is twice continuously differentiable and bounded away from zero on $\mathcal{I}$.

(ii) There exist some $\delta > 0$ and compactly supported kernel functions $K^{\dagger}(\cdot)$ and $\{K^{\ddagger,d}(\cdot)\}_{d\le\delta}$, such that (ii.1) $\sup_{u\in\mathbb{R}} |K^{\dagger}(u)| + \sup_{d\le\delta,u\in\mathbb{R}} |K^{\ddagger,d}(u)| < \infty$, (ii.2) the support of $K^{\ddagger,d}(\cdot)$ has Lebesgue measure bounded by Cd, where C is independent of d; and (ii.3) for all u and v such that $|u - v| \le \delta$,

$$|K(u) - K(v)| \le |u - v| \cdot K^{\dagger}(u) + K^{\ddagger,|u-v|}(u).$$

(iii) The basis function $R(\cdot)$ is Lipschitz continuous in $[-1, 1]$.

(iv) For all h sufficiently small, the minimum eigenvalues of $\Gamma_{h,\mathsf{x}}$ and $h^{-1}\Sigma_{h,\mathsf{x}}$ are bounded away from zero uniformly for $\mathsf{x} \in \mathcal{I}$.

The above strengthens and expands Assumption 1. Part (i) requires the density function to be reasonably smooth uniformly in $\mathcal{I}$. Part (ii) imposes additional requirements on the kernel function. Although seemingly technical, it permits a decomposition of the difference $|K(u) - K(v)|$ into two parts. The first part, $|u - v| \cdot K^{\dagger}(u)$, is a kernel function which vanishes uniformly as $|u - v|$ becomes small. Note that this will be the case for all piecewise smooth kernel functions, such as the triangular or the Epanechnikov kernel. However, difference of discontinuous kernels, such as the uniform kernel, cannot be made uniformly close to zero. This motivates the second term in the above decomposition. Part (iii) requires the basis function to be reasonably smooth. Together, parts (i)–(iii) imply that the estimator $\hat{\theta}(\mathsf{x})$ will be "smooth" in x, which is important to control the complexity of the process $T(\cdot)$. Finally, part (iv) implies that the matrices $\Gamma_{h,\mathsf{x}}$ and $\Sigma_{h,\mathsf{x}}$ are well-behaved uniformly for $\mathsf{x} \in \mathcal{I}$.

4.2. Strong approximation

We first discuss the covariance of the process $\mathfrak{T}(\cdot)$. It is straightforward to show that

$$\mathbb{C}\mathrm{ov}[\mathfrak{T}(\mathsf{x}),\mathfrak{T}(\mathsf{y})]=\frac{c'\Upsilon_h\Omega_{h,\mathsf{x},\mathsf{y}}\Upsilon_h c}{\sqrt{c'\Upsilon_h\Omega_{h,\mathsf{x}}\Upsilon_h c}\sqrt{c'\Upsilon_h\Omega_{h,\mathsf{y}}\Upsilon_h c}},\qquad \Omega_{h,\mathsf{x},\mathsf{y}}=\Gamma_{h,\mathsf{x}}^{-1}\Sigma_{h,\mathsf{x},\mathsf{y}}\Gamma_{h,\mathsf{y}}^{-1},$$

where

$$\Sigma_{h,\mathsf{x},\mathsf{y}}=\int_{\mathcal{X}_{h,\mathsf{y}}}\int_{\mathcal{X}_{h,\mathsf{x}}}R(u)R(v)'\Big[F(\min\{\mathsf{x}+hu,\mathsf{y}+hv\})-F(\mathsf{x}+hu)F(\mathsf{y}+hv)\Big]\cdot K(u)K(v)f(\mathsf{x}+hu)f(\mathsf{y}+hv)\mathrm{d}u\mathrm{d}v,$$

and $\Sigma_{h,\mathsf{x},\mathsf{x}}=\Sigma_{h,\mathsf{x}}$.

Now we state the second main distributional result of this paper in the following theorem.

Theorem 3 (*Strong Approximation*). *Suppose Assumptions 1 and 2 hold, and that* $h\to 0$ *and* $nh^2/\log(n)\to\infty$.

1. (11) *holds with*
$$r_{\mathrm{L},n}=\sqrt{\frac{n}{h}}\sup_{\mathsf{x}\in\mathcal{I}}\varrho(h,\mathsf{x})+\frac{\log(n)}{\sqrt{nh^2}}.$$
2. *On a possibly enlarged probability space, there exist a copy* $\tilde{\mathfrak{T}}(\cdot)$ *of* $\mathfrak{T}(\cdot)$, *and a centered Gaussian process,* $\{\mathfrak{B}(\mathsf{x}),\mathsf{x}\in\mathcal{I}\}$, *defined with the same covariance as* $\mathfrak{T}(\cdot)$, *such that* (12) *holds with*
$$r_{\mathrm{G},n}=\frac{\log(n)}{\sqrt{nh}}.$$

The first part of this theorem gives conditions such that the feasible Studentized process $T(\cdot)$ is well approximated by the infeasible (linear) process $\mathfrak{T}(\cdot)$, uniformly for $\mathsf{x}\in\mathcal{I}$. The latter process is mean zero, and takes a kernel-based form. However, standard strong approximation results for kernel-type estimators do not apply directly to the process $\mathfrak{T}(\cdot)$, as the implied (equivalent, Studentized) kernel $\mathcal{K}_{h,\mathsf{x}}(\cdot)$ depends not only on the bandwidth but also on the evaluation point in a non-standard way. That is, due to the boundary adaptive feature of the local regression distribution estimators, the shape of the implied kernel automatically changes for different evaluation points depending on whether they are interior or boundary points.

Putting the two results together, it follows that the distribution of $T(\cdot)$ is approximated by that of $\mathfrak{B}(\cdot)$, provided the following condition holds:

$$\sqrt{\frac{n}{h}}\sup_{\mathsf{x}\in\mathcal{I}}\varrho(h,\mathsf{x})+\frac{\log(n)}{\sqrt{nh^2}}\to 0.$$

To facilitate understanding of this rate restriction, we consider the local polynomial density estimation setting of Cattaneo et al. (2020b), where the basis function takes the form $R(u)=(1,u,u^2/2,\ldots,u^p/p!)'$ for some $p\geq 1$, and the second element of $\hat{\theta}(\mathsf{x})$ estimates the density $f(\mathsf{x})$. That is, $e_1'\hat{\theta}(\mathsf{x})=\hat{f}(\mathsf{x})\to_{\mathbb{P}}f(\mathsf{x})$ under Assumption 1, where $c=e_1$. By a Taylor expansion argument, it is easy to see that the smoothing bias has order h^{p+1} as long as the distribution function $F(\cdot)$ is suitably smooth. Then, the above rate restriction reduces to $\sqrt{nh^{2p+1}}+\frac{\log(n)}{\sqrt{nh^2}}\to 0$.

