Simple Local Polynomial Density Estimators*

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

September 27, 2017

Abstract

Empirical work in economics, statistics and many other disciplines often requires estimating one or more probability density functions, such as those of earnings or a poverty index, near or at a boundary point. Standard kernel density estimators cannot be used near boundary points due to their boundary bias, a fact that has led researchers to restrict attention to a region in the interior of the full support of the data or to employ other ad hoc smoothing or truncation methods. This paper presents an intuitive and easy-to-implement nonparametric density estimator based on local polynomial techniques, which does not require pre-binning or any other transformation of the data while still being fully boundary adaptive and automatic. This estimator is readily applicable to a variety of empirical contexts, including manipulation testing, counterfactual comparisons, treatment effects heterogeneity and specification, bunching, and auctions, just to mention a few obvious examples. We study the asymptotic properties of the proposed density estimator and use these results to provide fully automatic point estimation, inference and bandwidth selection methods. We apply these results to three specific empirical settings in program evaluation: discontinuity in density testing (McCrary, 2008), counterfactual analysis (DiNardo et al., 1996), and IV treatment effect specification and heterogeneity analysis (Abadie, 2003; Kitagawa, 2015). We showcase our methods with two empirical applications, and we also investigate their finite-sample performance in a Monte Carlo experiment. Our general results also cover estimation of the distribution function and derivatives of the density function, additional results useful in other nonparametric and semiparametric settings. Two distinct companion Stata and R software packages are provided.

Keywords: density estimation, local polynomial methods, regression discontinuity, manipulation test, counterfactual analysis, IV heterogeneity.

*A preliminary version of this paper circulated under the title "Simple Local Regression Distribution Estimators with an Application to Manipulation Testing". We thank Sebastian Calonico, Toru Kitagawa, Zhuan Pei, Andres Santos, Rocio Titiunik and Gonzalo Vazquez-Bare for useful comments that improved our work and software implementations. Cattaneo gratefully acknowledges financial support from the National Science Foundation (SES 1357561 and SES 1459931). Jansson gratefully acknowledges financial support from the National Science Foundation (SES 1459967) and the research support of CREATES (funded by the Danish National Research Foundation under grant no. DNRF78).

[†]Department of Economics and Department of Statistics, University of Michigan.

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

[§]Department of Economics and Department of Statistics, University of Michigan.

1 Introduction

Flexible (nonparametric) estimation of distribution functions, densities and derivatives thereof play an important role in empirical work in economics, statistics, and many other disciplines. Sometimes these quantities are the main objects of interest, while in other cases they are useful ingredients in forming other nonparametric or semiparametric inference procedures. For instance, in program evaluation, nonparametric density estimators are commonly used for manipulation testing (McCrary, 2008), counterfactual analysis (DiNardo et al., 1996), and IV treatment effect specification and heterogeneity analysis (Abadie, 2003; Kitagawa, 2015). Density-based presentation and testing methods are also used in program evaluation and causal inference to describe, for example, common support/overlap or distributional treatment effects (e.g., Imbens and Rubin, 2015, for a review and references). Furthermore, smooth estimates of probability density functions are used in many other literatures employing nonparametric and semiparametric methods (e.g., Ichimura and Todd, 2007).

A common problem faced by all density estimators in empirical work is the presence of boundary evaluation points on the support of the variable of interest: Whenever the density estimate is constructed at or near boundary points, which may or may not be known by the researcher, its finite- and large-sample statistical properties are affected. Standard kernel density estimators are invalid at or near boundary points, while other methods may remain valid but usually require choosing additional tuning parameters, transforming the data, a priori knowledge of the boundary point location, or some other boundary-related specific information or modification. Furthermore, it is usually the case that one type of density estimator must be used for evaluation points at or near the boundary, while a different type must be used for interior points. This has led to a proliferation of (mostly ad hoc) density estimation methods and/or corrections to address the ubiquitous boundary bias problem in practice. Perhaps the most common empirical approach is to restrict the analysis to an interior subset of the support of the variable of interest, something that at the minimum handicaps empirical work, and many times is not even feasible when the actual goal is to learn about the density at or near the boundary.

We introduce a novel nonparametric estimator of a density function constructed using local polynomial techniques (Fan and Gijbels, 1996), and then employ it to develop boundary adaptive and automatic density estimation and inference methods in program evaluation settings. Our estimator is intuitive and easy to implement, does not require pre-binning of the data or a priori knowledge of the boundary location, and enjoys all the desirable features associated with local polynomial regression estimation. In particular, the estimator automatically adapts to the (possibly unknown) boundaries of the support of the density without requiring specific data modification or additional tuning parameter choices, a feature that is unavailable for most other density estimators in the literature: see Karunamuni and Albert (2005) for a review on this topic. The most closely related approaches currently available in the literature are the local polynomial density estimators of Cheng et al. (1997) and Zhang and Karunamuni (1998), which require knowledge of the boundary location and pre-binning of the data (or, more generally, pre-estimation of the density near the boundary), and hence introduce additional tuning parameters that need to be chosen for implementation.

The heuristic idea underlying our estimator is quite simple: whereas other nonparametric density estimators are constructed by smoothing out a "rough" histogram estimate of the data, our estimator is constructed by smoothing out the empirical distribution function using local polynomial techniques. This leads to a density estimator that is constructed using a preliminary tuning-parameter-free and $\sqrt{n}$-consistent CDF estimator (where n denotes the sample size), and thus requires only choosing the bandwidth associated with the local polynomial fit at each evaluation point. Our general results cover estimation of the distribution function, density and derivatives thereof, for any polynomial order at both interior and boundary points, and formally give (i) asymptotic expansions of the leading bias and variance, (ii) asymptotic Gaussian distributional approximation and valid statistical inference, (iii) consistent standard error estimates, and (iv) consistent data-driven bandwidth selection based on an asymptotic mean squared error (MSE) expansion. All these results apply to both interior and boundary points in a fully automatic and data-driven way, without requiring a prior knowledge of the boundary location, transforming the estimator or the data in specific ways, or employing additional tuning parameters (beyond the main bandwidth present in any kernel-based nonparametric method).

While often overlooked by practitioners, automatic boundary adaptation in nonparametrics is of crucial importance because "in applications design points always have a bounded support" (Fan and Gijbels, 1996, p. 69), and in fact the boundary location is often unknown. Thus, our proposed density estimator offers a practically relevant approach for empirical work concerned with density estimation and inference. While our main results can be used in any nonparametric or semiparametric setting where estimators of distribution functions, densities and derivatives thereof are required, in this paper, we employ them to develop new estimation and inference methods in three program evaluation settings: manipulation testing, counterfactual analysis, and IV heterogeneity study. In each of these applications one or more density functions at or near a boundary point need to be estimated, and therefore our methods are particularly well suited.

Our first methodological application is related to manipulation testing, a problem that has recently attracted considerable attention in empirical work. Here the goal is to test for a discontinuity in the density of a random sample of units that has been divided in two disjoint groups, according to a hard-thresholding rule based on an observed random variable (usually called "index", "score" or "running variable") and a known cutoff point. McCrary (2008) proposed this clever idea in the context of regression discontinuity (RD) designs. The ultimate goal is to test formally whether units are systematically sorting around the cutoff point, and thus non-randomly selecting into one of the two groups (generally referred to as control and treatment groups). The key observation is that, in the absence of self-selection, the density of units near the cutoff would be continuous, and thus a statistical test can be formed to determine empirically whether there is evidence of sorting in empirical applications. This testing idea can be used not only as a falsification test in RD designs, but also as an empirical test for manipulation or self-selection in other impact evaluation settings.

Testing for manipulation naturally involves nonparametric density estimation at a boundary point – the cutoff point in the support of the running variable where group (or treatment) assignment is determined. Our density estimator is therefore particularly well-suited for constructing a new testing procedure in this context because it offers automatic boundary-adaptive density estimation in an intuitive and easy-to-implement way, requiring the choice of only one bandwidth. Using our main results, we develop a new manipulation test and establish its large-sample properties: the resulting testing procedure gives an alternative to the implementation in McCrary (2008), which employs the density estimator of Cheng et al. (1997) and thus requires choosing additional tuning parameters. We offer an empirical illustration of our methods employing the canonical Head Start data (Ludwig and Miller, 2007; Cattaneo et al., 2017c).

The other methodological applications given are concerned with counterfactual densities and IV treatment effect specification and heterogeneity, where researchers often want to estimate several density functions of reweighted data to then compare them across different evaluation points. In particular, here we consider the settings of DiNardo et al. (1996), Abadie (2003), and Kitagawa (2015), three papers where estimation of density functions naturally arise. Because our theoretical work allows for estimated weights, we can also develop new boundary adaptive density estimators applicable to these examples: our methods apply automatically to both interior and boundary points and therefore provide simple and easy-to-implement density estimation and inference methods that can be used over the entire support of the data, without requiring pre-binning, truncation or some other ad hoc transformation before the empirical analysis. To illustrate our proposed density methods for counterfactual and IV specification and heterogeneity analysis we report an empirical application using another canonical dataset in empirical microeconomics: the Job Training Partnership Act (JTPA) data.

All the density-based methods studied in this paper involve estimation and inference at or near a boundary point employing our proposed density estimator, possibly after including preliminary estimated weights, and are formally developed via the theoretical results given in the Supplemental Appendix. In fact, because our density estimators allow for $\sqrt{n}$-estimated weights such as those arising from inverse probability weighting, they can be used in many other empirically relevant contexts under the usual unconfoundedness or selection-on-observables assumption; typical areas of application include treatment effects, missing data, measurement error, and data combination. We do not describe these other applications in detail for brevity, and because they are very similar to the ones given below.

Finally, we also provide two general purpose software packages, for Stata and R, implementing the main results discussed in the paper. Cattaneo et al. (2017a) discusses the first package (lpdensity), which targets at generic density estimation over the support of the data, and Cattaneo et al. (2017b) discusses the second package (rddensity), which is specifically tailored to manipulation testing. In addition, we provide replication files of all the numerical results reported herein.

The rest of the paper is organized as follows. Section 2 introduces the density estimator, discusses the main intuition behind its construction, and outlines its applicability to program evaluation and related problems. Section 3 gives an overview of the main technical results developed in this paper. Section 4 applies our main theoretical results to the specific case of nonparametric testing of a discontinuity in a density at a point. Sections 5 and 6 develop the new density methods for counterfactual analysis and IV treatment effect heterogeneity, respectively. Section 7 offers a brief account of an extensive simulation study we conducted, while Section 8 discusses several potential extensions to our work and concludes. A long Supplemental Appendix reports more general theoretical results encompassing those discussed herein, includes the proofs of these general results, discusses additional methodological and technical results, and provides further simulation evidence.

