Local Regression Distribution Estimators*

Matias D. Cattaneo† Michael Jansson‡ Xinwei Ma§

December 4, 2019

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 more general L_2 least squares 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 provided.

Keywords: distribution and density estimation, local polynomial methods, uniform approximation, efficiency, optimal kernel, program evaluation.

*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.

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

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

§Department of Economics, UC San Diego.

1 Introduction

Kernel-based non-parametric estimation of distribution and density functions, as well as higher-order derivatives thereof, play an important role in econometrics. These non-parametric estimators often feature both as the main object of interest and as preliminary ingredients in multi-step semi-parametric procedures. Whitney Newey's path-breaking contributions to non-/semi-parametric econometrics employing kernel smoothing include Newey and Stoker (1993), Newey (1994a), Newey (1994b), Hausman and Newey (1995), Robins, Hsieh and Newey (1995), Newey, Hsieh and Robins (2004), Newey and Ruud (2005), Ichimura and Newey (2019), and Chernozhukov, Escanciano, Ichimura, Newey and Robins (2019). See also Newey and McFadden (1994) for an outstanding handbook chapter reviewing those methods. Our paper hopes to honor Whitney's 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-/semi-parametric 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 $\mathsf{x} \in \mathcal{X}$ is implemented via a kernel function and a bandwidth parameter. The local functional form approximation is done using a finite-dimension basis function, not necessarily of polynomial form. 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 the localization (or evaluation) point $\mathsf{x} \in \mathcal{X}$. If only a polynomial basis is used, then the resulting estimator reduces to the one recently proposed in Cattaneo, Jansson and Ma (2019c).

We present two main large sample distributional results for the local regression distribution estimators. First, we establish a pointwise (in $\mathsf{x} \in \mathcal{X}$) 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, Jansson and Ma (2019c) by allowing for arbitrary local basis functions, which is particularly useful for efficiency considerations, as we will discuss below.

Because the limiting asymptotic variance of the local polynomial density estimator is non-diagonal, we show that more efficient estimators can be constructed via a minimum distance approach based on "redundant" regressors. In particular, for estimation of the density and its derivatives, we show that our minimum distance construction can be used to recover 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 non-parametric kernel-based density estimation, but also a new class of more efficient boundary adaptive density estimators for practice.

Our second main large sample distributional result 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 $\mathsf{x} \in \mathcal{I}$, relying on a "coupling" result in Giné, Koltchinskii and Sakhanenko (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 allow us to deduce a distributional approximation for many functionals of the Studentized statistic, including its supremum, following ideas in Chernozhukov, Chetverikov and Kato (2014b). For further discussion and references on strong approximations and their applications to non-/semi-parametric statistics and econometrics see Chernozhukov, Chetverikov and Kato (2014a), Belloni, Chernozhukov, Chetverikov and Kato (2015), Belloni, Chernozhukov, Chetverikov and Fernandez-Val (2019), and Cattaneo, Farrell and Feng (2019b), Cattaneo, Crump, Farrell and Feng (2019a), 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, in one-sample and two-sample problems. Other applications of our results, not discussed here to conserve space, include parametric specification and shape restriction testing; see Cattaneo, Crump, Farrell and Feng (2019a) for an example using strong approximation methods in a different non-parametric setting. As a by-product, we also establish a linear approximation to the boundary

adaptive Studentized statistic, uniformly over $\mathsf{x} \in \mathcal{I}$, which gives uniform convergence rates and can be used for further theoretical developments.

In addition to our main large sample results for local regression distribution and related estimators, we briefly discuss two extensions. 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 empirically relevant as illustrated in Section 6. Second, we present and study an alternative class of estimators that employ a non-random L_2 loss function, 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 how these alternative estimators can be implemented to achieve maximum asymptotic efficiency in estimating the density function and its derivatives.

Finally, we illustrate our methods with two applications in program evaluation (see Abadie and Cattaneo, 2018, for a review). First, we discuss counterfactual density analysis following DiNardo, Fortin and Lemieux (1996); see also Chernozhukov, Fernandez-Val and Melly (2013) for closely 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 non-parametric density estimation to causal inference and program evaluation. In all these applications, we develop formal estimation and inference methods based on non-parametric 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, Angrist and Imbens (2002), corresponding to the Job Training Partnership Act (JTPA).

The paper proceeds as follows. Section 2 introduces the class of local regression distribution estimators. Section 3 establishes a pointwise distributional approximation, along with consistency of a standard errors 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 and related procedures. Section 5 discusses two extensions of our methodology, while Section 6 illustrate our new methods with two distinct program evaluation applications: counterfactual densities estimation (Section

6.1) and IV specification and heterogeneity density analysis (Section 6.2). Section 7 concludes. A supplemental appendix includes all proofs of our theoretical results as well as other technical, methodological and numerical results. Finally, a software package implementing the main results in this paper for `Stata` and `R` is discussed in Cattaneo, Jansson and Ma (2019d).

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 non-parametric estimators of $F(\cdot)$, and derivatives thereof, both pointwise at $\mathsf{x} \in \mathcal{X}$ and uniformly over $\mathcal{I} \subseteq \mathcal{X}$.

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:

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

where $R_h(\cdot)$ is a known local basis function, which can depend on a tuning parameter h, and θ is a parameter to be estimated. The generic formulation (1) is motivated in part by the important special case where $F(\cdot)$ is sufficiently smooth, in which case

$$F(x) \approx F(\mathsf{x}) + f(\mathsf{x})(x-\mathsf{x}) + \ldots + f^{(p-1)}(\mathsf{x})\frac{1}{p!}(x-\mathsf{x})^p \qquad \text{for } x \approx \mathsf{x}, \tag{2}$$

and $f^{(s)}(\mathsf{x}) = d^s f(x)/dx^s|_{x=\mathsf{x}}$ are higher-order density derivatives. Of course, the approximation (2) is of the form (1) with $R_h(u) = (1, u, \cdots, u^p/p!)'$, and hence $\theta(\mathsf{x}) = (F(\mathsf{x}), f(\mathsf{x}), \cdots, f^{(p-1)}(\mathsf{x}))'$. But, as further discussed below, other choices of $R_h(\cdot)$ and/or $\theta(\cdot)$ can be attractive, and as a consequence we take (1) as the starting point for our analysis.