Finally, if the goal is to approximate the distribution of $\sup_{\mathsf{x}\in\mathcal{I}}|T(\mathsf{x})|$, then an extra $\sqrt{\log(n)}$ factor is needed in the rate restriction, as discussed in Chernozhukov et al. (2014a). A formal statement of such result is given below, after we discuss how we can further approximate the infeasible Gaussian process $\mathfrak{B}(\cdot)$.

4.3. Confidence bands

Feasible inference cannot be based on the Gaussian process $\mathfrak{B}(\cdot)$, as its covariance structure is unknown and has to be estimated in practice. For estimation, first recall from Sections 2 and 3 that $W_i(\mathsf{x})=K((x_i-\mathsf{x})/h)/h$, $R_i(\mathsf{x})=R(x_i-\mathsf{x})$, and $\hat{\Gamma}(\mathsf{x})=\frac{1}{n}\sum_{i=1}^n W_i(\mathsf{x})R_i(\mathsf{x})R_i(\mathsf{x})'$. Then, we construct the plug-in estimator of $\Omega_{h,\mathsf{x},\mathsf{y}}$ as follows:

$$\hat{\Omega}_{h,\mathsf{x},\mathsf{y}}=n\Upsilon_h^{-1}\hat{\Gamma}(\mathsf{x})^{-1}\hat{\Sigma}(\mathsf{x},\mathsf{y})\hat{\Gamma}(\mathsf{y})^{-1}\Upsilon_h^{-1},\qquad \hat{\Sigma}(\mathsf{x},\mathsf{y})=\frac{1}{n^2}\sum_{i=1}^n\hat{\psi}_i(\mathsf{x})\hat{\psi}_i(\mathsf{y})'$$

where

$$\hat{\psi}_i(\mathsf{x})=\frac{1}{n}\sum_{j=1}^n W_j(\mathsf{x})R_j(\mathsf{x})(\mathbb{1}(x_i\leq x_j)-\hat{F}_j).$$

The following theorem combines previous results, and justifies the uniform confidence band constructed using critical values from $\sup_{\mathsf{x}\in\mathcal{I}}|\hat{\mathfrak{B}}(\mathsf{x})|$. Let $X_n=(x_1,x_2,\ldots,x_n)'$.

Theorem 4 (*Kolmogorov–Smirnov Distance*)**.** *Suppose Assumptions 1 and 2 hold, and that* $n \sup_{\mathsf{x}\in\mathcal{I}} \varrho(h,\mathsf{x})^2 \log(n)/h + \log(n)^5/(nh^2) \to 0$. *Then, conditional on* X_n, *there exists a centered Gaussian process* $\{\hat{\mathfrak{B}}(\mathsf{x}), \mathsf{x}\in\mathcal{I}\}$ *with covariance*

$$\mathbb{C}\mathrm{ov}\left[\hat{\mathfrak{B}}(\mathsf{x}), \hat{\mathfrak{B}}(\mathsf{y}) \middle| X_n\right] = \frac{c'\Upsilon_h\hat{\Omega}_{h,\mathsf{x},\mathsf{y}}\Upsilon_h c}{\sqrt{c'\Upsilon_h\hat{\Omega}_{h,\mathsf{x}}\Upsilon_h c}\sqrt{c'\Upsilon_h\hat{\Omega}_{h,\mathsf{y}}\Upsilon_h c}},$$

such that

$$\sup_{u\in\mathbb{R}}\left|\mathbb{P}\left[\sup_{\mathsf{x}\in\mathcal{I}}|T(\mathsf{x})|\le u\right] - \mathbb{P}\left[\sup_{\mathsf{x}\in\mathcal{I}}|\hat{\mathfrak{B}}(\mathsf{x})|\le u\middle| X_n\right]\right| = o_{\mathbb{P}}(1).$$

From Theorem 4, an asymptotically valid $100(1-\alpha)\%$ confidence band for $\{c'\theta(\mathsf{x}) : \mathsf{x}\in\mathcal{I}\}$ is given by

$$\mathrm{CI}_\alpha(\mathcal{I}) = \left\{\left[c'\hat{\theta}(\mathsf{x}) - \mathfrak{q}_{1-\alpha}\sqrt{c'\hat{\Omega}(\mathsf{x})c}\ ,\ c'\hat{\theta}(\mathsf{x}) + \mathfrak{q}_{1-\alpha}\sqrt{c'\hat{\Omega}(\mathsf{x})c}\right],\quad \mathsf{x}\in\mathcal{I}\right\},$$

where $\mathfrak{q}_{1-\alpha}$ is the $1-\alpha$ quantile of $\sup_{\mathsf{x}\in\mathcal{I}}|\hat{\mathfrak{B}}(\mathsf{x})|$, conditional on the data. That is,

$$\mathfrak{q}_a = \inf\left\{u\in\mathbb{R} : \mathbb{P}\left[\sup_{\mathsf{x}\in\mathcal{I}}|\hat{\mathfrak{B}}(\mathsf{x})|\le u\middle| X_n\right]\ge a\right\},$$

which can be obtained by simulating the process $\hat{\mathfrak{B}}(\cdot)$ on a dense grid.

As an alternative to analytic estimation of the covariance kernel, it is possible to consider resampling methods as in Chernozhukov et al. (2014a), Cheng and Chen (2019), Cattaneo et al. (2020a), and references therein. We relegate resampling-based inference for future research.

5. Extensions and other applications

We briefly outline some extensions of our main results. First, we introduce a re-weighted version of $\hat{\theta}$, which is useful in applications as illustrated in Section 6. Second, we discuss a new class of local regression estimators based on a non-random L^2 loss function, which has some interesting theoretical properties and may be of interest in some semiparametric settings. Finally, we discuss how to incorporate restrictions into the estimation procedure, employing the generic structure of the local basis $R(\cdot)$.

5.1. Re-weighted distribution estimator

Suppose $(x_1, w_1), (x_2, w_2), \ldots, (x_n, w_n)$ is a random sample, where x_i is a continuous random variable with a smooth cumulative distribution function, but now w_i is an additional "weighting" variable, possibly random and involving unknown parameters. We consider the generic weighted distribution parameter

$$H(\mathsf{x}) = \mathbb{E}[w_i\mathbb{1}(x_i\le\mathsf{x})],$$

whose practical interpretation depends on the specific choice of w_i.

We discuss some examples. If $w_i = 1$, $H(\cdot)$ becomes the distribution function $F(\cdot)$, and hence the results above apply. If w_i is set to be a certain ratio of propensity scores for subpopulation membership, then the derivative $\mathrm{d}H(\mathsf{x})/\mathrm{d}\mathsf{x}$ becomes a counterfactual density function, as in DiNardo et al. (1996); see Section 6.1. If w_i is set to be a combination of the treatment assignment and treatment status variables, then the resulting derivative can be used to conduct specification testing in IV models, or if w_i is set to be a certain ratio of propensity scores for a binary instrument, then the derivative can be used to identify distributions of compliers, as in Imbens and Rubin (1997), Abadie (2003), and Kitagawa (2015); see Section 6.2. Other examples of applicability of this extension include bunching, missing data, measurement error, data combination, and treatment effect settings.