2 Boundary Adaptive Density Estimation

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 over its possibly unknown support $\mathcal{X}$, and w_i is a weighting variable, possibly random and involving unknown parameters. We consider the generic parameter

$$f(x) = \frac{\partial}{\partial x}\mathbb{E}\Big[w_i \mathbb{1}(x_i \leq x)\Big],$$

whose practical interpretation depends on the specific choice of w_i. If $w_i = 1$, $f(\cdot)$ becomes the standard probability density function of the continuous variable of interest x_i. If $w_i = \mathbb{1}(x_i < \bar{x})$ or $w_i = \mathbb{1}(x_i \geq \bar{x})$, then $f(\cdot)$ is used to identify the left and right limits of the density function of x_i at the cutoff point $\bar{x}$ (i.e., $f_-(\bar{x}) = \lim_{x\uparrow\bar{x}} f(x)$ and $f_+(\bar{x}) = \lim_{x\downarrow\bar{x}} f(x)$), respectively, which is useful for manipulation testing (Section 4). If w_i is set to be a certain ratio of propensity scores for subpopulation membership, then $f(\cdot)$ becomes a counterfactual density function (Section 5). If w_i is set to be a combination of the treatment assignment and treatment status variables, then the resulting $f(\cdot)$ can be used to specification testing in IV settings, or if w_i is set to be a certain ratio of propensity scores for a binary instrument, then $f(\cdot)$ can be used to identify distributions of compliers in IV settings (Section 6).

Our results apply to known and unknown, as well as bounded or unbounded support $\mathcal{X}$, which is an important feature in most empirical applications employing density estimators. For example, in the context of manipulation testing (Section 4), the random variable x_i is a running variable, score or index, and the parameter of interest is the potential discontinuity of the density function, at an induced boundary point determined by the treatment eligibility cutoff. As another example, in counterfactual analysis or other related program evaluation contexts (e.g., Sections 5 and 6), the support of the data is unknown and often bounded, for example when x_i represents wage, earning or taxable income, in which case the boundary points are determined by the natural support (e.g., $x_i \geq 0$ if x_i is wage or earning) or by a policy (e.g., the value on the support $\mathcal{X}$ where a tax level/rate changes). These and other examples are discussed in upcoming sections.

We introduce a generic nonparametric estimator of $f(x)$, which is fully automatic, boundary adaptive for the possibly unknown support $\mathcal{X}$, and allows for $\sqrt{n}$-consistent estimated weights w_i. Our estimator requires only one tuning parameter choice and is very easy to construct and interpret. To describe it, first we define the plug-in weighted empirical CDF estimator

$$\tilde{F}(x) = \frac{1}{n}\sum_{i=1}^{n} w_i \mathbb{1}(x_i \leq x),$$

and then the proposed local polynomial density estimator $\hat{f}(x)$ is given by

$$\begin{bmatrix} \hat{\alpha}(x) \\ \hat{f}(x) \\ \hat{\boldsymbol{\gamma}}(x) \end{bmatrix} = \operatorname*{arg\,min}_{\alpha,\beta_1,\dots,\beta_p} \sum_{i=1}^{n} \Big(\tilde{F}(x_i) - \alpha - (x_i - x)\beta_1 - (x_i - x)^2\beta_2 - \cdots - (x_i - x)^p\beta_p\Big)^2 K\Big(\frac{x_i - x}{h}\Big),$$

where $K(\cdot)$ denotes a kernel function, h is a positive bandwidth, and $p \geq 1$. Our estimator takes the (weighted) empirical distribution function as an starting point but, instead of trying to numerically differentiate it, first constructs a simple and intuitive smooth local approximation to $\tilde{F}(x_i)$ using a polynomial expansion, and then obtains the density estimator as the slope coefficient in the local polynomial regression. To be specific, $\hat{\alpha}(x)$ is the intercept estimate, $\hat{f}(x)$ is the slope estimate (associated with β_1), and $\hat{\boldsymbol{\gamma}}(x)$ is a vector collecting estimated higher-order coefficients in the local polynomial approximation to $\tilde{F}(x_i)$. Using standard least squares algebra, all these coefficients can be given in closed form. Here we focus on the density estimator $\hat{f}(x)$, but in the Supplemental Appendix we also study the properties of $\hat{\alpha}(x)$ and $\hat{\boldsymbol{\gamma}}(x)$, as these objects may also be useful in some nonparametric and semiparametric applications (see Section 8 for further discussion).

The idea behind our density estimator $\hat{f}(x)$ is explained graphically in Figure 1, setting $w_i = 1$ only for simplicity. In this figure we consider three distinct evaluation points on $\mathcal{X} = [-1, 1]$: a is near the lower boundary, b is an interior point, and $c = 1$ is the upper boundary. Recall that the conventional kernel density estimator,

$$\hat{f}_{\mathtt{KD}}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right),$$

is valid for interior points, but otherwise inconsistent. See, e.g., Wand and Jones (1995) for a classical reference. On the other hand, our density estimator $\hat{f}(x)$ is valid for all evaluation points $x \in \mathcal{X}$ and can be used directly, without any modifications to approximate the unknown density. Figure 1 is constructed using $n = 500$ observations. The top panel plots one realization of the empirical distribution function $\hat{F}(x)$ in dark gray, and the local polynomial fits for the three evaluation points $x = a, b, c$ in red, the latter implemented with $p = 2$ (quadratic approximation) and bandwidth h (different value for each evaluation point considered). The vertical light gray areas highlight the localization region controlled by the bandwidth choice, that is, only observations falling in these regions are used to smooth-out the empirical distribution function via local polynomial approximation, depending on the evaluation point. Our proposed density estimator $\hat{f}(x)$ is the slope coefficient accompanying the first-order term in the local polynomial approximation, which is depicted in the bottom panel of Figure 1 as the solid line in red. The bottom panel also plots three other curves: dashed blue line corresponding to the population density function, dashed-dotted green line corresponding to the average of our density estimate over simulations, and dashed black line corresponding to average of the standard kernel density estimates obtained using $\hat{f}_{\mathtt{KD}}(x)$.

Figure 1 illustrates how our proposed density estimator adapts to (near) boundary points automatically, showing graphically its good performance in repeated samples. See the Supplemental Appendix for detailed simulation experiments corroborating these findings. Evaluation point b is an interior point and, consequently, a symmetric smoothing around that point is employed, just like the standard estimator $\hat{f}_{\mathtt{KD}}(x)$ does. On the other hand, evaluation points a and c both exhibit boundary bias if the standard kernel density estimator is used: point a is near the boundary and hence employs asymmetric smoothing, while point c is at the upper boundary and hence employs one-sided smoothing. In contrast, our proposed density estimator $\hat{f}(x)$ automatically adapts to the (possibly unknown) boundary point, as the bottom panel in Figure 1 illustrates. This feature makes $\hat{f}(x)$ particularly well-suited for empirical applications where there is known or unknown finite boundaries on the support of the data, which is arguably the case in most applications (and is always the case in finite samples). We are aware of only one other density estimator that exhibits automatic boundary carpentry: Cheng et al. (1997) introduced a local polynomial density estimator that requires knowledge of boundary location and is constructed using a preliminary histogram estimate, which by implication also requires several tuning parameters for implementation (i.e., histogram's bins length, location and total number, in addition to a bandwidth choice for the local polynomial fit). This estimator was popularized by McCrary (2008) in the context of RD designs; see Section 4.

More generally, when weights are allowed for, there is another potentially interesting connection between the estimand $f(x)$ and estimator $\hat{f}(x)$ described above, and the classical kernel-weighted averages featuring prominently in econometrics (e.g., Newey, 1994; Newey and McFadden, 1994). Because $f(x) = \mathbb{E}[w_i|x_i = x]g(x)$, with $g(x)$ denoting the probability density function of x_i, it follows that our proposed estimation approach gives an alternative boundary adaptive way of estimating density-weighted averages nonparametrically. Our proposed approach differs from standard kernel methods in that we first smooth out the $\sqrt{n}$-consistent empirical distribution function, which does not exhibit boundary problems, and then a nonparametric approximation to the desired derivative is extracted. This approach is conceptually distinct from the methods currently available in the literature, and it exhibits demonstrably superior properties such as automatic boundary adaptation for estimation and inference.

In the Supplemental Appendix we investigate the large sample properties of our proposed estimator $\hat{f}(x)$ when $w_i = w(\mathbf{z}_i;\boldsymbol{\theta}_0)$ is replaced by $\hat{w}_i = w(\mathbf{z}_i;\hat{\boldsymbol{\theta}})$ with $\hat{\boldsymbol{\theta}}$ a $\sqrt{n}$-consistent estimator of $\boldsymbol{\theta}_0$, that is, when estimated weights are used to construct the weighted empirical distribution function $\tilde{F}(x)$. This generalization is useful in counterfactual density estimation, IV treatment effects specification and heterogeneity analysis, bunching, missing data, and many other empirical problems of interest. Our general results include bias, variance and distributional approximations, standard error estimation, and optimal bandwidth selection and estimation, among other results. Because all these results are technical in nature, we relegate most details to the Supplemental Appendix. In Section 3 we include brief statements of our main results for completeness and reference. The main take-away is as follows: under regularity conditions (Assumptions 1, 2 and 3 below), which include mild restrictions on the possibly estimated weights w_i, and if $nh^2 \to \infty$ and $nh^{2p+1} \to 0$, then

$$\frac{\hat{f}(x)-f(x)}{\hat{\sigma}(x)} \rightsquigarrow \mathcal{N}(0,1), \qquad \hat{\sigma}(x)=\sqrt{\frac{1}{nh}\hat{\mathscr{V}}(x)}, \qquad \text{for all } x\in\mathcal{X},$$

where $\rightsquigarrow$ denotes convergence in distribution, and the exact formula of the variance estimator $\hat{\mathscr{V}}(x)$ is given in Section 3. Importantly, $\hat{\mathscr{V}}(x)$ is fully automatic and very easy to implement, and remains valid even when $\sqrt{n}$-consistent estimated weights are employed: We show that employing estimated weights has no first order impact on the nonparametric estimator $\hat{f}(x)$, which means in practice the weights can be treated as known for estimation, inference and bandwidth selection purposes. Therefore, our proposed estimation and inference methods are automatic and boundary adaptive, without requiring any specific modifications depending on the particular value x or the weighting scheme used.

The following section presents a technical summary of the main large-sample results we obtained for estimation and inference employing the proposed density estimator $\hat{f}(x)$. We then apply these generic results to specific problems of interest in empirical work in subsequent sections, where we also illustrate them using real data and applications. Other potential applications of our methods are briefly mentioned in Section 8.