As an estimator of $\theta(\mathsf{x})$ in (1), we consider the local regression estimator

$$\hat{\theta}(\mathsf{x}) = \operatorname*{argmin}_{\theta} \sum_{i=1}^{n} W_i(\hat{F}_i - R_i'\theta)^2, \tag{3}$$

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

$$\hat{F}_i = \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}\left(x_j \leq x_i\right) \tag{4}$$

is the empirical distribution function evaluated at x_i. The formulation (3) allows for a basis function $R_h(\cdot)$ that could explicitly depend on the tuning parameter h. The estimator $\hat{\theta}(\mathsf{x})$ is a generalization of the density estimator introduced in Cattaneo, Jansson and Ma (2019c) where it was assumed that $R_h(u) = (1, u, u^2/2, \cdots, u^p/p!)'$.

While not discussed herein to conserve space, the formulation (3) is general enough to allow for some interesting extensions/refinements in the definition of the local regression distribution estimators. For example, the local basis $R_h(u)$ can incorporate specific restrictions, such as continuity or lack thereof, on the distribution function, density function or higher-order derivatives at the evaluation point x. As a second example, shape constrains such as positivity or monotonicity can be incorporated by means of specific restrictions on the parameter space of θ. Finally, a third natural extension of (3) could allow for a non-identity link function, leading to a non-linear least squares formulation. We plan to investigate these and other extensions of our work in future research.

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 supplemental appendix 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) There exists a fixed and locally bounded function $R(\cdot)$, and a positive-definite diagonal matrix Υ_h for each $h > 0$, such that $\Upsilon_h R_h(u) = R(u/h)$.

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

$$\Gamma_{h,\mathsf{x}} = \int_{\mathcal{X}_h} R(u)R(u)'K(u)f(\mathsf{x}+hu)du.$$

(v) The minimum eigenvalue of $h^{-1}\Sigma_{h,\mathsf{x}}$ is bounded away from zero, where

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

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 near or at boundary points. Part (ii) requires standard restrictions on kernel-based estimators, which allows for all standard (compact supported) second-order kernel functions. Part (iii) requires that the local basis $R_h(\cdot)$ can be stabilized by a suitable normalization. Finally, parts (iv) and (v) give assumptions on various (nonrandom) matrices which will feature in the asymptotic distribution.

The error of the approximation in (1) depends on the choice of $R_h(\cdot)$ and θ, and is quantified by $\varrho(h)$ (we suppress the dependence on the evaluation point x). 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, 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.

3.2 Asymptotic Normality

We show that, under regularity conditions and if h vanishes at a suitable rate, 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

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

This result implies, in particular, 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 in general, but such dependence is suppressed in this section for simplicity.

Assuming $\hat{\Gamma}$ is invertible (with probability approaching one), we have

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

and result (5) follows if S is asymptotically mean-zero Gaussian in the sense that

$$\mathbb{V}[S]^{-1/2}S \rightsquigarrow \mathcal{N}(0,I) \tag{6}$$

and if the variance estimator $\hat{\Sigma}$ is consistent in the sense that $\mathbb{V}[S]^{-1}(\hat{\Sigma}-\mathbb{V}[S]) \to_{\mathbb{P}} 0$. Up to a leave-in bias term and a smoothing bias term, the statistic S can be written as

$$S=\frac{1}{n(n-1)}\sum_{i,j=1,i\neq j}^{n} W_j R_j\Big(\mathbb{1}\left(x_i \le x_j\right)-F(x_j)\Big), \tag{7}$$

that is, S is approximately a second-order U-statistic, so (6) should follow from a central limit theorem for (n-varying) U-statistics under suitable regularity conditions, including conditions ensuring that the approximation error in (1) is negligible. Moreover, the projection theorem for (varying) U-statistics suggests that if the approximation error in (1) is negligible, then

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

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

The following theorem formalizes the above intuition, and gives precise sufficient conditions. Here, and elsewhere in the sequel, we are considering asymptotics as $h \to 0$ and $n \to \infty$.

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 implies that the standard $(1-\alpha)\%$ confidence interval:

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

$$\mathfrak{q}_a = \inf\left\{ u \in \mathbb{R} : \mathbb{P}\left[\mathcal{N}(0,1) \leq u\right] \geq \alpha \right\},$$

is asymptotically valid, that is,

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

The above confidence interval is valid asymptotically for each evaluation point x, which is reflected by the notation $\mathrm{CI}_\alpha(\mathsf{x})$. Later, we will also develop 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), standard local polynomial regression estimators also have a limiting asymptotic variance of the "sandwich" form $A^{-1}BA^{-1}$, where

$$A \propto \int R(u)R(u)'K(u)du \qquad \text{and} \qquad B \propto \int R(u)R(u)'K(u)^2du,$$

which implies that setting $K(\cdot)$ to be the uniform kernel minimizes their asymptotic variance, at least in the sense that $A^{-1}BA^{-1} \geq B^{-1}$. See Granovsky and Müller (1991) for a more general discussion on the optimality of the uniform kernel for kernel-based estimation.

In the case of local regression distribution estimators, employing the uniform kernel does not exhaust the potential efficiency gains of choosing a kernel function to minimize their asymptotic variance. Heuristically, unlike the case of the asymptotic variance of local polynomial estimators, which is effectively of the form $A \propto B$ for appropriate choice of kernel function, the local regression distribution estimators exhibit a more complex and uneven asymptotic variance formula due to their construction. For example, in the case of local polynomial density and higher-order estimation (i.e., setting $R(u)$ polynomial of order $p > 1$) the asymptotic variance matrix of $\hat{\theta}$, excluding the intercept term for simplicity, takes the form $A^{-1}CA^{-1}$ with

$$C \propto \int\int \min\{u, v\} R(u) R(v)' K(u) K(v) du dv,$$

which implies that (the corresponding submatrix of) A is not proportional to (the corresponding submatrix of) C even when the kernel function is uniform. This discussion excludes the intercept because that ($\sqrt{n}$-consistent) estimator corresponds to the cumulative distribution function, and has a quite different asymptotic variance. The above heuristics apply to the general case where the local basis function $R(\cdot)$ needs not to be of polynomial form. 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 this idea. To motivate our approach, notice that in many cases it is possible to specify $R_h(\cdot)$ in such a way that θ can be taken to be of the form $\theta = (\theta_1', \theta_2')'$, 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 $R_{2,h}(\cdot)$ from $R_h(\cdot) = (R_{1,h}(\cdot)', R_{2,h}(\cdot)')'$ is given by