More generally, when weights are allowed for, there is another potentially interesting connection between the estimand $\mathrm{d}H(\mathsf{x})/\mathrm{d}\mathsf{x}$ and classical weighted averages featuring prominently in econometrics because $\mathrm{d}H(\mathsf{x})/\mathrm{d}\mathsf{x} = \mathbb{E}[w_i|x_i = \mathsf{x}]f(\mathsf{x})$, which is useful in the context of partial means and related problems as in Newey (1994b).

Our main results extend immediately to allow for $\sqrt{n}$-consistent estimated weights w_i or, more generally, to estimated weights that converge sufficiently fast. Specifically, we let $\hat{F}_{w,i}(x) = \frac{1}{n}\sum_{j=1}^n w_j\mathbb{1}(x_j\le x)$ in place of $\hat{F}_i$, and investigate the large sample properties of our proposed estimator in (2) when w_i is replaced by $\hat{w}_i = w_i(\hat{\beta})$ with $\hat{\beta}$ an a_n-consistent estimator, for some $a_n\to\infty$, and $w_i(\cdot)$ a known function of the data. That is, when estimated weights are used to construct the weighted empirical distribution function $\hat{F}_{w,i}(x)$. Provided that $a_n^{-1}\to 0$ sufficiently fast, this extra estimation step will not affect the asymptotic properties of our estimator of the density function or its derivatives (which will be true, for example, in parametric estimation cases, where $a_n = \sqrt{n}$ under regularity conditions). All the results reported in the previous sections apply to this extension, which we illustrate empirically below.

5.2. Local L^2 distribution estimators

The local regression distribution estimator is obtained from a least squares projection of the empirical distribution function onto a local basis, where the projection puts equal weights at all observations. That is, (2) employs an $L^2(\hat{F})$-projection

$$\hat{\theta}(\mathsf{x}) = \underset{\theta}{\operatorname{argmin}} \int \left(\hat{F}(u) - R(u-\mathsf{x})'\theta\right)^2 K\left(\frac{u-\mathsf{x}}{h}\right) \mathrm{d}\hat{F}(u).$$

This representation motivates a general class of local L^2 distribution estimators given by

$$\hat{\theta}_G(\mathsf{x}) = \underset{\theta}{\operatorname{argmin}} \int \left(\hat{F}(u) - R(u-\mathsf{x})'\theta\right)^2 K\left(\frac{u-\mathsf{x}}{h}\right) \mathrm{d}G(u)$$

for some measure G. We show in the SA that all our theoretical results continue to hold for $\hat{\theta}_G$, provided that G is absolutely continuous with respect to the Lebesgue measure and the Radon–Nikodym derivative is reasonably smooth. (Note that G does not need to be a proper distribution function.)

The estimator $\hat{\theta}_G$ involves only one average, while the local regression estimator $\hat{\theta}$ has two layers of averages (one from the construction of the empirical distribution function, and the other from the $L^2(\hat{F})$-projection/regression). As a result, with suitable centering and scaling, the local L^2 distribution estimator, $\hat{\theta}_G$, can be written as the sum of a mean-zero influence function and a smoothing bias term. Since $\hat{\theta}_G$ no longer involves a second order U-statistic (cf. (6)), or a leave-in bias, pointwise asymptotic normality can be established under weaker conditions: it is no longer needed to assume $nh^2 \to \infty$ (Theorem 1), and $nh \to \infty$ will suffice. Similarly, for the strong approximation results we only need to restrict $\log(n)/\sqrt{nh}$ as opposed to $\log(n)/\sqrt{nh^2}$ (part 1 of Theorem 3).

In addition, the local L^2 distribution estimator $\hat{\theta}_G$ is robust to "low" density. To see this, recall that the local regression estimator $\hat{\theta}$ involves regressing the empirical distribution on a local basis, which means that this estimator can be numerically unstable if there are only a few observations near the evaluation point. More precisely, the matrix $\hat{\Gamma}$ will be close to singular if the effective sample size is small.

Although the local L^2 distribution estimator $\hat{\theta}_G$ takes a simpler form, is robust to low density, and its large sample properties can be established under weaker bandwidth conditions, it does have one drawback: it requires knowledge of the support $\mathcal{X}$. To be more precise, let G be the Lebesgue measure, then the local L^2 distribution estimator may be biased at or near boundaries of $\mathcal{X}$ if it is compact. In contrast, the local regression distribution estimator is fully boundary adaptive, even in cases where the location of the boundary is unknown. See Cattaneo et al. (2020b) for further discussion for the case of density estimation.

5.3. Incorporating restrictions

The formulation (2) is general enough to allow for some interesting extensions in the definition of the local regression distribution estimator. The key observation is that the estimator has a weighted least squares representation with a generic local basis function $R(\cdot)$, which allows for deploying well-known results from linear regression models. We briefly illustrate this idea with three examples.

First, consider the case where the local basis $R(u)$ incorporates specific restrictions, such as continuity or lack thereof, on the distribution function, density function or higher-order derivatives at the evaluation point x. To give a concrete example, suppose that $F(\mathsf{x})$ and $f(\mathsf{x})$ are known to be continuous at some interior point $\mathsf{x} \in \mathcal{X}$, while no information is available for the higher-order derivatives. Then, these restrictions can be effortlessly incorporated into the local regression distribution estimator by considering the local basis function

$$R(u) = \left(1, u, \frac{u^2}{2}\mathbb{1}(u < \mathsf{x}), \frac{u^2}{2}\mathbb{1}(u \geq \mathsf{x}), \frac{u^3}{6}\mathbb{1}(u < \mathsf{x}), \frac{u^3}{6}\mathbb{1}(u \geq \mathsf{x}), \ldots, \frac{u^p}{p!}\mathbb{1}(u < \mathsf{x}), \frac{u^p}{p!}\mathbb{1}(u \geq \mathsf{x})\right)'.$$

It follows that $\hat{f}(\mathsf{x}) = e_1'\hat{\theta}(\mathsf{x})$ consistently estimates the density $f(\mathsf{x})$ at the kink point x, while $e_2'\hat{\theta}(\mathsf{x})$ and $e_3'\hat{\theta}(\mathsf{x})$ are consistent estimators of the left and the right derivative of the density function, respectively (and similarly for other higher-order one-sided derivatives). In this example, the generalized formulation not only reduces the bias of $\hat{f}(\mathsf{x}) = e_1'\hat{\theta}(\mathsf{x})$ even in the absence of continuity of higher-order derivatives, but also provides the basis for testing procedures for continuity of higher-order derivatives; e.g., by considering a statistic based on $(e_2 - e_3)'\hat{\theta}(\mathsf{x})$. This provides a concrete illustration of the advantages of allowing for generic local basis. A distinct example was developed in Cattaneo et al. (2018, 2020b) for density discontinuity testing in regression discontinuity designs.