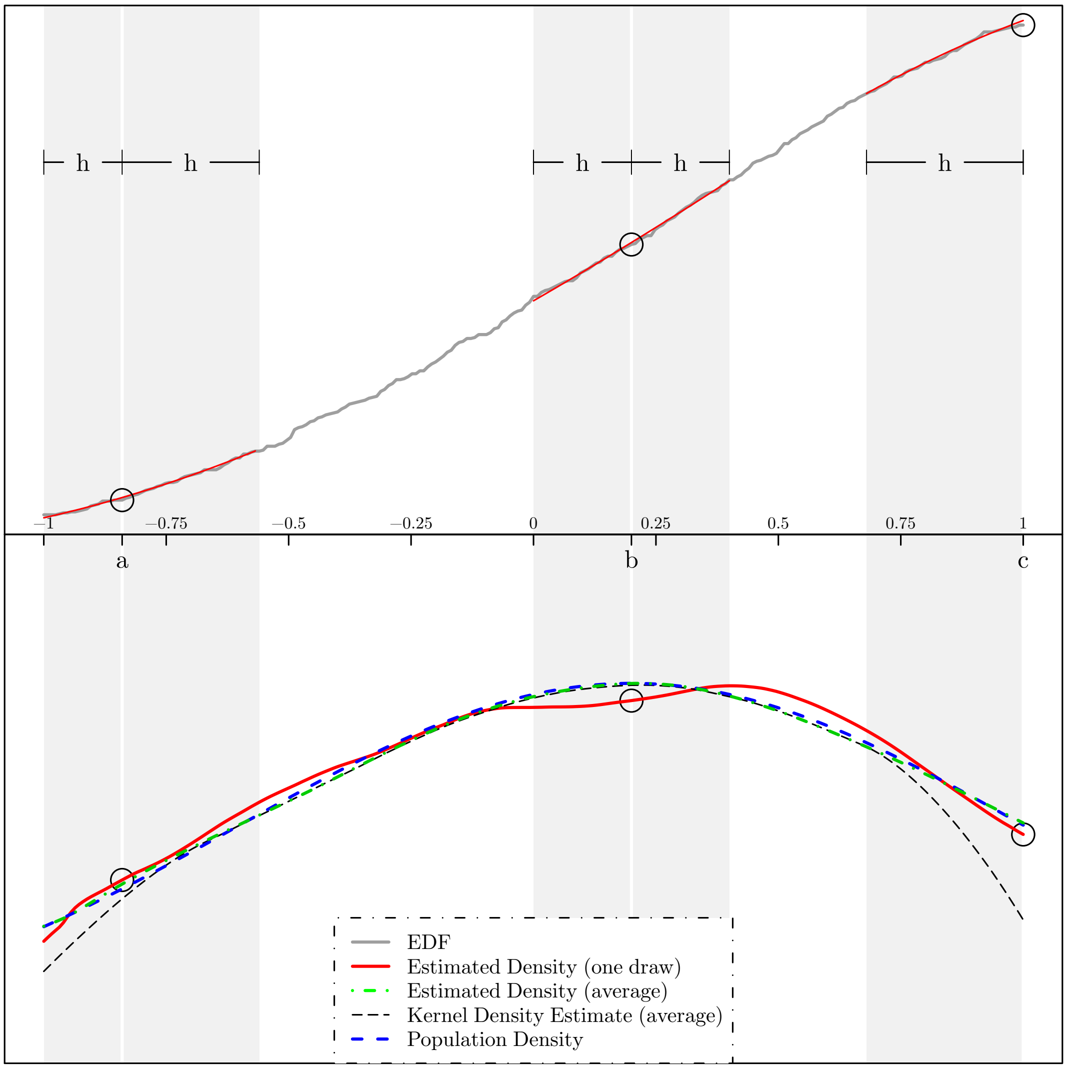
3 Overview of Technical Results

We summarize two main technical results on our density estimator: (i) an asymptotic distributional approximation with precise leading bias and variance characterizations, and (ii) a consistent standard error estimator which is also data-driven and fully-automatic. Both results are fully boundary adaptive and do not require prior knowledge of the shape of $\mathcal{X}$. We leave preliminary lemmas and detailed proofs to the Supplemental Appendix to conserve space. Properties of other estimators obtained by our method, including a smoothed distribution function estimator and estimators of higher order derivatives, are also available in the Supplemental Appendix.

Before stating the results, we give the regularity conditions employed.

Assumption 1 (DGP). $\{x_1, x_2, \cdots, x_n\}$ is a random sample of size n, with distribution function G that is $p+1$ times continuously differentiable for some $p\geq 1$ in a neighborhood of the evaluation point x, and the probability density function of x_i, denoted by g, is positive in a neighborhood of

Figure 1. Graphical Illustration of Density Estimator.



Notes: (i) Constructed using companion R (and Stata) package described in Cattaneo et al. (2017a) with simulated data.

the evaluation point x.

Note that G is generally different from F, except if uniform weighting $w_i = 1$ is used. To allow for estimated weighting schemes, we assume the weights take the form $w_i = w(\mathbf{z}_i; \boldsymbol{\theta}_0)$, where $\mathbf{z}_i$ are additional variables (which can include x_i), and $\boldsymbol{\theta}_0$ is a finite dimensional parameter.

Assumption 2 (Estimated Weights).
(i) $\{\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_n\}$ is a random sample of size n, $w_i = w(\mathbf{z}_i; \boldsymbol{\theta}_0)$, with $\mathbb{E}[w_i|x_i]$ being p times continuously differentiable, $\mathbb{E}[w_i^2|x_i]$ continuous, and $\mathbb{E}[w_i^4] < \infty$.
(ii) $\boldsymbol{\theta} \mapsto w(\cdot; \boldsymbol{\theta})$ is twice continuously differentiable and, for some $\delta > 0$, $\mathbb{E}[\sup_{|\theta - \theta_0| \le \delta} |w(\mathbf{z}_i; \boldsymbol{\theta})| + |\dot{\mathbf{w}}(\mathbf{z}_i; \boldsymbol{\theta})| + |\ddot{\mathbf{w}}(\mathbf{z}_i; \boldsymbol{\theta})|] < \infty$, where $\dot{\mathbf{w}}(\mathbf{z}_i; \boldsymbol{\theta}) = \partial w(\mathbf{z}_i; \boldsymbol{\theta})/\partial \boldsymbol{\theta}'$ and $\ddot{\mathbf{w}}_i(\mathbf{z}_i; \boldsymbol{\theta}) = \partial^2 w(\mathbf{z}_i; \boldsymbol{\theta})/\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'$.
(iii) $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ admits an asymptotic linear representation.

Part (i) ensures that $f(x)$ is well-defined and possesses certain smoothness, and finite fourth moment of the weights is used to justify a Lindeberg Condition for asymptotic normality. Part (ii) is used in the asymptotic variance formula. Collectively, this assumption provides regularity conditions on the estimated weights, ensuring summability of certain quantities and asymptotic expansions. To save notation, we set $\dot{\mathbf{w}}_i = \dot{\mathbf{w}}(\mathbf{z}_i; \boldsymbol{\theta}_0)$ and $\ddot{\mathbf{w}}_i = \ddot{\mathbf{w}}(\mathbf{z}_i; \boldsymbol{\theta}_0)$. Observed (possibly random) weights are allowed as a special case where $\mathbf{z}_i = w_i$ and $w(\cdot)$ is the identity map.

Assumption 3 (Kernel). The kernel function $K(\cdot)$ is nonnegative, symmetric, and continuous on its support $[-1, 1]$.

This assumption is standard in nonparametric estimation, and is satisfied for common kernel functions. We exclude kernels with unbounded support (for example the Gaussian kernel) for simplicity, since such kernels will always hit boundaries. Our results, however, remain to hold for common unbounded kernels with careful analysis, albeit the notation becomes more cumbersome.

Theorem 1 (Distributional Approximation). Suppose Assumption 1–3 hold with either observed weights w_i or estimated $\hat{w}_i = w_i(\mathbf{z}_i; \hat{\boldsymbol{\theta}})$. If $nh^2 \to \infty$ and $nh^{2p+1} = O(1)$, then

$$\frac{\hat{f}(x) - f(x) - h^p \mathscr{B}(x)}{\sqrt{\frac{1}{nh} \mathscr{V}(x)}} \rightsquigarrow \mathcal{N}(0, 1), \qquad \text{for all } x \in \mathcal{X},$$

where the asymptotic bias and variance are defined as

$$\mathscr{B}(x) = \mathbf{e}_1'\mathbf{A}(x)^{-1}\mathbf{a}(x), \qquad \mathscr{V}(x) = \mathbf{e}_1'\mathbf{A}(x)^{-1}\mathbf{B}(x)\mathbf{A}(x)^{-1}\mathbf{e}_1,$$

with

$$\begin{aligned}
\mathbf{A}(x) &= g(x)\int_{h^{-1}(\mathcal{X}-x)} \mathbf{r}_p(u)\,\mathbf{r}_p(u)'\,K(u)\,\mathrm{d}u,\\
\mathbf{a}(x) &= g(x)\frac{F^{(p+1)}(x)}{(p+1)!}\int_{h^{-1}(\mathcal{X}-x)} u^{p+1}\mathbf{r}_p(u)\,K(u)\,\mathrm{d}u,\\
\mathbf{B}(x) &= g(x)^2 H^{(1)}(x)\iint_{h^{-1}(\mathcal{X}-x)} \min\{u,v\}\mathbf{r}_p(u)\,\mathbf{r}_p(v)'\,K(u)K(v)\mathrm{d}u\mathrm{d}v,
\end{aligned}$$

$\mathbf{r}_p(u) = (1, u, u^2, \cdots, u^p)'$ being the p-th order polynomial expansion, $\mathbf{e}_1 = (0, 1, 0, \cdots, 0)'$ the second unit vector, and $H(x) = \mathbb{E}[w_i^2\mathbb{1}(x_i \le x)]$.

We make two remarks here. First the integration region reflects the effect of boundaries. Recall that the kernel function is compactly supported, and if x is an interior point, we have, in large samples, $h^{-1}(\mathcal{X}-x) \supset [-1,1]$, so that the kernel function is not truncated and the local approximation is symmetric around x. On the other hand, for x near or at boundaries, $h^{-1}(\mathcal{X}-x) \not\supset [-1,1]$, and the local approximation is asymmetric or even one-sided.

Second, both the unweighted distribution G and the weighted distribution F feature in the matrices defined above. The unweighted density of x_i shows up reflecting the "design" of the local regression, since weighting is only conducted in $\tilde{F}$. The weighted distribution feature as part of the smoothing bias. Interestingly, the unweighted distribution *does not* feature in the asymptotic distribution (bias or variance), since it is canceled in $\mathbf{A}(x)^{-1}\mathbf{a}(x)$ and $\mathbf{A}(x)^{-1}\mathbf{B}(x)\mathbf{A}(x)^{-1}$: This is desirable as our proposed estimator is design adaptive.

Next we show how standard error can be constructed. The one we propose is an appealing companion to our main estimator and is highly relevant for empirical applications, since it does not require knowledge of relative positioning of the evaluation point to boundaries of $\mathcal{X}$ (if any). For notational simplicity, we normalize the observations as $\check{x}_i = h^{-1}(x_i - x)$, and define

$$\hat{\mathbf{A}}(x) = \frac{1}{nh}\sum_{i=1}^{n} \mathbf{r}_p(\check{x}_i)\,\mathbf{r}_p(\check{x}_i)'\,K(\check{x}_i)$$

$$\hat{\mathbf{B}}(x) = \frac{1}{n^3h^3} \sum_{i,j,k=1}^{n} \mathbf{r}_p\left(\check{x}_j\right)\mathbf{r}_p\left(\check{x}_k\right)' K\left(\check{x}_j\right) K\left(\check{x}_k\right) \hat{w}_i^2 \Big(\mathbb{1}[x_i \le x_j] - \tilde{F}(x_j)\Big)\Big(\mathbb{1}[x_i \le x_k] - \tilde{F}(x_k)\Big),$$

and

$$\hat{\mathscr{V}}(x) = \mathbf{e}_1'\hat{\mathbf{A}}(x)^{-1}\hat{\mathbf{B}}(x)\hat{\mathbf{A}}(x)^{-1}\mathbf{e}_1.$$

Then we have the following:

Theorem 2 (Variance Estimation)**.** If the conditions in Theorem 1 hold, then $\hat{\mathscr{V}}(x) \to_{\mathbb{P}} \mathscr{V}(x)$ for all $x \in \mathcal{X}$.