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

while an optimal minimum distance estimator of θ_1 is given by

$$\hat{\theta}_{\mathtt{MD},1} = \underset{\theta_1}{\operatorname{argmin}} \begin{pmatrix} \hat{\theta}_1 - \theta_1 \\ \hat{\theta}_2 \end{pmatrix}' \hat{\Omega}^{-1} \begin{pmatrix} \hat{\theta}_1 - \theta_1 \\ \hat{\theta}_2 \end{pmatrix} = \hat{\theta}_1 - \hat{\Omega}_{12}\hat{\Omega}_{22}^{-1}\hat{\theta}_2. \tag{8}$$

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

$$\hat{\Omega}_{11}^{-1/2}(\hat{\theta}_1 - \theta_1) \rightsquigarrow \mathcal{N}(0, I),$$

$$\hat{\Omega}_{\mathtt{R},11}^{-1/2}(\hat{\theta}_{\mathtt{R},1} - \theta_1) \rightsquigarrow \mathcal{N}(0, I), \qquad \hat{\Omega}_{\mathtt{R},11} = \hat{\Gamma}_{11}^{-1}\hat{\Sigma}_{11}\hat{\Gamma}_{11}^{-1}$$

and

$$\hat{\Omega}_{\mathtt{MD},11}^{-1/2}(\hat{\theta}_{\mathtt{MD},1} - \theta_1) \rightsquigarrow \mathcal{N}(0, I), \qquad \hat{\Omega}_{\mathtt{MD},11} = \hat{\Omega}_{11} - \hat{\Omega}_{12}\hat{\Omega}_{22}^{-1}\hat{\Omega}_{21},$$

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

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

and

$$\hat{\theta}_{\mathtt{R},1} = \underset{\theta_1}{\operatorname{argmin}} \begin{pmatrix} \hat{\theta}_1 - \theta_1 \\ \hat{\theta}_2 \end{pmatrix}' \hat{\Omega} \begin{pmatrix} \hat{\theta}_1 - \theta_1 \\ \hat{\theta}_2 \end{pmatrix},$$

each estimator admits a minimum distance interpretation, but only $\hat{\theta}_{\mathtt{MD},1}$ can be interpreted as an optimal minimum distance estimator based on $\hat{\theta}$.

As a consequence, we investigate whether an appropriately implemented $\hat{\theta}_{\mathtt{MD},1}$ can lead to asymptotic efficiency gains relative to $\hat{\theta}_1$ and $\hat{\theta}_{\mathtt{R},1}$. 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, (2) is satisfied

and a natural choice of $R_h(\cdot)$ is

$$R_h(u) = \Big(R_{1,h}(u)', R_{2,h}(u)'\Big)' = \Big(1, P(u)', Q(u)'\Big)' = R(u), \tag{9}$$

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

$$\theta = \Big(\underbrace{F(\mathsf{x})}_{\text{intercept}}, \underbrace{f(\mathsf{x}), \cdots, f^{(p-1)}(\mathsf{x})}_{\text{slope, } R_{1,h}(\cdot)=P(\cdot)}, \underbrace{0, \cdots, 0}_{\text{redundant, } R_{2,h}(\cdot)=Q(\cdot)} \Big)', \tag{10}$$

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

With (9) and (10), we define the minimum distance density estimator as $\hat{f}_{\mathtt{MD}}(\mathsf{x}) = e_1'\hat{\theta}_{\mathtt{MD},1}$, where e_ℓ is the $(\ell+1)$th unit vector. Similarly, we have $\hat{f}(\mathsf{x}) = e_1'\hat{\theta}_1$ and $\hat{f}_{\mathtt{R}}(\mathsf{x}) = e_1'\hat{\theta}_{\mathtt{R},1}$. 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: $Q(u) = (u^{p+1}/(p+1)!, \cdots, u^{p+q}/(p+q)!)$. By redefining the parameters as $\theta = (F(\mathsf{x}),\ f(\mathsf{x}),\ \cdots,\ f^{(p+q-1)}(\mathsf{x}))'$, the smoothing error will be of order $\varrho(h) = o(h^{p+q})$. Notice that, in this case, $\hat{f}(\mathsf{x})$ and $\hat{f}_{\mathtt{R}}(\mathsf{x})$ correspond to the density estimator introduced in Cattaneo, Jansson and Ma (2019c) implemented with $R_h(u) = (1, u, \cdots, u^{p+q}/(p+q)!)'$ and $R(u) = (1, u, \cdots, u^p/p!)'$, 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 $\hat{f}_{\mathtt{R}}(\mathsf{x})$ are (weakly) asymptotically inefficient relative to $\hat{f}_{\mathtt{MD}}(\mathsf{x})$ 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 the natural minimal regularity conditions (e.g., lack of perfect collinearity between P and Q) the asymptotic variance of the minimum distance density estimator (and more generally, its ℓ-th derivative) is

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

where

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

Because we consider an interior evaluation point x where $F(\cdot)$ is assumed to be smooth, it is possible to simplify the expression of Γ_h and Σ_h. In particular, it suffices to consider the following:

$$\Gamma_h = f(\mathsf{x})\int_{-1}^{1} R(u)R(u)'K(u)du, \quad \Sigma_h = hf(\mathsf{x})^3\int_{-1}^{1}\int_{-1}^{1} R(u)R(v)'K(u)K(v)\min\{u,v\}dudv.$$

See the SA for omitted details.

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

$$\sup_{Q\in\mathcal{Q}} \frac{\left[\int_{-1}^{1}\int_{-1}^{1} P_\ell(u)Q(v)\min\{u,v\}K(u)K(v)dudv\right]^2}{\int_{-1}^{1}\int_{-1}^{1} Q(u)Q(v)\min\{u,v\}K(u)K(v)dudv} \tag{11}$$

where

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

with $P_\ell(u) = e_\ell'\Big(\int_{-1}^{1} P(u)P(u)'K(u)du\Big)^{-1}P(u)$ and $\ell = 1, 2, \ldots, p$. 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 estimator among all possible choices of redundant regressors Q. In addition, it shows that this lower bound can be (approximately) achieved by setting the redundant regressor $Q(\cdot)$ to include certain higher order polynomial functions.

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*

$$\inf_{Q \in \mathcal{Q}} \mathsf{AsyVar}[\hat{f}^{(\ell)}_{\mathtt{MD}}(\mathsf{x})] \geq \nu_\ell, \qquad \nu_\ell = e_\ell' \left(\int_{-1}^{1} \dot{P}(u)\dot{P}(u)' du \right)^{-1} e_\ell.$$

Furthermore, by direct calculation for each $p = 1, 2, 3, \cdots$, $\lim_{j\to\infty} \mathsf{AsyVar}[\hat{f}^{(\ell)}_{\mathtt{MD},j}(\mathsf{x})] = \nu_\ell$, *where the minimum distance estimator* $\hat{f}^{(\ell)}_{\mathtt{MD},j}(\mathsf{x}) = e_\ell' \hat{\theta}_{\mathtt{MD},j}$ *is constructed with*

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

or

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

and $K(\cdot)$ *being the uniform kernel.*

The first part, and the main result in 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 provides an alternative way of characterizing minimum variance results for non-parametric kernel-based estimators of the density function and its derivatives.