As a second example, consider imposing shape constraints, such as positivity or monotonicity, in the construction of the local regression distribution estimator. Such constraints amount to specific restrictions on the parameter space of θ, which naturally leads to restricted weighted least squares estimation in the context of our estimator. To be concrete, consider constructing a local polynomial density estimator which is non-negative, in which case $R(u)$ is a polynomial basis

of order $p \geq 1$ and (2) is extended to:

$$\hat{\theta}(\mathsf{x}) = \operatorname*{argmin}_{\theta} \sum_{i=1}^{n} W_i(\hat{F}_i - R_i'\theta)^2 \qquad \text{subject to } T\theta \geq 0,$$

where T denotes a matrix of restrictions; in this example, $T = e_1'$ to ensure that $\hat{f}(\mathsf{x}) = e_1'\hat{\theta}(\mathsf{x}) \geq 0$. This example showcases the advantages of the weighted least squares formulation of our estimator. Local monotonicity constraints, for instance, could also be easily incorporated in a similar fashion.

The final example of extensions of our basic local regression distribution estimation approach concerns non-identity link functions, leading to a non-linear least squares formulation. Specifically, (2) can be generalized to $\hat{\theta}(\mathsf{x}) = \operatorname{argmin}_{\theta} \sum_{i=1}^{n} W_i(\hat{F}_i - \Lambda(R_i'\theta))^2$ for some known link function $\Lambda(\cdot)$. For instance, such extension may be useful to model distributions with large support or to impose specific local shape constraints.

All of the examples above, as well as many others, can be analyzed using the large sample results developed in this paper and proper extensions thereof. We plan to investigate these and other extensions in future research.

6. Applications

We discuss two applications of our main results in the context of program evaluation (see Abadie and Cattaneo, 2018, and references therein).

6.1. Counterfactual densities

In this first application, the objects of interest are density functions over their entire support, including boundaries and near-boundary regions, which are estimated using estimated weighting schemes, as this is a key feature needed for counterfactual analysis (and many other applications). Our general estimation strategy is specialized to the counterfactual density approach originally proposed by DiNardo et al. (1996). We focus on density estimation, and we refer readers to Chernozhukov et al. (2013) for related methods based on distribution functions as well as for an overview of the literature on counterfactual analysis.

To construct a counterfactual density or, more generally, re-weighted density estimators, we simply need to set the weights $(w_1, w_2, \ldots, w_n)$ appropriately. In most applications, this also requires constructing preliminary consistent estimators of these weights, as we illustrate in this section. Suppose the observed data is $(x_i, t_i, z_i')'$, $i = 1, 2, \ldots, n$, where x_i continues to be the main outcome variable, z_i collects other covariates, and t_i is a binary variable indicating to which group unit i belongs. For concreteness, we call these two groups control and treatment, though our discussion does not need to bear any causal interpretation.

The marginal distribution of the outcome variable x_i for the full sample can be easily estimated without weights (that is, $w_i = 1$). In addition, two conditional densities, one for each group, can be estimated using $w_i^1 = t_i/\mathbb{P}[t_i = 1]$ for the treatment group and $w_i^0 = (1 - t_i)/\mathbb{P}[t_i = 0]$ for the control group, and are denote by $\hat{f}_1(x)$ and $\hat{f}_0(x)$, respectively. For example, in the context of randomized controlled trials, these density estimators can be useful to depict the distribution of the outcome variables for control and treatment units.

A more challenging question is: what would the outcome distribution have been, had the treated units had the same covariates distribution as the control units? The resulting density is called the counterfactual density for the treated, which is denoted by $f_{1\triangleright 0}(x)$. Knowledge about this distribution is important for understanding differences between $f_1(x)$ and $f_0(x)$, as the outcome distribution is affected by both group status and covariates distribution. Furthermore, the counterfactual distribution has another useful interpretation: Assume the outcome variable is generated from potential outcomes, $x_i = t_i x_i(1) + (1 - t_i)x_i(0)$, then under unconfoundedness, that is, assuming t_i is independent of $(x_i(0), x_i(1))'$ conditional on the covariates z_i, $f_{1\triangleright 0}(x)$ is the counterfactual distribution for the control group: it is the density function associated with the distribution of $x_i(1)$ conditional on $t_i = 0$.

Regardless of the interpretation taken, $f_{1\triangleright 0}(x)$ is of interest and can be estimated using our generic density estimator $\hat{f}(x)$ with the following weights:

$$w_i^{1\triangleright 0} = t_i \cdot \frac{\mathbb{P}[t_i = 0|z_i]}{\mathbb{P}[t_i = 1|z_i]} \frac{\mathbb{P}[t_i = 1]}{\mathbb{P}[t_i = 0]}.$$

In practice, this choice of weighting scheme is unknown because the conditional probability $\mathbb{P}[t_i = 1|z_i]$, a.k.a. the propensity score, is not observed. Thus, researchers estimate this quantity using a flexible parametric model, such as Probit or Logit. Our technical results allow for these estimated weights to form counterfactual density estimators after replacing the theoretical weights by their estimated counterparts, provided the estimated weights converge sufficiently fast to their population counterparts.

To be more precise, we can model $\mathbb{P}[t_i = 1|z_i] = G(b(z_i)'\beta)$ for some known link function $G(\cdot)$, such as Logit or Probit, and K-dimensional basis expansion $b(z_i)$, such as power series or B-splines. If the model is correctly specified for some fixed K and basis function $b(\cdot)$, then $\max_{1\leq i\leq n} |w_i - \hat{w}_i| = O_{\mathbb{P}}(a_n^{-1})$ with $a_n = \sqrt{n}$ under mild regularity conditions, and all our results carry over to the setting with estimated weights mentioned in Section 5.1. Alternatively, from a nonparametric perspective, if $K \to \infty$ as $n \to \infty$, and for appropriate basis function $b(\cdot)$ and regularity conditions, $\max_{1\leq i\leq n} |w_i - \hat{w}_i| = O_{\mathbb{P}}(a_n^{-1})$ with a_n depending on both K and n. Then, as in the parametric case, our main results carry over if $a_n^{-1} \to 0$ fast enough. The exact rate requirements can be deduced from the main theorems above.

Table 1
Summary Statistics for the JTPA data.

	Full	JTPA offer		JTPA enrollment	
		N	Y	N	Y
Income	17949.20	17191.13	18321.59	17015.58	19098.44
HS or GED	0.72	0.71	0.72	0.70	0.74
Male	0.46	0.47	0.46	0.48	0.45
Nonwhite	0.36	0.36	0.36	0.36	0.37
Married	0.28	0.27	0.29	0.27	0.29
Work $\leq$ 12	0.44	0.43	0.44	0.44	0.44
AFDC	0.17	0.17	0.17	0.16	0.19
Age					
22–25	0.24	0.25	0.24	0.24	0.25
26–29	0.21	0.20	0.21	0.21	0.21
30–35	0.24	0.25	0.24	0.24	0.25
36–44	0.19	0.19	0.19	0.20	0.19
45–54	0.08	0.08	0.08	0.08	0.07
Sample size	9872	3252	6620	5447	4425

Columns: (i) Full: full sample; (ii) JTPA Offer: whether offered JTPA services; (iii) JTPA Enrollment: whether enrolled in JTPA. *Rows*: (i) Income: cumulative income over 30-month period post random selection; (ii) HS or GED: whether has high school degree or GED; (iii) Male: gender being male; (iv) Nonwhite: black or Hispanic; (v) Married: whether married; (vi) Work $\leq$ 12: worked less than 12 weeks during one year period prior to random assignment; (vii) Age: age groups.