Finally, using the result above and under regularity conditions, it follows that the pointwise (approximate) MSE-optimal bandwidth choice for our proposed density estimator is

$$h_{\mathtt{MSE}}(x) = \left(\frac{\mathscr{V}(x)}{2p\mathscr{B}(x)^2}\right)^{1/(1+2p)} n^{-1/(1+2p)}.$$

The Supplemental Appendix offers details on this choice, presents analogous integrated version, and discusses other related results such as valid implementation and consistent bandwidth estimation. Furthermore, in the Supplemental Appendix we present analogous results to those above for the smooth CDF $\hat{\alpha}(x)$ and derivatives estimators $\hat{\boldsymbol{\gamma}}(x)$, and also present uniform asymptotic results, among other technical developments omitted here to conserve space. `Stata` and `R` general purpose software implementing all these results is discussed in Cattaneo et al. (2017a).

4 Application: Manipulation Testing

One of the main features of the density estimator introduced in this paper is that it automatically reduces boundary bias, while at the same time avoiding the need for choosing additional tuning and smoothing parameters necessarily present in other related procedures (e.g., bins structure in pre-binning estimators or distance to the boundary in boundary-corrected estimators). In other words, this intuitive estimator automatically generates a boundary-corrected kernel density estimator requiring the choice of only one tuning parameter: the main bandwidth h. As a first empirical application of our proposed density estimator exploiting these features, we consider density discontinuity testing at a cutoff point inducing a change in treatment status, a form of manipulation testing originally introduced by McCrary (2008) in the context of RD designs. See also Frandsen (2017) for a complementary manipulation test with discrete running variable in RD designs.

Testing for manipulation is quite useful when units are assigned to two (or more) distinct groups using a hard-thresholding rule, as it provides an intuitive and simple method to check empirically whether units are able to alter (i.e., manipulate) their assignment. Manipulation tests are used in empirical work both as falsification tests of RD designs and as empirical tests with substantive implications. The implementations available in the literature require choosing multiple tuning parameters (McCrary, 2008) or employ empirical likelihood methods together with boundary-corrected kernels (Otsu et al., 2014). The method we introduced in this section, on the other hand, requires choosing only one tuning parameter, avoids pre-binning the data, and permits the use of simple well-known weighting schemes (i.e., kernel functions), and thus it removes the need of choosing the length and positions of bins, or employing complicated boundary kernels directly. For example, regular kernels such as the uniform or triangular kernel can be used for implementation. In addition, our method is quite intuitive and easy to implement, while being fully data-driven and principled, as bandwidth selection methods are also formally developed and implemented.

To describe the manipulation testing setup, suppose units are assigned to one group ("control") if $x_i < \bar{x}$ and to another group ("treatment") if $x_i \geq \bar{x}$. For example, in the application discussed below we employ the Head Start data from Ludwig and Miller (2007), where x_i is a poverty index at the county level, $\bar{x} = 59.1984$ is a fixed cutoff determining eligibility to the program (see panel (a) in Figure 2 below). The goal is to test formally whether the density $f(x)$ is continuous at $\bar{x}$, using the two subsamples $\{x_i : x_i < \bar{x}\}$ and $\{x_i : x_i \geq \bar{x}\}$. Formally, the null and alternative hypotheses are:

$$\mathsf{H}_0 : \lim_{x\uparrow\bar{x}} f(x) = \lim_{x\downarrow\bar{x}} f(x) \qquad \text{vs} \qquad \mathsf{H}_1 : \lim_{x\uparrow\bar{x}} f(x) \neq \lim_{x\downarrow\bar{x}} f(x).$$

This hypothesis testing problem, of course, induces a nonparametric boundary point problem at $x = \bar{x}$ because two distinct densities need to be estimated, one from the left and the other from the right. This problem renders standard kernel density estimator inapplicable, but our proposed density estimator $\hat{f}(x)$ is readily applicable in this context because it is boundary adaptive and fully automatic. Furthermore, our estimator can be used to plot the density near the cutoff in an automatic way; see panel (b) of Figure 2 below for an example using the Head Start data.

To be more precise, first define $\tilde{F}_-(x)$ and $\tilde{F}_+(x)$ to be the weighted empirical distribution functions constructed using $w_i^- = \mathbb{1}(x_i < \bar{x})$ and $w_i^+ = \mathbb{1}(x_i \geq \bar{x})$, respectively. Then, our proposed density estimator can be applied twice, to the data below and above the cutoff, to obtain two estimators of the density at the boundary point $\bar{x}$, which we denote by $\hat{f}_-(x)$ and $\hat{f}_+(x)$, respectively. Thus, our new manipulation test statistic takes the form:

$$T(h) = \frac{\hat{f}_+(\bar{x}) - \hat{f}_-(\bar{x})}{\sqrt{\hat{\sigma}_+^2(\bar{x}) + \hat{\sigma}_-^2(\bar{x})}},$$

where $\hat{\sigma}_-(x)$ and $\hat{\sigma}_+(x)$ denote the standard error estimators mentioned previously but now computed with the weighting choice $w_i^- = \mathbb{1}(x_i < \bar{x})$ and $w_i^+ = \mathbb{1}(x_i \geq \bar{x})$, respectively. The exact formula is given in Section 3, and all other technical details are discuss in the Supplemental Appendix.

Therefore, employing our main theoretical results, we obtain conditions so that the finite sample distribution of $T(h)$ can be approximated by the standard normal distribution, which leads to the following result: under assumptions specified in Section 3, and the vanishing bandwidth sequence satisfies $nh^2 \to \infty$ and $nh^{1+2p} \to 0$, then

(1) Under H_0, $\lim_{n\to\infty} \mathbb{P}[|T(h)| \geq \Phi_{1-\alpha/2}] = \alpha$,

(2) Under H_1, $\lim_{n\to\infty} \mathbb{P}[|T(h)| \geq \Phi_{1-\alpha/2}] = 1$,

where Φ_α denotes the α-quantile of the standard Gaussian distribution. This result establishes asymptotic validity of the α-level testing procedure that rejects H_0 iff $|T(h)| \geq \Phi_{1-\alpha/2}$, $\alpha \in (0, 1)$, and also shows its consistency. The result follows immediately from our generic asymptotic approximations for $\hat{f}(x)$ after using $w_i^- = \mathbb{1}(x_i < \bar{x})$ and $w_i^+ = \mathbb{1}(x_i \geq \bar{x})$, and evaluating at $x = \bar{x}$; see the Supplemental Appendix for detailed proofs and related technical and implementation issues.

A key implementation issue of our manipulation test is the choice of bandwidth h, a problem common to all nonparametric manipulation tests available in the literature. On the other hand, an important feature of our method is that this bandwidth h is the only tuning parameter needed for implementation, unlike other manipulation tests available in the literature. To select h in an automatic and data-driven way, we obtain in the Supplemental Appendix a mean squared error optimal (MSE-optimal) bandwidth choice for the point estimator $\hat{f}_+(\bar{x}) - \hat{f}_-(\bar{x})$ and we proposed a consistent implementation thereof, which is denoted by $\hat{h}_p$. We also present other alternatives such MSE-optimal bandwidth selectors for each-side density estimator separately.

Given the data-driven bandwidth choice $\hat{h}_p$, we propose a simple robust bias-corrected test statistic implementation, employing ideas in Calonico et al. (2014) and Calonico et al. (2017); see the later reference for theoretical results on higher-order refinements and the important role of pre-asymptotic variance estimation. Specifically, our proposed data-driven robust bias-corrected test statistic is $T_{p+1}(\hat{h}_p)$, which rejects H_0 iff $|T_{p+1}(\hat{h}_p)| \geq \Phi_{1-\alpha/2}$ for a nominal α-level test. This approach corresponds to a special case of manual bias-correction together with the corresponding adjustment of Studentization. In practice, most common choices are $p = 2$. Companion general purpose software in Stata and R is presented in Cattaneo et al. (2017b) for discontinuity in density test.

4.1 Empirical Illustration

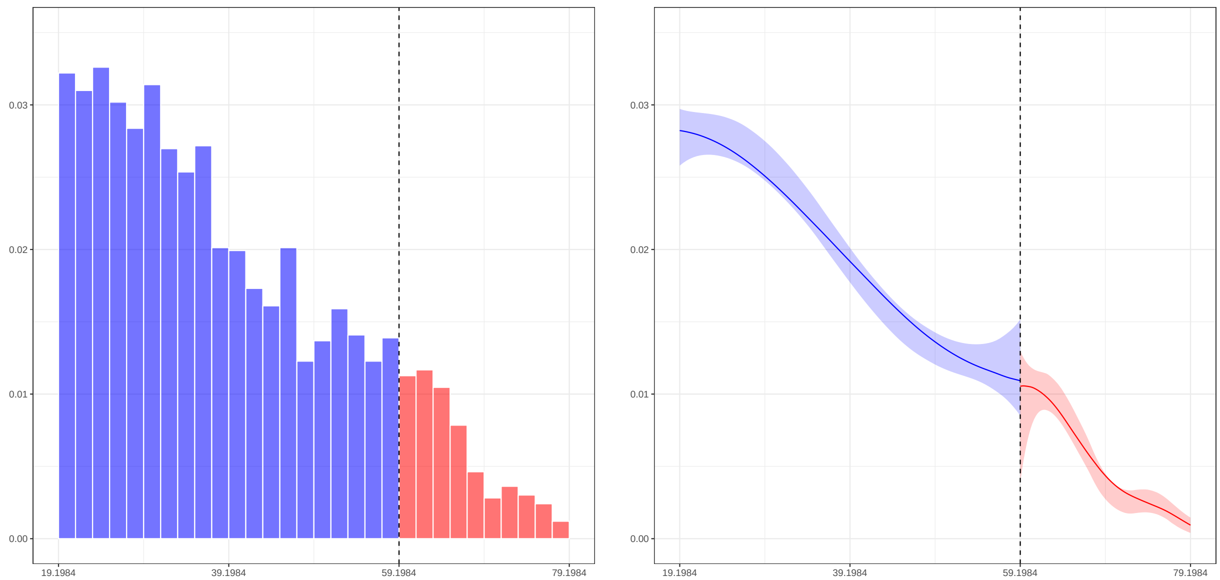
We apply our proposed manipulation test to the data of Ludwig and Miller (2007) on the original Head Start implementation in the U.S. In this empirical application, the data captures a discontinuity on access to program funds at the county level, which occurred in 1965 when the program was first implemented: to ensure that applications from the poorest communities would be represented in a nationwide grant competition for the program's funds, the federal government provided assistance to the 300 poorest counties in the U.S. to write and submit applications for Head Start funding. This led to increased Head Start participation and funding rates in these counties, creating a discontinuity in program participation at the 300th poorest county. Using our notation, x_i denotes the poverty index for county i, which was computed in 1965 using 1960 Census variables, and $\bar{x} = 59.1984$ is the cutoff point and poverty index of the 300th poorest municipality.