The second part of the theorem gives a simple recipe for implementation. Specifically, it proposes a specific choice of $Q(\cdot)$ so that the corresponding minimum distance estimator approximately achieves the variance bound for j large enough. We assume $Q(\cdot)$ is orthogonal to $P(\cdot)$ for theoretical convenience: to implement this estimator only a local polynomial regression of the empirical distribution function on a constant, the polynomial basis $P(\cdot)$, and one additional regressor u^{2j+1} or u^{2j+2} (depending on the choice of ℓ) needs to be implemented, to then apply (8) with the corresponding estimated variance-covariance matrix.

In Figure 1, we consider the local linear/quadratic density estimator ($\ell = 1$) 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 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 kernel density estimators employing second order kernels. In other words, the asymptotic variance of our proposed minimum distance local polynomial density estimator converges to the optimal asymptotic variance as $j \to \infty$.

4 Uniform Distribution Theory

The distributional result presented in Theorem 1 is valid pointwise for $\mathsf{x} \in \mathcal{X}$. In this section we develop an 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), and where c is a conformable vector and $\mathcal{I} \subseteq \mathcal{X}$ is some prespecified region. This stochastic process is not tight, and hence does not converge in distribution. Our approximation proceeds in two steps. First, for an positive (vanishing) r_n sequence, we establish an uniform "linearization" of the process $T(\cdot)$ of the form:

$$\sup_{\mathsf{x}\in\mathcal{I}} |T(\mathsf{x}) - \mathfrak{T}(\mathsf{x})| = O_{\mathbb{P}}(r_n), \tag{12}$$

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

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

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 (12), we develop new uniform approximations with sharp rate of convergence, which may be of independent interest in semi-parametric estimation and inference settings. See the SA for more details.

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

structure, such that

$$\sup_{\mathsf{x}\in\mathcal{I}}\left|\tilde{\mathfrak{T}}(\mathsf{x})-\mathfrak{B}(\mathsf{x})\right|=O_{\mathbb{P}}(r_n). \tag{13}$$

This type of strong approximation result, when established with suitably fast rate $r_n \to 0$, 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 (13), we employ a result originally established by Rio (1994), and later extended in Giné, Koltchinskii and Sakhanenko (2004); see also Giné and Nickl (2010, Proof of Proposition 5).

In this section we consider a fixed the 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 the previous section. 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 (e.g., 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 exists some $\delta>0$ and compactly supported kernel functions $K^{\dagger}(\cdot)$ and $\{K^{\ddagger,d}(\cdot)\}_{d\leq\delta}$, such that (ii.1) $\sup_{u\in\mathbb{R}}|K^{\dagger}(u)|+\sup_{d\leq\delta,u\in\mathbb{R}}|K^{\ddagger,d}(u)|<\infty$, (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 $|u-v|\leq\delta$,

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

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

(iv) For all h sufficiently small, $\Sigma_{h,\mathsf{x}}$ is positive definite, and 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 $\mathcal{I}$. This assumption can be easily justified if the density function belongs to a Hölder ball. 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

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

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. *(12) holds with*
$$r_n=\sqrt{\frac{n}{h}}\sup_{\mathsf{x}\in\mathcal{I}}\varrho(h,\mathsf{x})+\frac{\log n}{\sqrt{nh^2}}.$$
2. *On a possibly enlarged probability space, there exists a copy $\tilde{\mathfrak{T}}(\cdot)$ of $\mathfrak{T}(\cdot)$, and a tight and*

centered Gaussian process, $\{\mathfrak{B}(\mathsf{x}), \mathsf{x} \in \mathcal{I}\}$*, defined with the same covariance as* $\mathfrak{T}(\cdot)$*, such that* (13) *holds with*

$$r_n = \frac{\log n}{\sqrt{nh}}.$$

The first part of this theorem gives conditions such that the feasible t-statistic process $T(\cdot)$ is well approximated by the infeasible (linear) t-statistic 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 distribution regression estimators, the shape of the implied kernel automatically changes for different evaluation points depending on whether they are interior or boundary points. Nevertheless, we are able to obtain a valid strong approximation result, which is reported in the second part of the Theorem.

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, Jansson and Ma (2019c), where the basis function takes the form: $R_h(u) = R(u) = (1, u, u^2/2, \cdots, u^p/p!)$ 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 $c = e_1$ is the second unit vector. 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, Chetverikov and Kato (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 $W_i(\mathsf{x}) = K((x_i - \mathsf{x})/h)/h$, $R_i(\mathsf{x}) = R_h(x_i - \mathsf{x})$, 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

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

The following lemma characterizes the approximation error from replacing the infeasible process $\mathfrak{B}(\cdot)$ by a plug-in feasible version. (Conditioning on data means conditioning on $x_1, x_2, \cdots, x_n$.)

Lemma 1 *Suppose Assumptions 1 and 2 hold, and that $h \to 0$ and $nh^2/\log n \to \infty$. Then, there exist two tight and centered Gaussian processes, $\{\tilde{\mathfrak{B}}(\mathsf{x}), \mathsf{x} \in \mathcal{I}\}$ and $\{\hat{\mathfrak{B}}(\mathsf{x}), \mathsf{x} \in \mathcal{I}\}$, such that (i) $\tilde{\mathfrak{B}}(\cdot)$ has the same distribution as $\mathfrak{B}(\cdot)$ and is independent of the data; (ii) conditional on the data, $\hat{\mathfrak{B}}(\cdot)$ has covariance*

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

and (iii)

$$\mathbb{E}\left[\sup_{\mathsf{x}\in\mathcal{I}}|\tilde{\mathfrak{B}}(\mathsf{x}) - \hat{\mathfrak{B}}(\mathsf{x})| \middle| \text{Data}\right] = O_{\mathbb{P}}\left(\left(\frac{\log^2 n}{\sqrt{nh^3}}\right)^{1/3}\right).$$