6.1.1. Empirical illustration

We demonstrate empirically how marginal, conditional, and counterfactual densities can be estimated with our proposed method. We consider the effect of education on earnings using a subsample of the data in Abadie et al. (2002). The data consists of individuals who did not enroll in the Job Training Partnership Act (JTPA). The main outcome variable is the sum of earnings in a 30-month period, and individuals are split into two groups according to their education attainment: $t_i = 1$ for those with high school degree or GED, and $t_i = 0$ otherwise. Also available are demographic characteristics, including gender, ethnicity, age, marital status, AFDC receipt (for women), and a dummy indicating whether the individual worked at least 12 weeks during a one-year period. The sample size is 5447, with 3927 being either high school graduates or GED. Summary statistics are available as the fourth column in Table 1. We leave further details on the JTPA program to Section 6.2, where we utilize a larger sample and conduct distribution estimation in a randomized controlled (intention-to-treat) and instrumental variables (imperfect compliance) setting.

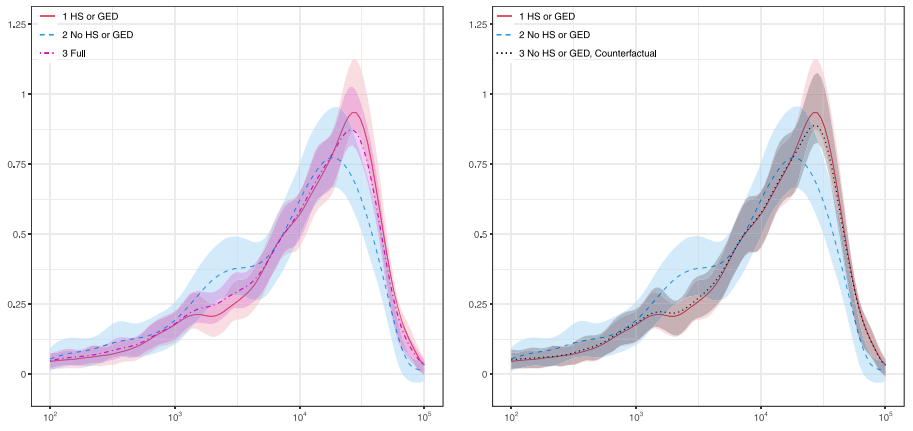
It is well-known that education has significant impact on labor income, and we first plot earning distributions separately for subsamples with and without high school degree or GED. The two estimates, $\hat{f}_1(x)$ and $\hat{f}_0(x)$, are plotted in panel (a) of Fig. 2. There, it is apparent that the earning distribution for high school graduates is very different compared to those without high school degree. More specifically, both the mean and median of $\hat{f}_1(x)$ are higher than $\hat{f}_0(x)$, and $\hat{f}_1(x)$ seems to have much thinner left tail and thicker right tail.

As mentioned earlier, direct comparison between $\hat{f}_1(x)$ and $\hat{f}_0(x)$ does not reveal the impact of having high school degree on earning, since the difference is confounded by the fact that individuals with high school degree can have very different characteristics (measured by covariates) compared to those without. We employ covariates adjustments, and ask the following question: what would the earning distribution have been for high school graduates, had they had the same characteristics as those without such degree?

We estimate the counterfactual distribution $f_{1\triangleright 0}(x)$ by our proposed method, and is shown in panel (b) of Fig. 2. The difference between $\hat{f}_{1\triangleright 0}(x)$ and $\hat{f}_1(x)$ is not very profound, although it seems $\hat{f}_{1\triangleright 0}(x)$ has smaller mean and median. On the other hand, difference between $\hat{f}_0(x)$ and $\hat{f}_{1\triangleright 0}(x)$ remains highly nontrivial. Our empirical finding is compatible with existing literature on return to education: it is generally believed that education leads to significant accumulation of human capital, hence increase in labor income. As a result, educational attainment is usually one of the most important "explanatory variables" for differences in income.

6.2. IV specification and heterogeneity

Self-selection and treatment effect heterogeneity are important concerns in causal inference and studies of socioeconomic programs. It is now well understood that classical treatment parameters, such as the average treatment effect or the treatment effect on the treated, are not identifiable even when treatment assignment is fully randomized due to imperfect compliance. Indeed, what can be recovered is either an intention-to-treat parameter or, using the instrumental variables method, some other more local treatment effect, specific to a subpopulation: the "compliers". See Imbens and Rubin (2015) and references therein for further discussion. Practically, this poses two issues for empirical work employing instrumental variables methods focusing on local average treatment effects. First, since compliers are usually not identified, it is crucial to understand how different their characteristics are compared to the population as a whole. Second, it is often desirable



(a) Marginal Distributions (b) Counterfactual Distribution

Fig. 2. Earning distributions by education, JTPA. *Notes*: (i) Full: earning distribution for the full sample ($n = 5447$); (ii) HS or GED: earning distributions for subgroups with and without high school degree or GED ($n = 1520$ and 3927, respectively); (iii) No HS or GED, Counterfactual: counterfactual earning distribution. Point estimates are obtained by using local polynomial regression with order 2, and robust confidence intervals are obtained with local polynomial of order 3. Bandwidths are chosen by minimizing integrated mean squared errors. All estimates are obtained using companion R (and Stata) package described in Cattaneo et al. (2021b).

to have a thorough estimate of the distribution of potential outcomes, which provides information not only on the mean or median, but also on its dispersion, overall shape, or local curvatures.

Motivated by these observations, and to illustrate the applicability of our density estimation methods, we now consider two related problems. First, we investigate specification testing in the context of local average treatment effects based on comparison of two (rescaled) densities as discussed by Kitagawa (2015). This method requires estimating two densities nonparametrically. Second, we consider estimating the density of potential outcomes for compliers in the IV setting of Abadie (2003), which allows for conditioning on covariates. The resulting density plots not only provide visual guides on treatment effects, but also can be used for further analysis to construct a rich set of summary statistics or as inputs for semiparametric procedures. Both methods require estimated weights.