A manipulation test in this context amounts to testing whether there is a disproportional number of counties are situated above $\bar{x}$ relative to those present below the cutoff, which can be formally tested by employing our proposed discontinuity in density test. To begin, Figure 2(a) presents the histogram of counties below and above the cutoff. This rough density estimate is the preliminary data processing used in the original test proposed by McCrary (2008), which requires choosing first both the bin length of the histogram and the number (and location) of bins, and then a bandwidth for the second-step local polynomial fit. Figure 2(b) presents our smooth local polynomial density estimate along with pointwise confidence intervals for a grid of points near the cutoff $\bar{x}$. This fits is fully automatic, as it is constructed using a local data-driven bandwidth estimate and robust bias-correction, and is obtained using our general purpose software described in Cattaneo et al. (2017a).

Table 1 presents the empirical results from our manipulation test. This considers two main approaches, both covered by our theoretical work and available in our software implementation: (i) using two distinct bandwidths on each side of the cutoff ($h_- \neq h_+$) and (ii) using a common bandwidth for each side of the cutoff ($h_- = h_+$), with h_- and h_+ denoting the bandwidth on the left and on the right, respectively. For each of these approaches, we consider three distinct implementations of our manipulation test, which varies the degree of polynomial approximation used to smooth-out the empirical distribution function. Specifically, $T_q(h_p)$ denotes the test statistic constructed using a q-th order local polynomial density estimator, with bandwidth choice that is MSE-optimal for p-th order local polynomial density estimator. For example, our recommended choice is $T_3(h_2)$, with either common bandwidth or two different bandwidths, which amounts to first choose MSE-optimal bandwidth(s) for local quadratic fit, and then conduct inference using a cubic approximation instead. This approach, as mentioned before, is a simple implementation of the robust bias-correction method (Calonico et al., 2014, 2017), and has been shown to deliver not only valid first-order inference but also higher-order improvements in related settings. Notice that $T_p(h_p)$ does not lead to a valid inference approach, in general, because a first-order bias will make the test over-reject the null hypothesis (for example, see the simulations reported in Cattaneo et al., 2017b, Section 6).

In this application, our empirical results show no evidence of manipulation. In fact, this finding is consistent with the underlying institutional knowledge of the program: the poverty index was constructed in 1965 at the federal level using county level information from the 1960 Census, which implies it is indeed highly implausible that individual counties could have manipulated their assigned poverty index. Results in Table 1 suggests that this finding is also robust against different bandwidth and local polynomial order specifications.

Figure 2. Manipulation Testing, Head Start Data.



(a) Histogram Estimate

(b) Local Polynomial Density Estimate

Notes: (i) panel (a) reports histogram estimate of the running variable (poverty index) computed with default values in R, and panel (b) reports local polynomial density using companion R (and Stata) package described in Cattaneo et al. (2017b); and (ii) $n_- = 2,504$, $n_+ = 300$, and $\bar{x} = 59.1984$.

Table 1. Manipulation Testing, Head Start Data.

	Bandwidths		**Eff.** n		**Test**	
	left	right	left	right	T	p-val
$h_- \neq h_+$						
$T_2(\hat{h}_1)$	15.771	2.326	581	65	0.024	0.981
$T_3(\hat{h}_2)$	19.776	8.296	762	210	-1.146	0.252
$T_4(\hat{h}_3)$	32.487	10.808	1598	232	-1.083	0.279
$h_- = h_+$						
$T_2(\hat{h}_1)$	3.274	3.274	99	95	-1.355	0.175
$T_3(\hat{h}_2)$	9.213	9.213	316	221	-0.515	0.607
$T_4(\hat{h}_3)$	12.270	12.270	419	243	-0.712	0.477

Notes: (i) $T_p(h)$ denotes the manipulation test statistic using p-th order density estimators with bandwidth choice h (which could be common on both sides or different on either side of the cutoff), and $\hat{h}_p$ denotes the estimated MSE-optimal bandwidths for p-th order density estimator or difference of estimators (depending on the case considered); (ii) Columns under "Bandwidths" report estimated MSE-optimal bandwidths, Columns under "Eff. n" report effective sample size on either side of the cutoff, and Columns under "Test" report value of test statistic (T) and two-sided p-value (p-val); and (iii) first three rows allow for different bandwidths on each side of the cutoff, while last three rows employ a common bandwidth on both sides of the cutoff (chosen to be MSE-optimal for the difference of density estimates). All estimates are obtained using companion R (and Stata) package described in Cattaneo et al. (2017b).

5 Application: Counterfactual Densities

The previous application on manipulation testing focused on density estimation and inference at a boundary point. The density was also estimated at other points near the boundary but only for graphical presentation (Figure 2(b)). In this second application, the object of interest are density functions over their entire support, including boundaries and near-boundary regions. In addition, we now employ local polynomial density estimation with 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). The focus of this section is on density estimation, and we refer readers to Chernozhukov et al. (2013), and references therein, 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, reweighted 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. Following DiNardo et al. (1996), suppose the observed data is $\{(x_i, t_i, \mathbf{z}_i')' : 1 \leq i \leq n\}$, where x_i continues to be the main outcome variable, $\mathbf{z}_i$ 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 interpretations.

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 $\mathbf{z}_i$, $f_{1\rhd 0}(x)$ 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 which interpretation the researcher takes, $f_{1\rhd 0}(x)$ is of interest and can be estimated using our generic density estimator $\hat{f}(x)$ with the following weights:

$$w_i^{1\rhd 0} = t_i \cdot \frac{\mathbb{P}[t_i = 0|\mathbf{z}_i]}{\mathbb{P}[t_i = 1|\mathbf{z}_i]} \frac{\mathbb{P}[t_i = 1]}{\mathbb{P}[t_i = 0]}.$$

In practice, this choice of weighting scheme is unknown because the conditional probability $\mathbb{P}[t_i = 1|\mathbf{z}_i]$, 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. All our theoretical results presented in the Supplement Appendix, including distributional approximations and consistent bandwidth selection, continue to apply in this case.

5.1 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 $5,447$, with $3,927$ being either high school graduates or GED. Summary statistics are available as the fourth column in Table 2. We leave further details on the JTPA program to Section 6, 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 3. 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.

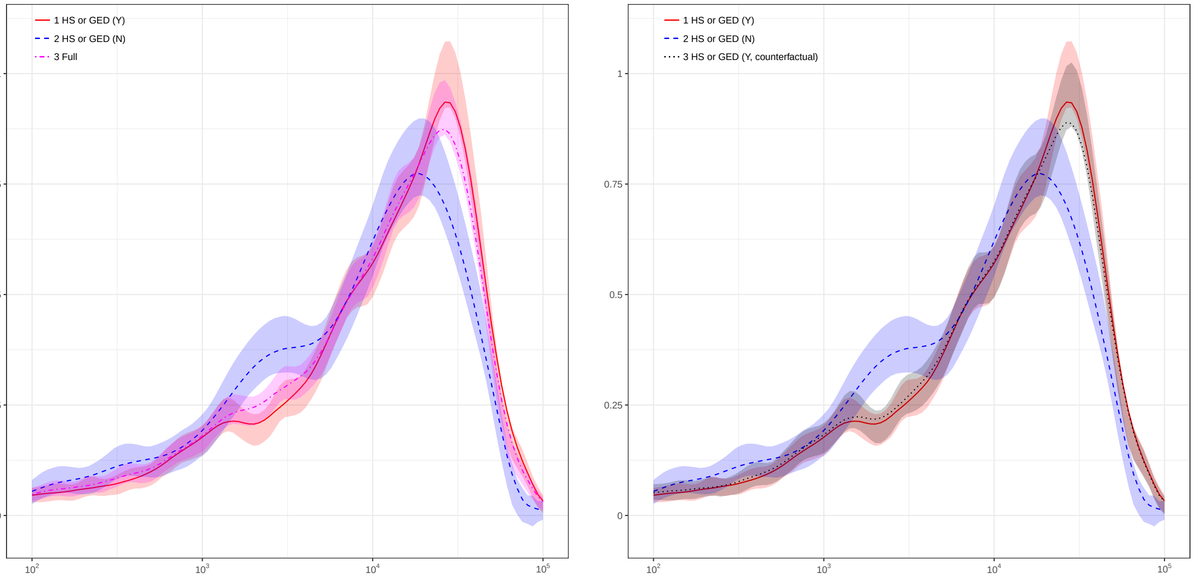
As mentioned earlier, direct comparison between $\hat{f}_1(x)$ and $\hat{f}_0(x)$ does not reveal the impact of having high school degree on earning, since the difference is confounded by the fact that individuals with high school degree can have very different characteristics (measured by covariates) 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 3. 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 significant. 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 difference in income and earning.

6 Application: 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

Figure 3. 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 et al. (2017a).

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 nonparametrically with non-estimated weights. 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. This density method requires 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 $\mathbf{z}_i$. We assume that the observed data is a random sample $\{(x_i, t_i, d_i, \mathbf{z}_i')' : 1 \leq i \leq n\}$.

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 Balke and Pearl (1997) and Heckman and Vytlacil (2005), 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 $B \subset \mathbb{R}$,

$$\begin{aligned} &\mathbb{P}[x_i \in B,\ t_i = 1 | d_i = 1] \geq \mathbb{P}[x_i \in B,\ t_i = 1 | d_i = 0], \\ \text{and} \qquad &\mathbb{P}[x_i \in B,\ t_i = 0 | d_i = 0] \geq \mathbb{P}[x_i \in B,\ t_i = 0 | d_i = 1]. \end{aligned}$$

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

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

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 reweighted 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:

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

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

6.1 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 trainings, the treatment takeup, however, was only about 67% among those who

Table 2. 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

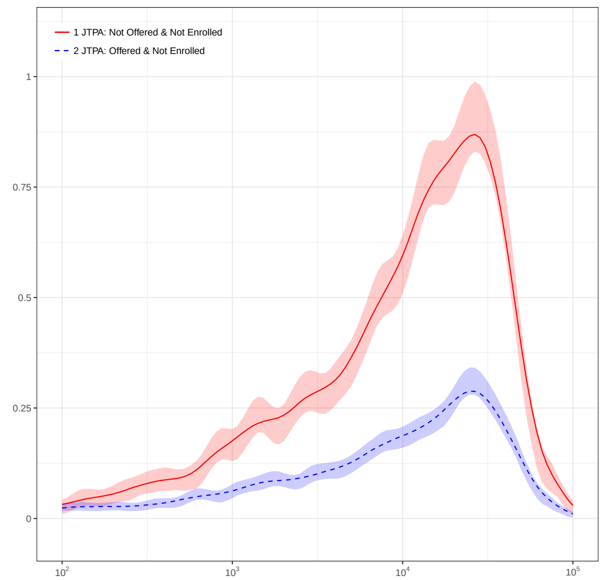
Columns: (i) Full: full sample; (ii) JTPA Offer: whether offered JTPA services; (iii) JTPA Enrollment: whether enrolled in JTPA.
Rows: (i) Income: cumulative income over 30-month period post random selection; (ii) HS or GED: whether has high school degree or GED; (iii) Male: gender being male; (iv) Nonwhite: 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.