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

Theorem 4 (Kolmogorov-Smirnov Distance) *Suppose Assumptions 1 and 2 hold, and that*

$$\frac{n\log n}{h}\sup_{\mathsf{x}\in\mathcal{I}}\varrho(h,\mathsf{x})^2 + \frac{(\log n)^7}{nh^3} = o(1),$$

then

$$\sup_{u\in\mathbb{R}}\left|\mathbb{P}\Big[\sup_{\mathsf{x}\in\mathcal{I}}|T(\mathsf{x})|\leq u\Big]-\mathbb{P}\Big[\sup_{\mathsf{x}\in\mathcal{I}}|\hat{\mathfrak{B}}(\mathsf{x})|\leq u\Big|\text{Data}\Big]\right|=o(1),$$

where $\hat{\mathfrak{B}}(\cdot)$ *is a tight and centered Gaussian process defined in Lemma 1.*

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

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

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,

$$\mathfrak{q}_a=\inf\left\{u\in\mathbb{R}:\mathbb{P}\left[\sup_{\mathsf{x}\in\mathcal{I}}|\hat{\mathfrak{B}}(\mathsf{x})|\leq u\Big|\text{Data}\right]\geq\alpha\right\},$$

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

5 Extensions

We briefly outline two extensions of the main results presented above. 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 least-squares loss function, which has some interesting theoretical properties and may be of interest in some semi-parametric settings.

5.1 Re-weighted Distribution Estimator

Suppose $(x_1,w_1),(x_2,w_2),\cdots,(x_n,w_n)$ 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

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

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 $\frac{d}{d\mathsf{x}}H(\mathsf{x})$ becomes a counterfactual density function, as in DiNardo, Fortin and

Lemieux (1996); see Section 6.1 below. If w_i is set to be a combination of the treatment assignment and treatment status variables, then the resulting density 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 density can be used to identify distributions of compliers, as in Imbens and Rubin (1997), Abadie (2003), and Kitagawa (2015); see Section 6.2 below. 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 $\frac{d}{d\mathsf{x}}H(\mathsf{x})$ and classical weighted averages featuring prominently in econometrics because $\frac{d}{d\mathsf{x}}H(\mathsf{x}) = \mathbb{E}[w_i|x_i = \mathsf{x}]f(\mathsf{x})$, 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. Specifically, we let $\hat{F}_{w,i}(x) = \frac{1}{n}\sum_{j=1}^{n} w_j \mathbb{1}(x_j \leq x_i)$ in place of $\hat{F}_i$ in (4), and investigate the large sample properties of our proposed estimator in (3) when w_i is replaced by $\hat{w}_i = w_i(\hat{\beta})$ with $\hat{\beta}$ a $\sqrt{n}$-consistent estimator and $w_i(\cdot)$ a known function of the data, that is, when estimated weights are used to construct the weighted empirical distribution function $\hat{F}_{w,i}(x)$. All the results reported in the previous sections apply to this extension, which is heavily used in the upcoming applications.

5.2 Local Projection Distribution Estimators

The local regression distribution estimator is obtained from a least squares projection of the empirical distribution function to a local basis, where the projection puts equal weights at all observations. That is, (3) employs an $L_2(\hat{F})$-projection

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

This representation motivates a general class of local projection distribution estimators given by

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

for some distribution function G. We show in the SA that all our theoretical results continue to hold for $\hat{\theta}_G$, provided that the distribution G is absolutely continuous with respect to the Lebesgue measure and the Radon-Nikodym derivative is reasonably smooth.

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 projection 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 (c.f. (7)), or 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), but rather $nh \to \infty$. Similarly, for the strong approximation results we only need to assume $\log n/\sqrt{nh}$ as opposed to $\log n/\sqrt{nh^2}$ (part 1 of Theorem 3).

In addition, the local projection 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 projection 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 support $\mathcal{X}$. In contrast, the local regression distribution estimator is fully boundary adaptive, even in cases where the location of the boundary is unknown. See Cattaneo, Jansson and Ma (2019c) for further discussion for the case of density estimation.

6 Applications

We discuss two applications of our main results in the context of program evaluation (see Abadie and Cattaneo, 2018, and references therein). These applications also employ the first extension discussed previously.

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, Fortin and Lemieux (1996). We focus on density estimation, and we refer

readers to Chernozhukov, Fernandez-Val and Melly (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 $(w_1, w_2, \cdots, w_n)$ 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 to. 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.

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, Angrist and Imbens (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 $5,447$, with $3,927$ 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 Figure 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 Figure 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 to have a thorough estimate of the distribution of potential outcomes, which provides information not only on the mean or median, but also 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 densities as discussed by Kitagawa (2015). This method requires estimating two densities non-parametrically. 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 semi-parametric 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 $\{(x_i, t_i, d_i, z_i')' : 1 \leq i \leq n\}$.

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 $\mathcal{I} \subset \mathbb{R}$,

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

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

$$\mathbb{P}[t_i = 0|d_i = 0] \cdot f_{d=0,t=0}(x) \geq \mathbb{P}[t_i = 0|d_i = 1] \cdot f_{d=1,t=0}(x), \qquad \text{for all } x \in \mathcal{I}.$$

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 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 subsample. 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 $\hat{f}(x)$ using, respectively, the following weights:

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

Here, the weights need to be estimated in practice, unless precise knowledge about the treatment assignment mechanism is available. Our results again allow for $\sqrt{n}$-consistent estimated weights such as those obtained by fitting a flexible Logit or Probit model to approximate the propensity score $\mathbb{P}[d_i = 1|z_i]$.

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, Angrist and Imbens (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, $\$1,130$ 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 $\hat{f}(x)$: one first constructs two density estimates using the weights given earlier, $w_i^{d=0,t=0}$ and $w_i^{d=1,t=0}$. We plot the two estimated (scaled) densities in Figure 3. A simple eyeball test suggests no evidence against instrumental variable validity. A formal hypothesis test, justified using our theoretical results, confirms this finding.

Second, we estimate the density of the potential outcomes for compliers. In panel (a) of Figure 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 Figure 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 different is much larger. Simple mean comparison implies that enrolling in JTPA is associated with $\$2,083$ 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 Figure 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 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 estimator, 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 two extensions of our work: re-weighted local regression estimators, and local L_2 least squares projection estimators.

Although beyond the scope of this paper, it would be useful to generalized 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 and econometrics, including for instance estimation of auction models.

References

Abadie, A. (2003), "Semiparametric Instrumental Variable Estimation of Treatment Response Models," *Journal of Econometrics*, 113, 231–263.

Abadie, A., Angrist, J., and Imbens, G. (2002), "Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings," *Econometrica*, 70, 91–117.