We first introduce the notation and the potential outcomes framework. For each individual there is a binary indicator of treatment assignment (a.k.a. the instrument), denoted by d_i. The actual treatment (takeup), however, can be different, due to imperfect compliance. More specifically, let $t_i(0)$ and $t_i(1)$ be the two potential treatments, corresponding to $d_i = 0$ and 1, then the observed binary treatment indicator is $t_i = d_i t_i(1)+(1-d_i)t_i(0)$. We also have a pair of potential outcomes, $x_i(0)$ and $x_i(1)$, associated with $t_i = 0$ and 1, and what is observed is $x_i = t_i x_i(1) + (1 - t_i)x_i(0)$. Finally, also available are some covariates, collected in z_i. We assume that the observed data is a random sample $\{(x_i, t_i, d_i, z_i')' : 1 \le i \le n\}$.

There are three important assumptions for identification. First, the instrument has to be exogenous, meaning that conditional on covariates, it is independent of the potential treatments and outcomes. Second, the instrument has to be relevant, meaning that conditional on covariates, the instrument should be able to induce changes in treatment takeups. Third, there are no defiers (a.k.a. the monotonicity assumption). We do not reproduce the exact details of those assumptions and other technical requirements for identification; see the references given for more details.

Building on Imbens and Rubin (1997), Kitagawa (2015) discusses interesting testable implications in this IV setting, which can be easily adapted to test instrument validity using our density estimator. In the current context, the testable implications take the following form: for any (measurable) set $\mathcal{I} \subset \mathbb{R}$,

$$\mathbb{P}[x_i \in \mathcal{I},\ t_i = 1|d_i = 1] \ge \mathbb{P}[x_i \in \mathcal{I},\ t_i = 1|d_i = 0],$$
$$\text{and} \quad \mathbb{P}[x_i \in \mathcal{I},\ t_i = 0|d_i = 0] \ge \mathbb{P}[x_i \in \mathcal{I},\ t_i = 0|d_i = 1].$$

The first requirement holds trivially in the JTPA context, since the program does not allow enrollment without being offered (that is, $\mathbb{P}[t_i = 1|d_i = 0] = 0$). Therefore we demonstrate the second with our density estimator. Let $f_{d=0,t=0}(x)$ be the earning density for the subsample $d_i = 0$ and $t_i = 0$, that is, for individuals without JTPA offer and not enrolled. Similarly let $f_{d=1,t=0}(x)$ be the earning density for individuals offered JTPA but not enrolled. Then the second inequality in the above display is equivalent to, for all $x \in \mathbb{R}$,

$$\mathbb{P}[t_i = 0|d_i = 0] \cdot f_{d=0,t=0}(x) \ge \mathbb{P}[t_i = 0|d_i = 1] \cdot f_{d=1,t=0}(x).$$

Thus, our density estimator can be used directly, where $f_{d=0,t=0}(x)$ is consistently estimated with weights $w_i^{d=0,t=0} = (1 - d_i)(1 - t_i)/\mathbb{P}[d_i = 0, t_i = 0]$, and $f_{d=1,t=0}(x)$ is consistently estimated with $w_i^{d=1,t=0} = d_i(1 - t_i)/\mathbb{P}[d_i = 1, t_i = 0]$.

Abadie (2003) showed that the distributional characteristics of compliers are identified, and can be expressed as re-weighted marginal quantities. We focus on three distributional parameters here. The first one is the distribution of the

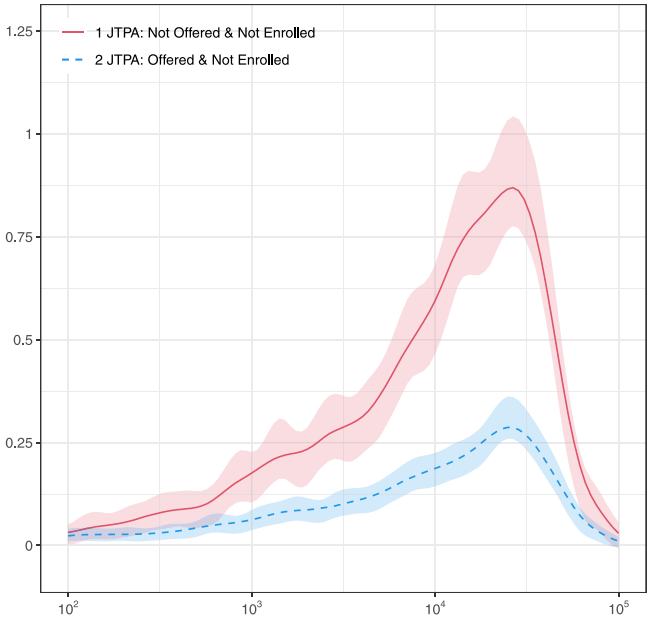


Fig. 3. Testing validity of instruments, JTPA. *Notes*: (i) JTPA: Not Offered & Not Enrolled: the scaled density estimate $\frac{\sum_i \mathbb{1}(t_i=0,d_i=0)}{\sum_i \mathbb{1}(d_i=0)}\hat{f}_{d=0,t=0}(x)$; (ii) JTPA: Offered & Not Enrolled: the scaled density estimate $\frac{\sum_i \mathbb{1}(t_i=0,d_i=1)}{\sum_i \mathbb{1}(d_i=1)}\hat{f}_{d=1,t=0}(x)$. Point estimates are obtained by using local polynomial regression with order 2, and robust confidence bands are obtained with local polynomial of order 3. Bandwidths are chosen by minimizing integrated mean squared errors. All estimates are obtained using companion R (and Stata) package described in Cattaneo et al. (2021b).

observed outcome variable, x_i, for compliers, which is denoted by f_c. This parameter is important for understanding the overall characteristics of compliers, and how different it is from the populations. The other two parameters are distributions of the potential outcomes, $x_i(0)$ and $x_i(1)$, for compliers, since the difference thereof reveals the effect of treatment for this subgroup. They are denoted by $f_{c,0}$ and $f_{c,1}$, respectively. The three density functions can also be estimated using our proposed local polynomial density estimator $\hat{f}(x)$ with, respectively, the following weights:

$$w_i^c = \frac{1}{\mathbb{P}[t_i(1) > t_i(0)]} \cdot \left(1 - \frac{t_i(1-d_i)}{\mathbb{P}[d_i=0|z_i]} - \frac{(1-t_i)d_i}{\mathbb{P}[d_i=1|z_i]}\right),$$
$$w_i^{c,0} = \frac{1}{\mathbb{P}[t_i(1) > t_i(0)]} \cdot (1-t_i) \cdot \frac{1-d_i-\mathbb{P}[d_i=0|z_i]}{\mathbb{P}[d_i=0|z_i]\mathbb{P}[d_i=1|z_i]},$$
$$w_i^{c,1} = \frac{1}{\mathbb{P}[t_i(1) > t_i(0)]} \cdot t_i \cdot \frac{d_i-\mathbb{P}[d_i=1|z_i]}{\mathbb{P}[d_i=0|z_i]\mathbb{P}[d_i=1|z_i]}.$$