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 5, also available are the treatment takeup (JTPA enrollment) and the instrument (JTPA Offer). See Table 2 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 $w_i^{d=1,t=0}$, and then scales down the density estimates by the corresponding conditional probabilities.

Figure 4. 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 et al. (2017a).

We plot the two estimated (scaled) densities in Figure 4. 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 5, 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 eaning 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 5. 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 5, 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".

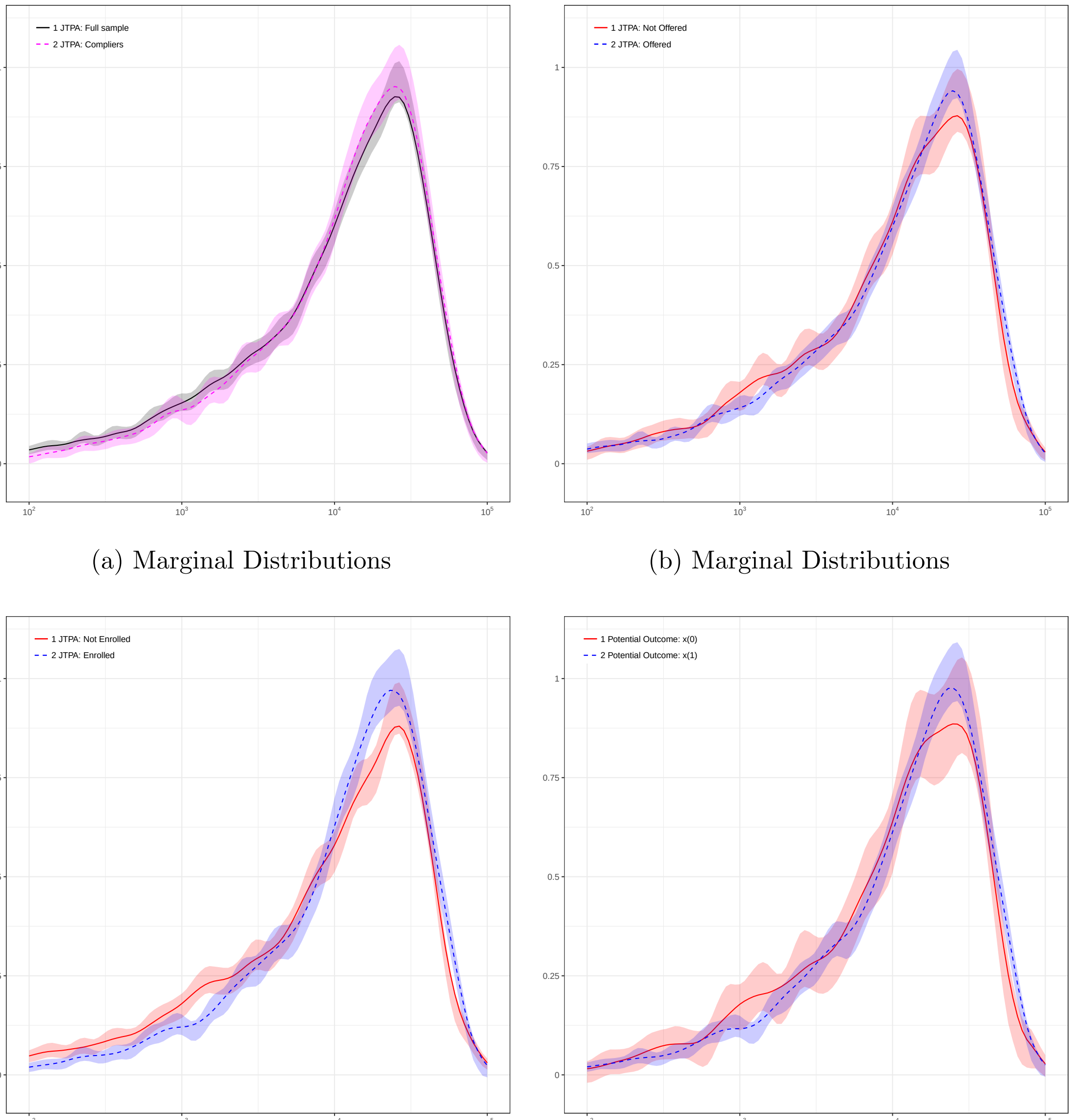
7 Simulation Evidence

We discuss briefly a representative example of the simulation results reported in the Supplemental Appendix, which includes more comprehensive Monte Carlo evidence employing with different data generating process, evaluation point, sample size, and polynomial order, among other features.

We generate i.i.d. sample of size 2,000 from the standard exponential distribution, which has distribution function $F(x) = 1-e^{-x}$ and density e^{-x} with support $[0, \infty)$. We choose two evaluation points, $x = 0$ and 1.5, corresponding to boundary and interior cases, respectively. For estimating density, we use the triangular kernel $K(x) = (1-|x|)\mathbb{1}(|x| \leq 1)$ and local polynomial of order 2. For bandwidth, we consider both a fixed bandwidth grid as multiples of the MSE optimal bandwidth, as well a as data-driven bandwidth estimated from data. Details about our bandwidth selection procedure are also available in the Supplemental Appendix.

We collect the simulation results in Table 3. First, note that the bias increases as the bandwidth gets larger, while variance decreases. This is compatible with classical results in nonparametrics: undersmoothing tend to make the estimator more biased but less volatile. Second, standard errors constructed from Theorem 2 works extremely well. Indeed, average standard errors across simulations (column "mean") match the simulated variability of our estimator (column "sd") almost perfectly. Third, empirical size is well controlled. For fixed bandwidths, the empirical rejection rate is very close to the nominal 5% level, while for estimated bandwidth we have slight over-rejection,

Figure 5. 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 et al. (2017a).

due to the extra variability introduced by estimating bandwidth. Note that we center the test statistic at simulated average of our estimator, to eliminate the impact of smoothing bias.

Table 3. Simulation Results.

(a) $x = 0$ (boundary point)

	$\hat{f}$ bias	$\hat{f}$ sd	$\hat{f}$ $\sqrt{\text{mse}}$	SE mean	SE size
$h_{\texttt{MSE}}\times$					
0.1	0.000	0.187	0.187	0.185	5.36
0.3	0.000	0.103	0.103	0.103	5.54
0.5	−0.007	0.078	0.078	0.077	5.08
0.7	−0.017	0.063	0.065	0.063	4.90
0.9	−0.028	0.054	0.061	0.053	4.92
1	−0.034	0.051	0.061	0.050	4.88
1.1	−0.039	0.048	0.062	0.047	4.90
1.3	−0.052	0.042	0.067	0.041	5.12
1.5	−0.065	0.038	0.075	0.037	5.36
1.7	−0.078	0.035	0.085	0.034	5.86
1.9	−0.091	0.032	0.097	0.031	5.88
$\hat{h}$	−0.031	0.073	0.079	0.064	8.70

(b) $x = 1.5$ (interior point)

	$\hat{f}$ bias	$\hat{f}$ sd	$\hat{f}$ $\sqrt{\text{mse}}$	SE mean	SE size
$h_{\texttt{MSE}}\times$					
0.1	0.001	0.037	0.037	0.036	6.12
0.3	0.001	0.021	0.021	0.020	5.96
0.5	0.002	0.016	0.016	0.016	5.62
0.7	0.003	0.013	0.013	0.013	5.14
0.9	0.005	0.011	0.012	0.011	4.92
1	0.006	0.011	0.012	0.011	5.04
1.1	0.007	0.010	0.012	0.010	5.00
1.3	0.010	0.009	0.013	0.009	4.94
1.5	0.013	0.008	0.015	0.008	4.90
1.7	0.016	0.008	0.018	0.008	4.54
1.9	0.020	0.007	0.021	0.007	4.60
$\hat{h}$	0.005	0.012	0.013	0.011	8.24

Notes: (i) bias: empirical bias of the estimators; (ii) sd: empirical standard deviation of the estimators; (iii) $\sqrt{\text{mse}}$: empirical MSE of the estimators; (iv) mean: empirical average of the estimated standard errors; (v) size: empirical size of testing the hypothesis at nominal 5% level, the test statistic is centered at $\mathbb{E}\hat{f}$. For each simulation, we use sample of size $n = 2,000$, triangular kernel, and local polynomial of order 2. 5,000 Monte Carlo repetitions are used.

Additional simulations for density estimation are available in the Supplemental Appendix, and simulation evidence specialized to the density continuity test are given in our companion paper Cattaneo et al. (2017b).

8 Conclusion

We introduced a new kernel-based density estimator employing local polynomial approximation, which is intuitive, easy to implement and boundary adaptive. It requires choosing only one tuning parameter, and it avoids the need for data transformation (such as pre-binning), additional tuning parameter choices, or other boundary-specific transformations. Furthermore, the estimator can be used directly for all evaluation points on the support of the variable of interest (boundary, near-boundary or interior). From a technical perspective, we developed valid bias and variance approximations, large sample distributional approximations, consistent standard errors, consistent data-driven bandwidth selectors, all these results allowing for root-n estimable weighted distributions. All these results were illustrated with three methodological applications: discontinuity-in-density testing, counterfactual comparisons, and specification testing and compliers heterogeneity analysis in IV settings.

Our methods can also be applied to many other contexts of interest in empirical work, including auctions, bunching, missing data, measurement error, and data combination, just to mention a few more. In fact, our distribution, density and derivatives thereof estimators can be used in any nonparametric and semiparametric setting where these objected need to be estimated. To make our methods as accessible as possible we also provide general purpose software in Stata and R, as described in Cattaneo et al. (2017a,b).

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.

Balke, A., and Pearl, J. (1997), "Bounds on Treatment Effects from Studies with Imperfect Compliance," *Journal of the American Statistical Association*, 92, 1171–1176.

Calonico, S., Cattaneo, M. D., and Farrell, M. H. (2017), "On the Effect of Bias Estimation on Coverage Accuracy in Nonparametric Inference," *Journal of the American Statistical Association*, forthcoming.

Calonico, S., Cattaneo, M. D., and Titiunik, R. (2014), "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs," *Econometrica*, 82, 2295–2326.

Cattaneo, M. D., Jansson, M., and Ma, X. (2017a), "`lpdensity`: Local Polynomial Density Estimation and Inference," working paper, University of Michigan.

——— (2017b), "`rddensity`: Manipulation Testing based on Density Discontinuity," working paper, University of Michigan.

Cattaneo, M. D., Titiunik, R., and Vazquez-Bare, G. (2017c), "Comparing Inference Approaches for RD Designs: A Reexamination of the Effect of Head Start on Child Mortality," *Journal of Policy Analysis and Management*, 36, 643–681.

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

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.