Abadie, A., and Cattaneo, M. D. (2018), "Econometric Methods for Program Evaluation," *Annual Review of Economics*, 10, 465–503.

Belloni, A., Chernozhukov, V., Chetverikov, D., and Fernandez-Val, I. (2019), "Conditional Quantile Processes based on Series or Many Regressors," *Journal of Econometrics*, forthcoming.

Belloni, A., Chernozhukov, V., Chetverikov, D., and Kato, K. (2015), "Some New Asymptotic Theory for Least Squares Series: Pointwise and Uniform Results," *Journal of Econometrics*, 186, 345–366.

Cattaneo, M. D., Crump, R. K., Farrell, M. H., and Feng, Y. (2019a), "On Binscatter," arXiv:1902.09608.

Cattaneo, M. D., Farrell, M. H., and Feng, Y. (2019b), "Large Sample Properties of Partitioning-Based Estimators," *Annals of Statistics*, forthcoming.

Cattaneo, M. D., Jansson, M., and Ma, X. (2019c), "Simple Local Polynomial Density Estimators," *Journal of the American Statistical Association*, forthcoming.

——— (2019d), "`lpdensity`: Local Polynomial Density Estimation and Inference," working paper.

Cheng, M.-Y., Fan, J., and Marron, J. S. (1997), "On Automatic Boundary Corrections," *Annals of Statistics*, 25, 1691–1708.

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

——— (2014b), "Gaussian Approximation of Suprema of Empirical Processes," *Annals of Statistics*, 42, 1564–1597.

Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K., and Robins, J. M. (2019), "Locally Robust Semiparametric Estimation," arXiv:1608.00033.

Chernozhukov, V., Fernandez-Val, I., and Melly, B. (2013), "Inference on Counterfactual Distributions," *Econometrica*, 81, 2205–2268.

DiNardo, J., Fortin, N. M., and Lemieux, T. (1996), "Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach," *Econometrica*, 64, 1001–1044.

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

Giné, E., Koltchinskii, V., and Sakhanenko, L. (2004), "Kernel Density Estimators: Convergence in Distribution for Weighted Sup-Norms," *Probability Theory and Related Fields*, 130, 167–198.

Giné, E., and Nickl, R. (2010), "Confidence Bands in Density Estimation," *Annals of Statistics*, 38, 1122–1170.

Granovsky, B. L., and Müller, H.-G. (1991), "Optimizing Kernel Methods: A Unifying Variational Principle," *International Statistical Review/Revue Internationale de Statistique*, 373–388.

Hausman, J. A., and Newey, W. K. (1995), "Nonparametric Estimation of Exact Consumers Surplus and Deadweight Loss," *Econometrica*, 63, 1445–1476.

Ichimura, H., and Newey, W. K. (2019), "The Influence Function of Semiparametric Estimators," *arXiv:1508.01378*.

Imbens, G. W., and Rubin, D. B. (1997), "Estimating Outcome Distributions for Compliers in Instrumental Variables Models," *Review of Economic Studies*, 64, 555–574.

——— (2015), *Causal Inference in Statistics, Social, and Biomedical Sciences*, New York: Cambridge University Press.

Kitagawa, T. (2015), "A Test for Instrument Validity," *Econometrica*, 83, 2043–2063.

Newey, W. K. (1994a), "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62, 1349–1382.

——— (1994b), "Kernel Estimation of Partial Means and a General Variance Estimator," *Econometric Theory*, 10, 233–253.

Newey, W. K., Hsieh, F., and Robins, J. M. (2004), "Twicing Kernels and a Small Bias Property of Semiparametric Estimators," *Econometrica*, 72, 947–962.

Newey, W. K., and McFadden, D. L. (1994), "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics, Volume IV*, eds. R. F. Engle and D. L. McFadden, New York: Elsevier Science B.V., pp. 2111–2245.

Newey, W. K., and Ruud, P. A. (2005), "Density Weighted Linear Least Squares," in *Identification and Inference in Econometric Models: Essays in Honor of Thomas Rothenberg*, eds. D. Andrews and J. Stock, Cambridge: Cambridge University Press, pp. 554–573.

Newey, W. K., and Stoker, T. M. (1993), "Efficiency of Weighted Average Derivative Estimators and Index Models," *Econometrica*, 61, 1199–1223.

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

Robins, J. M., Hsieh, F., and Newey, W. K. (1995), "Semiparametric Efficient Estimation of a Conditional Density with Missing or Mismeasured Covariates," *Journal of the Royal Statistical Society: Series B (Methodological)*, 57, 409–424.

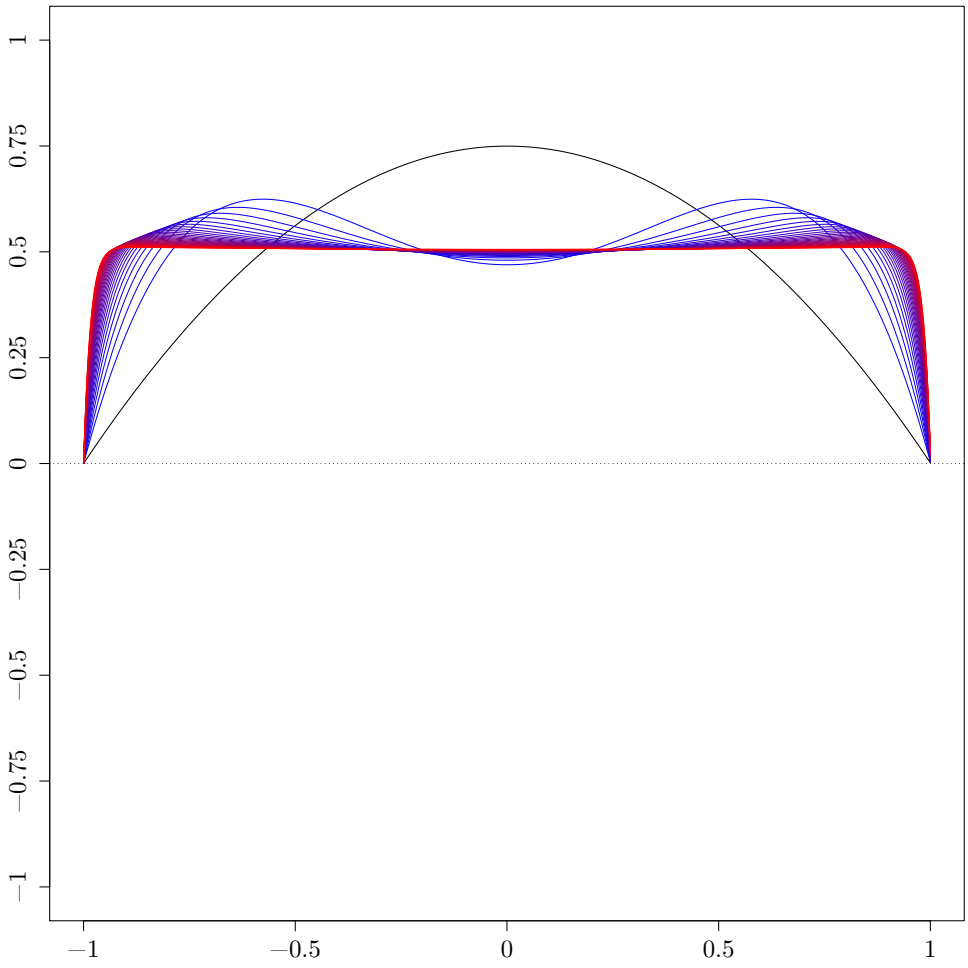
Zaitsev, A. Y. (2013), "The Accuracy of Strong Gaussian Approximation for Sums of Independent Random Vectors," *Russian Mathematical Surveys*, 68, 721–761.