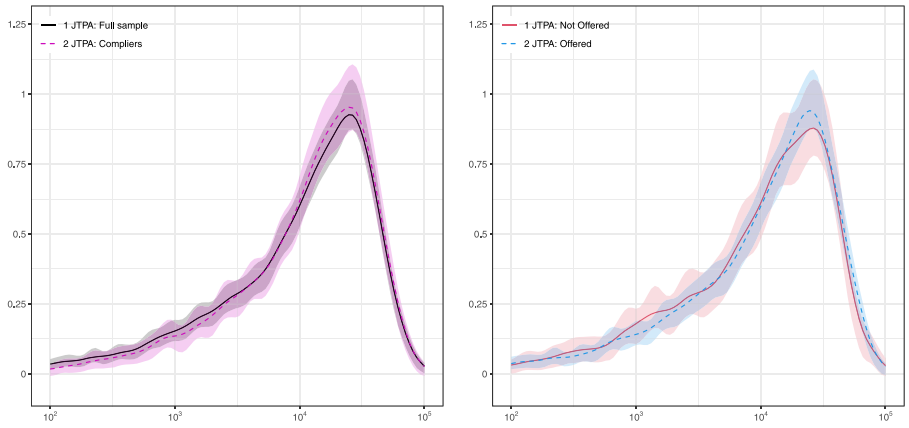
Here, the weights need to be estimated in practice, unless precise knowledge about the treatment assignment mechanism is available. As mentioned previously, our results allow for estimated weights such as those obtained by fitting a flexible Logit or Probit model to approximate the propensity score $\mathbb{P}[d_i = 1|z_i]$ so long as they converge sufficiently fast to their population counterparts.

6.2.1. Empirical illustration

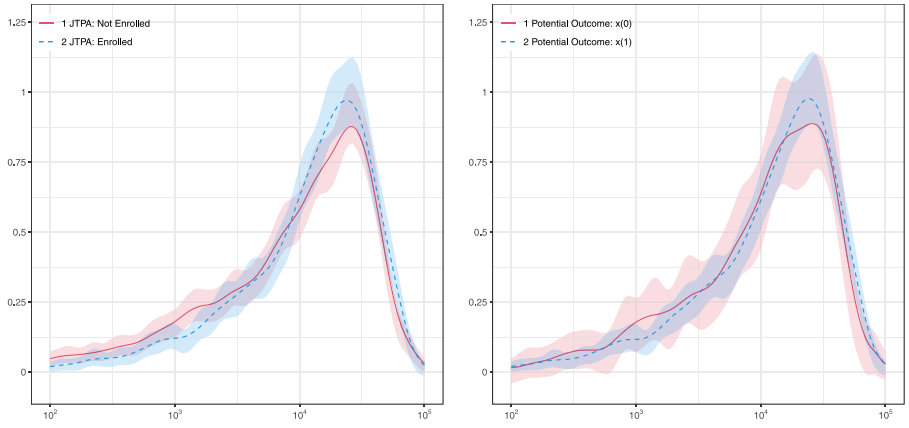
The JTPA is a large publicly funded job training program targeting at individuals who are economically disadvantaged and/or facing significant barriers to employment. Individuals were randomly offered JTPA training, the treatment take-up, however, was only about 67% among those who were offered. Therefore the JTPA offer provides valid instrument to study the impact of the job training program. We continue to use the same data as Abadie et al. (2002), who analyzed quantile treatment effects on earning distributions.

Besides the main outcome variable and covariates already introduced in Section 6.1, also available are the treatment take-up (JTPA enrollment) and the instrument (JTPA Offer). See Table 1 for summary statistics for the full sample and separately for subgroups. As the JTPA offers were randomly assigned, it is possible to estimate the intent-to-treat effect by mean comparison. Indeed, individuals who are offered JTPA services earned, on average, $1130 more than those not offered. On the other hand, due to imperfect compliance, it is in general not possible to estimate the effect of job training (i.e. the effect of JTPA enrollment), unless one is willing to impose strong assumptions such as constant treatment effect.

We first implement the IV specification test, which is straightforward using our density estimator $\hat{f}(x)$. We plot the two estimated (rescaled) densities in Fig. 3. A simple eyeball test suggests no evidence against instrumental variable validity. A formal hypothesis test, justified using our theoretical results, confirms this finding.



(a) Marginal Distributions (b) Marginal Distributions



(c) Marginal Distributions (d) Potential Outcome Distributions

Fig. 4. Earning distributions, JTPA. *Notes*: (a) earning distributions in the full sample and for compliers; (b) earning distributions by JTPA offer; (c) earning distributions by JTPA enrollment; (d) distributions of potential outcomes for compliers. Point estimates are obtained by using local polynomial regression with order 2, and robust confidence bands are obtained with local polynomial of order 3. Bandwidths are chosen by minimizing integrated mean squared errors. All estimates are obtained using companion `R` (and `Stata`) package described in Cattaneo et al. (2021b).

Second, we estimate the density of the potential outcomes for compliers. In panel (a) of Fig. 4, we plot earning distributions for the full sample and that for the compliers, where the second is estimated using the weights w_i^c, introduced earlier. The two distributions seem quite similar, while compliers tend to have higher mean and thinner left tail in the earning distribution. Next we consider the intent-to-treat effect, as the difference in earning distributions for subgroups with and without JTPA offer (a.k.a. the reduced form estimate in the 2SLS context). This is given in panel (b) of Fig. 4. The effect is significant, albeit not very large. We also plot earning distributions for individuals enrolled (and not) in JTPA in panel (c). Not surprisingly, the difference is much larger. Simple mean comparison implies that enrolling in JTPA is associated with $2083 more income.

Unfortunately, neither panel (b) nor (c) reveals information on distribution of potential outcomes. To see the reason, note that in panel (b) earning distributions are estimated according to treatment assignment, but potential outcomes are defined according to treatment takeup. And panel (c) does not give potential outcome distributions since treatment takeup is not randomly assigned. In panel (d) of Fig. 4, we use weighting schemes $w_i^{c,0}$ and $w_i^{c,1}$ to construct potential earning distributions for compliers, which estimates the identified distributional treatment effect in this IV setting. Indeed, treatment effect on compliers is larger than the intent-to-treat effect, but is smaller than that in panel (c). The result is compatible with the fact that JTPA has positive and nontrivial effect on earning. Moreover, it demonstrates the presence of self-selection: those who participated in JTPA on average would benefit the most, followed by compliers who are regarded as "on the margin of indifference".

7. Conclusion

We introduced a new class of local regression distribution estimators, which can be used to construct distribution, density, and higher-order derivatives estimators. We established valid large sample distributional approximations, both pointwise and uniform over their support. Pointwise on the evaluation point, we characterized a minimum distance implementation based on redundant regressors leading to asymptotic efficiency improvements, and gave precise results in terms of (tight) lower bounds for interior points. Uniformly over the evaluation points, we obtained valid linearizations and strong approximations, and constructed confidence bands. Finally, we discussed several extensions of our work.

Although beyond the scope of this paper, it would be useful to generalize our results to the case of multivariate regressors $x_i \in \mathbb{R}^d$. Boundary adaptation is substantially more difficult in multiple dimensions, and hence our proposed methods are potentially very useful in such setting. In addition, multidimensional density estimation can be used to construct new conditional distribution, density and higher derivative estimators in a straightforward way. These new estimators would be useful in several areas of economics, including for instance estimation of auction models.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2021.01.006.