Frandsen, B. (2017), "Party Bias in Union Representation Elections: Testing for Manipulation in the Regression Discontinuity Design When the Running Variable is Discrete," in *Regression Discontinuity Designs: Theory and Applications (Advances in Econometrics, volume 38)*, eds. M. D. Cattaneo and J. C. Escanciano, Emerald Group Publishing, pp. 281–315.

Heckman, J. J., and Vytlacil, E. J. (2005), "Structural Equations, Treatment Effects and Econometric Policy Evaluation," *Econometrica*, 73, 669–738.

Ichimura, H., and Todd, P. E. (2007), "Implementing Nonparametric and Semiparametric Estimators," in *Handbook of Econometrics*, eds. J. Heckman and E. Leamer, Vol. 6B of *Handbook of Econometrics*, chapter 74, Elsevier.

Imbens, G. W., and Rubin, D. B. (2015), *Causal Inference in Statistics, Social, and Biomedical Sciences*, Cambridge University Press.

Karunamuni, R., and Albert, T. (2005), "On Boundary Correction in Kernel Density Estimation," *Statistical Methodology*, 2, 191–212.

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

Ludwig, J., and Miller, D. L. (2007), "Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design," *Quarterly Journal of Economics*, 122, 159–208.

McCrary, J. (2008), "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test," *Journal of Econometrics*, 142, 698–714.

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

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.

Otsu, T., Xu, K.-L., and Matsushita, Y. (2014), "Estimation and Inference of Discontinuity in Density," *Journal of Business and Economic Statistics*, 31, 507–524.

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

Zhang, S., and Karunamuni, R. J. (1998), "On Kernel Density Estimation Near Endpoints," *Journal of Statistical Planning and Inference*, 70, 301–316.

Simple Local Polynomial Density Estimators
Supplemental Appendix

Matias D. Cattaneo[*] Michael Jansson[†] Xinwei Ma[‡]

September 27, 2017

Abstract

This Supplemental Appendix contains general theoretical results encompassing those discussed in the main paper, includes proofs of those general results, discusses additional methodological and technical results, and reports detailed simulation evidence.

[*]Department of Economics, Department of Statistics, University of Michigan.
[†]Department of Economics, UC Berkeley and *CREATES*.
[‡]Department of Economics, Department of Statistics, University of Michigan.

Contents

1 Setup

We repeat the setup in the main paper for completeness. Recall that $\{x_i\}_{1\leq i\leq n}$ is a random sample from the distribution G, supported on $\mathcal{X} = [x_{\mathtt{L}}, x_{\mathtt{U}}]$. Note that it is possible to have $x_{\mathtt{L}} = -\infty$ and/or $x_{\mathtt{U}} = \infty$, and we only need the extra requirement that $\mathbb{P}[x_i = \pm\infty] = 0$ so that G is a tight distribution. We will assume both $x_{\mathtt{L}}$ and $x_{\mathtt{U}}$ are finite, to facilitate discussion on boundary estimation issues. Since the method we propose is local in nature, whether or not G has bounded support is not relevant, and we introduce the two "end points" $x_{\mathtt{L}}$ and $x_{\mathtt{U}}$ to simplify notation and discussions later.

Without loss of generality, we assume there is a companion set of weights $\{w_i\}_{1\leq i\leq n}$, such that $F(u) := \mathbb{E}[\mathbb{1}[x_i \leq u]w_i]$ is a well-defined distribution function. Detailed assumptions are postponed to a later section. Note that when $w_i \equiv 1$, F reduces to G. We also allow the weights to be estimated, and discussions thereof is also postponed.

Define the empirical distribution function (hereafter e.d.f.)

$$\hat{F}(u) = \sum_i w_i \mathbb{1}[x_i \leq u] \Big/ \sum_i w_i,$$

and summations are understood as from 1 to n, unless otherwise specified. $\hat{F}$ has appealing properties such as it is 0 below the first order statistic, and 1 above the largest one.

Remark 1 (Alternative: $\tilde{F}$). In the main paper we used another specification of the e.d.f., as

$$\tilde{F}(u) = \frac{1}{n}\sum_i w_i \mathbb{1}[x_i \leq u].$$

The difference between $\hat{F}$ and $\tilde{F}$ is the scaling factor, and is negligible for most purposes. We note, however, that there are some subtle differences.

First, $\tilde{F}$, viewed as a process, does not converge to a Brownian bridge unless $w_i = 1$. To see this, simply plugin $u = x_{\mathtt{U}}$, leading to $\tilde{F}(x_{\mathtt{U}}) = \sum_i w_i/n$ which has nondegenerate distribution asymptotically. If one is interested in nonparametric estimates such as density or further derivatives, using both $\hat{F}$ and $\tilde{F}$ are fine. The "asymmetry" in $\tilde{F}$ will only affect the estimated intercept in our local polynomial regression.

The major difference between $\hat{F}$ and $\tilde{F}$ emerges when the weights do not sum up to 1, i.e. $0 < \mathbb{E}[w_i] < 1$. To see this, consider the density test example introduced in the main paper. There the object of interest is the density at one point, $\bar{x}$, estimated from left (or right). Consider the weights $w_i = \mathbb{1}[x_i < \bar{x}]$, which effectively restricts to the subsample to the left of cutoff. Then $\tilde{F}$ constructed from the weights is *not* a proper distribution function, since it starts from 0 and reaches maximum $\sum_i \mathbb{1}[x_i \leq \bar{x}]/n$ at the cutoff, hence the density estimated thereof is not proper, as does not integrates to 1. On the other hand, using $\hat{F}$ will give a proper density. The difference between those two densities is simply a scaling factor.

In the main paper we use $\tilde{F}$ to simplify exposition and discussion, while in this Supplemental Appendix we use $\hat{F}$ to develop the general theory, as it has better mathematical properties. We note that using either will deliver asymptotically equivalent nonparametric estimates.

Given $p \in \mathbb{N}$, our local polynomial distribution estimator is defined as

$$\hat{\boldsymbol{\beta}}_p(x) = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}} \sum_i \left(\hat{F}(x_i) - \mathbf{r}_p(x_i - x)'\boldsymbol{\beta}\right)^2 K\left(\frac{x_i - x}{h}\right),$$

where $\mathbf{r}_p(u) = [1, u, u^2, \cdots, u^p]$ is a (one-dimensional) polynomial expansion; K is a kernel function whose properties are to be specified later; and $h = h_n$ is a bandwidth sequence. The estimator, $\hat{\boldsymbol{\beta}}_p(x)$, is motivated as a local Taylor series expansion, hence the target parameter is (i.e. the population counterpart, assuming exists)

$$\boldsymbol{\beta}_p(x) = \left[\frac{1}{0!}F(x),\ \frac{1}{1!}F^{(1)}(x),\ \cdots,\ \frac{1}{p!}F^{(p)}(x)\right]'.$$

Therefore, we also write

$$\hat{\boldsymbol{\beta}}_p(x) = \left[\frac{1}{0!}\hat{F}_p(x),\ \frac{1}{1!}\hat{F}_p^{(1)}(x),\ \cdots,\ \frac{1}{p!}\hat{F}_p^{(p)}(x)\right]',$$

or equivalently, $\hat{F}_p^{(v)} = v!\mathbf{e}_v'\hat{\boldsymbol{\beta}}_p(x)$, provided that $v \le p$, and $\mathbf{e}_v$ is the $(v+1)$-th unit vector of $\mathbb{R}^{p+1}$. We also use $f = F^{(1)}$ to denote the corresponding probability density function (hereafter p.d.f) for convenience.

The estimator has the following matrix form, which we will utilize:

$$\hat{\boldsymbol{\beta}}_p(x) = \mathbf{H}^{-1}\left(\frac{1}{n}\mathbf{X}_h'\mathbf{K}_h\mathbf{X}_h\right)^{-1}\left(\frac{1}{n}\mathbf{X}_h'\mathbf{K}_h\mathbf{Y}\right),$$

where

$$\mathbf{X}_h = \left[\left(\frac{x_i - x}{h}\right)^j\right]_{1\le i\le n,\ 0\le j\le p},$$

$\mathbf{K}_h$ is a diagonal matrix collecting $\{h^{-1}K((x_i - x)/h)\}_{1\le i\le n}$, and $\mathbf{Y}$ is a column vector collecting $\{\hat{F}(x_i)\}_{1\le i\le n}$. We also use the convention $K_h(u) = h^{-1}K(u/h)$.

Before giving an overview of our results, we make a short digression on definition of boundary regions. Boundary region is defined as $[x_{\mathtt{L}}, x_{\mathtt{L}} + h) \cup (x_{\mathtt{U}} - h, x_{\mathtt{U}}]$, and the two segments are called lower and upper boundaries, respectively. As the bandwidth vanishes as the sample size n increases, boundary region is really a finite sample concept. To facilitate discussion on boundary issues, it is common to consider a drifting sequence of evaluation points, $x = x_{\mathtt{L}} + ch$ with $0 \le c < 1$ or $x = x_{\mathtt{U}} - ch$. We call such evaluation points in the lower and upper boundary region, respectively. Therefore we allow the evaluation point x to depend on h (hence implicitly n), but do not make it explicit to conserve notation. Remarks will be made when it is crucial to distinguish whether x is fixed or a drifting sequence.

Remark 2 (More general notion of interior points). We assumed the support of the sample being a (possibly unbounded) line segment in $\mathbb{R}$ purely for notational convenience. Assume the support is a general measurable set $\mathcal{X} \subset \mathbb{R}$, interior points are then $\{x \in \mathcal{X} :\ B(x, h) \subset \mathcal{X}\}$, where $B(x, h) = \{y \in \mathbb{R} :\ |y - x| < h\}$. We don't find this level of generality very useful, but note all our results easily adapt.

1.1 Additional Notation

In this Supplemental Appendix, we use n to denote sample size, and limits are taken with $n \to \infty$, unless otherwise specified. Euclidean norms are denoted by $|\cdot|$, and other norms will be defined at their first appearances.

For sequence of numbers (or random variables), $a_n \precsim b_n$ implies $\limsup_n |a_n/b_n|$ is finite, and $a_n \asymp b_n$ implies both directions. The notation $a_n \precsim_{\mathbb{P}} b_n$ is used to denote that $|a_n/b_n|$ is asymptotically tight: $\limsup_{\varepsilon\uparrow\infty} \limsup_n \mathbb{P}[|a_n/b_n| \geq \varepsilon] = 0$. $a_n \asymp_{\mathbb{P}} b_n$ implies both $a_n \precsim_{\mathbb{P}} b_n$ and $b_n \precsim_{\mathbb{P}} a_n$. When b_n is a sequence of nonnegative numbers, $a_n = O(b_n)$ is sometimes used for $a_n \precsim b_n$, so does $a_n = O_{\mathbb{P}}(b_n)$.

For probabilistic convergence, we use $\to_{\mathbb{P}}$ for convergence in probability and $\rightsquigarrow$ for weak convergence (convergence in distribution). Standard normal distribution is denoted as $\mathcal{N}(0,1)$, with c.d.f. Φ and p.d.f. ϕ.

throughout, we use C to denote generic constant factor which does not depend on sample size. The exact value can change in different contexts.