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.

Figure 1: Equivalent Kernels of the Minimum Distance Estimators.

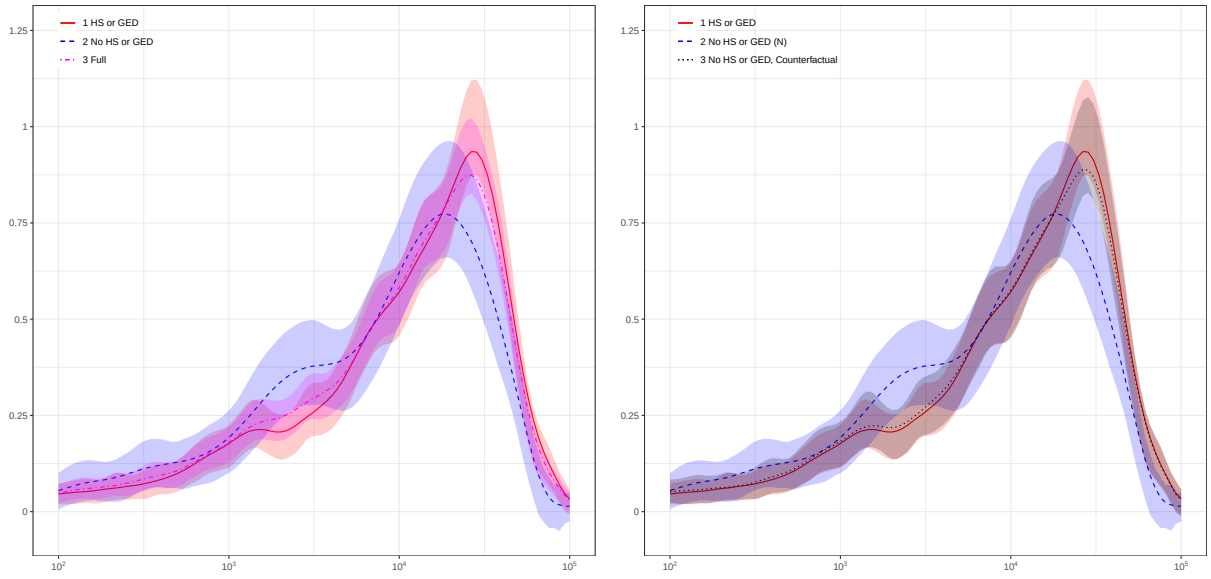


Notes. We set $P(u) = u$ or $P(u) = (u, u^2/2)'$, and K uniform. The redundant regressor is $Q(u) = u^{2j+1}$ for $j = 1, 2, \cdots, 30$. The initial equivalent kernel is quadratic (black solid line), and the minimum variance kernel is uniform (red solid line).

Figure 2: Earning Distributions by Education, JTPA.

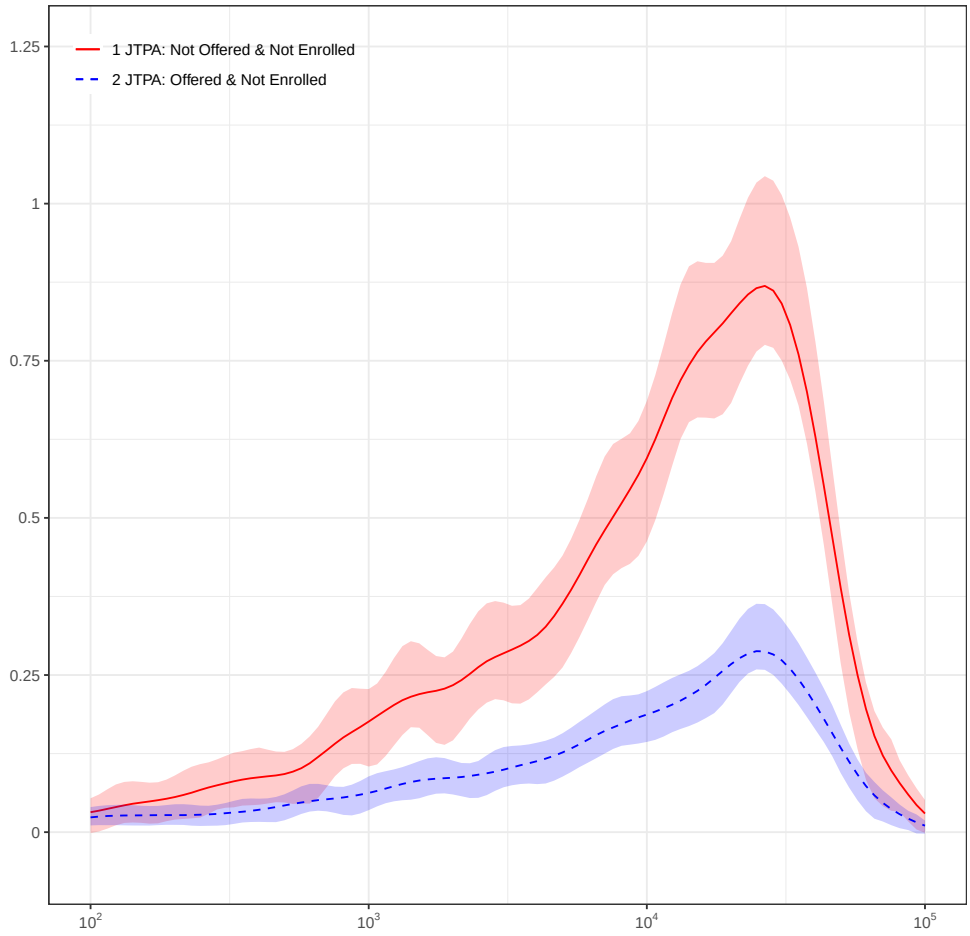
(a) Marginal Distributions

(b) Counterfactual Distribution



Notes: (i) Full: earning distribution for the full sample ($n = 5,447$); (ii) HS or GED (N/Y): earning distributions for subgroups without and with high school degree or GED ($n = 1,520$ and $3,927$, respectively); (iii) HS or GED (Y, 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, Jansson and Ma (2019d).