References

Abadie, A., 2003. Semiparametric instrumental variable estimation of treatment response models. J. Econometrics 113 (2), 231–263.
Abadie, A., Angrist, J., Imbens, G., 2002. Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. Econometrica 70 (1), 91–117.
Abadie, A., Cattaneo, M.D., 2018. Econometric methods for program evaluation. Annu. Rev. Econ. 10, 465–503.
Belloni, A., Chernozhukov, V., Chetverikov, D., Fernandez-Val, I., 2019. Conditional quantile processes based on series or many regressors. J. Econometrics 213 (1), 4–29.
Belloni, A., Chernozhukov, V., Chetverikov, D., Kato, K., 2015. Some new asymptotic theory for least squares series: Pointwise and uniform results. J. Econometrics 186 (2), 345–366.
Calonico, S., Cattaneo, M.D., Farrell, M.H., 2018. On the effect of bias estimation on coverage accuracy in nonparametric inference. J. Amer. Statist. Assoc. 113 (522), 767–779.
Calonico, S., Cattaneo, M.D., Farrell, M.H., 2020. Coverage error optimal confidence intervals for local polynomial regression. arXiv:1808.01398.
Cattaneo, M.D., Crump, R.K., Farrell, M.H., Feng, Y., 2021a. On binscatter. arXiv:1902.09608.
Cattaneo, M.D., Farrell, M.H., Feng, Y., 2020a. Large sample properties of partitioning-based estimators. Ann. Statist. 48 (3), 1718–1741.
Cattaneo, M.D., Jansson, M., Ma, X., 2018. Manipulation testing based on density discontinuity. Stata J. 18 (1), 234–261.
Cattaneo, M.D., Jansson, M., Ma, X., 2020b. Simple local polynomial density estimators. J. Amer. Statist. Assoc. 115 (531), 1449–1455.
Cattaneo, M.D., Jansson, M., Ma, X., 2021b. `lpdensity`: Local polynomial density estimation and inference. J. Stat. Softw. forthcoming.
Cheng, G., Chen, Y.-C., 2019. Nonparametric inference via bootstrapping the debiased estimator. Electron. J. Stat. 13 (1), 2194–2256.
Cheng, M.-Y., Fan, J., Marron, J.S., 1997. On automatic boundary corrections. Ann. Statist. 25 (4), 1691–1708.
Chernozhukov, V., Chetverikov, D., Kato, K., 2014a. Anti-concentration and honest adaptive confidence bands. Ann. Statist. 42 (5), 1787–1818.
Chernozhukov, V., Chetverikov, D., Kato, K., 2014b. Gaussian approximation of suprema of empirical processes. Ann. Statist. 42 (4), 1564–1597.
Chernozhukov, V., Escanciano, J.C., Ichimura, H., Newey, W.K., Robins, J.M., 2020. Locally robust semiparametric estimation. arXiv:1608.00033.
Chernozhukov, V., Fernandez-Val, I., Melly, B., 2013. Inference on counterfactual distributions. Econometrica 81 (6), 2205–2268.
DiNardo, J., Fortin, N.M., Lemieux, T., 1996. Labor market institutions and the distribution of wages, 1973-1992: A semiparametric approach. Econometrica 64 (5), 1001–1044.
Fan, J., Gijbels, I., 1996. Local Polynomial Modelling and Its Applications. Chapman & Hall/CRC, New York.
Giné, E., Koltchinskii, V., Sakhanenko, L., 2004. Kernel density estimators: Convergence in distribution for weighted sup-norms. Probab. Theory Related Fields 130 (2), 167–198.
Giné, E., Nickl, R., 2010. Confidence bands in density estimation. Ann. Statist. 38 (2), 1122–1170.
Granovsky, B.L., Müller, H.-G., 1991. Optimizing kernel methods: A unifying variational principle. Int. Stat. Rev. (Rev. Int. Stat.) 59 (3), 373–388.
Hausman, J.A., Newey, W.K., 1995. Nonparametric estimation of exact consumers surplus and deadweight loss. Econometrica 63 (6), 1445–1476.
Ichimura, H., Newey, W.K., 2020. The influence function of semiparametric estimators. arXiv:1508.01378.
Ichimura, H., Todd, P.E., 2007. Implementing nonparametric and semiparametric estimators. In: Heckman, J.J., Leamer, E.E. (Eds.), Handbook of Econometrics, Volume 6B. Elsevier Science B.V., New York, pp. 5369–5468.
Imbens, G.W., Rubin, D.B., 1997. Estimating outcome distributions for compliers in instrumental variables models. Rev. Econom. Stud. 64 (4), 555–574.
Imbens, G.W., Rubin, D.B., 2015. Causal Inference in Statistics, Social, and Biomedical Sciences. Cambridge University Press, New York.
Karunamuni, R.J., Zhang, S., 2008. Some improvements on a boundary corrected kernel density estimator. Statist. Probab. Lett. 78 (5), 499–507.
Kitagawa, T., 2015. A test for instrument validity. Econometrica 83 (5), 2043–2063.
Newey, W.K., 1994a. The asymptotic variance of semiparametric estimators. Econometrica 62 (6), 1349–1382.
Newey, W.K., 1994b. Kernel estimation of partial means and a general variance estimator. Econometric Theory 10 (2), 233–253.
Newey, W.K., Hsieh, F., Robins, J.M., 2004. Twicing kernels and a small bias property of semiparametric estimators. Econometrica 72 (3), 947–962.
Newey, W.K., McFadden, D.L., 1994. Large sample estimation and hypothesis testing. In: Engle, R.F., McFadden, D.L. (Eds.), Handbook of Econometrics, Volume 5. Elsevier Science B.V., New York, pp. 2111–2245.
Newey, W.K., Ruud, P.A., 2005. Density weighted linear least squares. In: Andrews, D., Stock, J. (Eds.), Identification and Inference in Econometric Models: Essays in Honor of Thomas Rothenberg. Cambridge University Press, Cambridge, pp. 554–573.
Newey, W.K., Stoker, T.M., 1993. Efficiency of weighted average derivative estimators and index models. Econometrica 61 (5), 1199–1223.
Rio, E., 1994. Local invariance principles and their application to density estimation. Probab. Theory Related Fields 98 (1), 21–45.
Robins, J.M., Hsieh, F., Newey, W.K., 1995. Semiparametric efficient estimation of a conditional density with missing or mismeasured covariates. J. R. Stat. Soc. Ser. B Stat. Methodol. 57 (2), 409–424.
Zaitsev, A., 2013. The accuracy of strong Gaussian approximation for sums of independent random vectors. Russian Math. Surveys 68 (4), 721–761.
Zhang, S., Karunamuni, R.J., 1998. On kernel density estimation near endpoints. J. Statist. Plann. Inference 70 (1), 301–316.