Figure 3: Testing Validity of Instruments, JTPA.

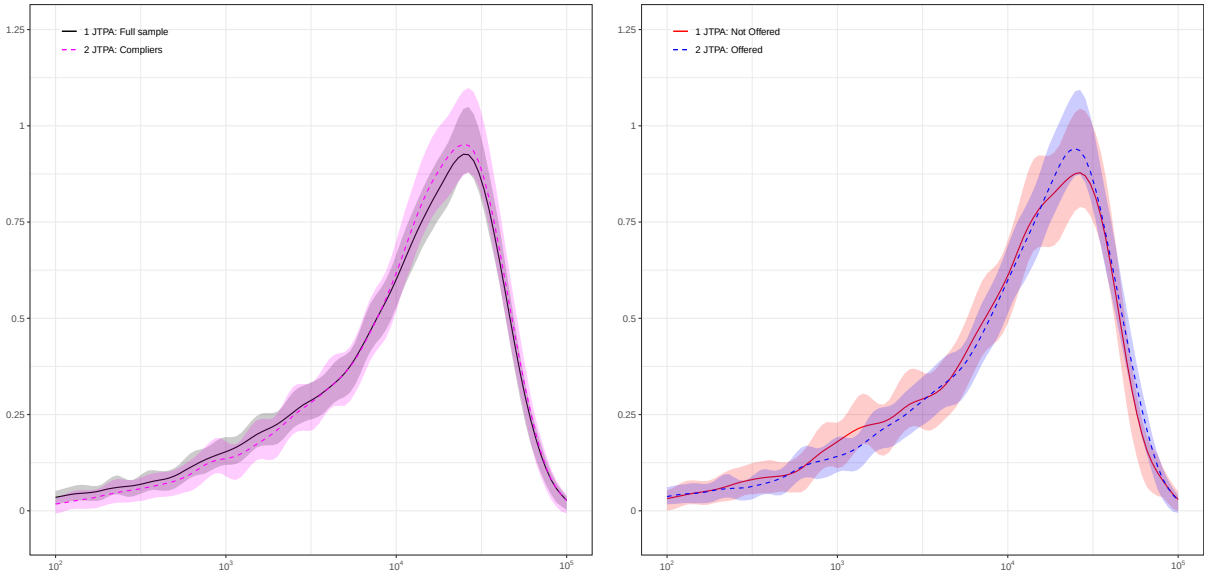


Notes: (i) JTPA: Not Offered & Not Enrolled: the scaled density estimate $\frac{\sum_i \mathbb{1}(t_i=0,d_i=0)}{\sum_i \mathbb{1}(d_i=0)} \hat{f}_{d=0,t=0}(x)$; (ii) JTPA: Offered & Not Enrolled: the scaled density estimate $\frac{\sum_i \mathbb{1}(t_i=0,d_i=1)}{\sum_i \mathbb{1}(d_i=1)} \hat{f}_{d=1,t=0}(x)$. Point estimates are obtained by using local polynomial regression with order 2, and robust confidence 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, Jansson and Ma (2019d).

Figure 4: Earning Distributions, JTPA.

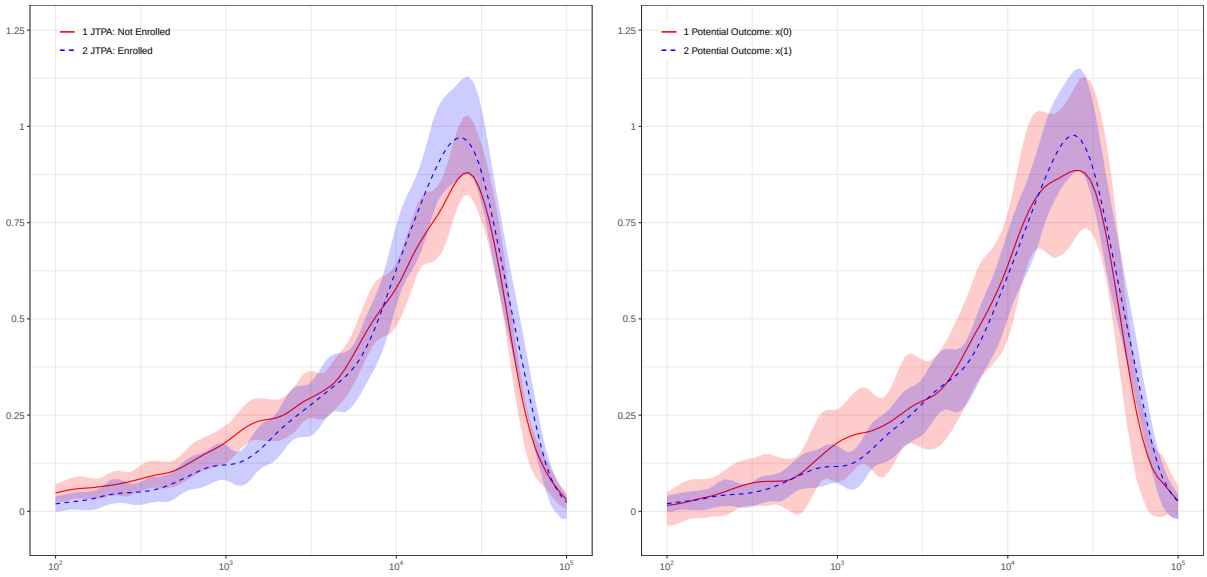
(a) Marginal Distributions

(b) Marginal Distributions



(c) Marginal Distributions

(d) Potential Outcome Distributions



Notes: panel (a) earning distributions in the full sample and for compliers; panel (b) earning distributions by JTPA offer; panel (c) earning distributions by JTPA enrollment; panel (d) distributions of potential outcomes for compliers. 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, Jansson and Ma (2019d).