1.2 Overview of Main Results

In this subsection, we give an overview of our results, including a (first order) mean squared error (hereafter m.s.e.) expansion, and asymptotic normality. Fix some $v \geq 1$ and p, we have the following:

$$\left|\hat{F}_p^{(v)}(x) - F^{(v)}(x)\right| = O_{\mathbb{P}}\left(h^{p+1-v}\mathscr{B}_{p,v,x} + h^{p+2-v}\tilde{\mathscr{B}}_{p,v,x} + \sqrt{\frac{1}{nh^{2v-1}}\mathscr{V}_{p,v,x}}\right).$$

The previous result gives m.s.e. expansion for nonparametric derivative estimators, $1 \leq v \leq p$, but not for $v = 0$. With $v = 0$, $\hat{F}_p(x)$ is essentially a smoothed e.d.f., which estimates the c.d.f. $F(x)$. Since $F(x)$ is $\sqrt{n}$-estimable, it should be expected that it has very different properties compared to the nonparametric components. Indeed, we have

$$\left|\hat{F}_p(x) - F(x)\right| = O_{\mathbb{P}}\left(h^{p+1}\mathscr{B}_{p,0,x} + h^{p+2}\tilde{\mathscr{B}}_{p,0,x} + \sqrt{\frac{1}{n}\mathscr{V}_{p,0,x}}\right).$$

There is another complication, however, when x is in the boundary region. For a drifting sequence x in the boundary region, the e.d.f. $\hat{F}(x)$ is "super-consistent" in the sense that it converges at rate $\sqrt{h/n}$. The reason is that when x is near $x_{\mathtt{L}}$ or $x_{\mathtt{U}}$, $\hat{F}(x)$ is essentially estimating 0 or 1, and the variance, $F(x)(1-F(x))$ vanishes asymptotically, giving rise to the additional factor $\sqrt{h}$. This is shared by our estimator: for $v = 0$ and x in the boundary region, the c.d.f. estimator $\hat{F}_p(x)$ is super-consistent, with $\mathscr{V}_{p,0,x} \asymp h$.

Also note that for the m.s.e. expansion, we provide not only the first order bias, but also the second order bias. We will only use the second order bias for bandwidth selection, since it is well-known that in some cases the first order bias can vanish.

The m.s.e. expansion provides rate of convergence of our estimators. The following shows that, under suitable regularity conditions, they are also asymptotically normal. Again first consider $v \geq 1$.

$$\sqrt{nh^{2v-1}}\left(\hat{F}_p^{(v)}(x) - F^{(v)}(x) - h^{p+1-v}\mathscr{B}_{p,v,x}\right) \rightsquigarrow \mathcal{N}\left(0, \mathscr{V}_{p,v,x}\right),$$

provided that the bandwidth is not too large, so that after scaling, the remaining bias does not explode. For $v = 0$, i.e. the smoothed e.d.f., we have

$$\sqrt{\frac{n}{\mathscr{V}_{p,0,x}}}\left(\hat{F}_p(x) - F(x) - h^{p+1}\mathscr{B}_{p,0,x}\right) \rightsquigarrow \mathcal{N}\left(0, 1\right),$$

where we moved the variance $\mathscr{V}_{p,0,x}$ as a scaling factor in the above display, to encompass the situation where x lies in boundary region (recall from the previous subsection that in this case the scaling factor has order $\sqrt{n/h}$).

1.3 Some Matrices

In this subsection we collect some matrices which will be used throughout this Supplemental Appendix. They show up in asymptotic results as components of bias and variance. Recall that x can be either a fixed point or a drifting sequence, and for the latter, it takes the form $x = x_{\mathtt{L}} + ch$ or $x = x_{\mathtt{U}} - ch$ for some $c \in [0, 1)$.

$$\begin{aligned}
\mathbf{S}_{p,x} &= \int_{h^{-1}(\mathcal{X}-x)} \mathbf{r}_p(u)\mathbf{r}_p(u)'K(u)\mathrm{d}u &&= \int_{\frac{x_{\mathtt{L}}-x}{h}}^{\frac{x_{\mathtt{U}}-x}{h}} \mathbf{r}_p(u)\mathbf{r}_p(u)'K(u)\mathrm{d}u,\\
\mathbf{c}_{p,x} &= \int_{h^{-1}(\mathcal{X}-x)} \mathbf{r}_p(u)u^{p+1}K(u)\mathrm{d}u &&= \int_{\frac{x_{\mathtt{L}}-x}{h}}^{\frac{x_{\mathtt{U}}-x}{h}} \mathbf{r}_p(u)u^{p+1}K(u)\mathrm{d}u,\\
\tilde{\mathbf{c}}_{p,x} &= \int_{h^{-1}(\mathcal{X}-x)} \mathbf{r}_p(u)u^{p+2}K(u)\mathrm{d}u &&= \int_{\frac{x_{\mathtt{L}}-x}{h}}^{\frac{x_{\mathtt{U}}-x}{h}} \mathbf{r}_p(u)u^{p+1}K(u)\mathrm{d}u,\\
\tilde{\mathbf{c}}_{p,x} &= \int_{h^{-1}(\mathcal{X}-x)} \mathbf{r}_p(u)u^{p+2}K(u)\mathrm{d}u &&= \int_{\frac{x_{\mathtt{L}}-x}{h}}^{\frac{x_{\mathtt{U}}-x}{h}} \mathbf{r}_p(u)u^{p+2}K(u)\mathrm{d}u,\\
\mathbf{\Gamma}_{p,x} &= \iint_{h^{-1}(\mathcal{X}-x)} (u \wedge v)\mathbf{r}_p(u)\mathbf{r}_p(v)K(u)K(v)\mathrm{d}u\mathrm{d}v &&= \iint_{\frac{x_{\mathtt{L}}-x}{h}}^{\frac{x_{\mathtt{U}}-x}{h}} (u \wedge v)\mathbf{r}_p(u)\mathbf{r}_p(v)K(u)K(v)\mathrm{d}u\mathrm{d}v,\\
\mathbf{T}_{p,x} &= \int_{h^{-1}(\mathcal{X}-x)} \mathbf{r}_p(u)\mathbf{r}_p(u)'K(u)^2\mathrm{d}u &&= \int_{\frac{x_{\mathtt{L}}-x}{h}}^{\frac{x_{\mathtt{U}}-x}{h}} \mathbf{r}_p(u)\mathbf{r}_p(u)'K(u)^2\mathrm{d}u,
\end{aligned}$$

where $h^{-1}(\mathcal{X} - x) = \{h^{-1}(y - x) : \ y \in \mathcal{X}\}$. Later we will assume the kernel function K being supported on $[-1, 1]$, hence with bandwidth $h \downarrow 0$, the region of integration in the above display can be replaced by

x	$(x_{\mathtt{L}} - x)/h$	$(x_{\mathtt{U}} - x)/h$
x interior	-1	$+1$
$x = x_{\mathtt{L}} + ch$ in lower boundary	$-c$	$+1$
$x = x_{\mathtt{U}} - ch$ in upper boundary	-1	$+c$

Since we do not allow $x_{\mathtt{L}} = x_{\mathtt{U}}$, no drifting sequence x can be in both boundary regions, at least asymptotically.

1.4 Assumptions

In this section we give detailed assumptions supporting results, including preliminary lemmas and our main results. Other specific assumptions will be given in corresponding sections.

Let $\mathcal{O}$ be a subset of Euclidean space with nonempty interior, $\mathcal{C}^s(\mathcal{O})$ denotes functions that are at least s-times continuously differentiable in the interior of $\mathcal{O}$, and that the derivatives can be continuously extended to the boundary of $\mathcal{O}$.

Assumption 1 (DGP).
(i) $\{x_i, w_i\}_{1\leq i\leq n}$ *is a random sample.*
(ii) x_i *has support* $\mathcal{X} = [x_{\mathtt{L}}, x_{\mathtt{U}}]$ *with* $x_{\mathtt{U}} > x_{\mathtt{L}}$ *and distribution* G. *Further,* $G \in \mathcal{C}^{\alpha_x}(\mathcal{X})$.
(iii) Let $w_s(x) = \mathbb{E}[w_i^s|x_i = x]$, *and* $w_s \in \mathcal{C}^{\alpha_{w,s}}(\mathcal{X})$.
(iv) $\mathbb{E}[w_i] = 1$ *and* $\mathbb{E}[w_i|x_i]$ *is nonnegative almost surely.*

Part (i) is standard. Part (ii) and (iii) together implies smoothness of F. Part (iv) ensures that F is a proper distribution function. To see this, note that $F(x_{\mathtt{U}}) = \mathbb{E}[w_i] = 1$, and for any Borel subset A, $F(A) = \int_A w_1(x)\mathrm{d}G(x)$, hence by construction F is absolute continuous with respect to G, and $w_1(x) \geq 0$ almost surely implies F is a positive measure. For notational convenience, we use $w(\cdot) = w_1(\cdot)$.

Technically, part (iv) is not essential for our theory. It is possible to drop this assumption entirely, then the object of interest will be a general Radon-Nikodym derivative (and derivatives thereof) that can be negative.

Assumption 2 (Kernel).
The kernel function $K(\cdot)$ *is nonnegative, symmetric, and belongs to* $\mathcal{C}^0([-1,1])$. *Further, it integrates to one:* $\int_{\mathbb{R}} K(u)\mathrm{d}u = 1$.

Assumption 2 is standard in nonparametric estimation, and is satisfied for common kernel functions. We exclude kernels with unbounded support (for example the Gaussian kernel) for simplicity, since such kernels will always hit boundaries. Our results, however, remain to hold with careful analysis, albeit the notation becomes more cumbersome.

Also note that if we simply have $\int_{\mathbb{R}} K(u)\mathrm{d}u > 0$, i.e. the last part of the previous assumption is violated, we can simply redefine $\tilde{K}(u) = K(u)/\int_{\mathbb{R}} K(u)\mathrm{d}u$. This is not essential since least squares is invariant to multiplicative scaling.

Assumption 3 (Positive density).
$G^{(1)}(x) > 0$ *for* $x \in \mathcal{X}$.

Technically, we do not need the density to be positive for all the support $\mathcal{X}$. Since all our results are local in nature, it suffices to have $G^{(1)}(x) > 0$ for the evaluation point (hence strictly positive in a neighborhood by continuity). We also use g to denote the density $G^{(1)}$ just to follow conventions.

2 Large Sample Properties with Observed Weights

2.1 Preliminary Lemmas

We first consider the object $\mathbf{X}_h'\mathbf{K}_h\mathbf{X}_h/n$

Lemma 1. *Assume Assumptions 1–3 hold with $\alpha_x \geq 1$. Further $h \to 0$ and $nh \to \infty$. Then*

$$\frac{1}{n}\mathbf{X}_h'\mathbf{K}_h\mathbf{X}_h = g(x)\mathbf{S}_{p,x} + o(1) + O_{\mathbb{P}}\left(1/\sqrt{nh}\right).$$

Note that with Lemma 1, the quantity $\mathbf{X}_h'\mathbf{K}_h\mathbf{X}_h/n$ is asymptotically invertible. Since the density $g(x)$ enters as a multiplicative factor, it also shows why we need Assumption 3. Also note that this result covers both interior x and boundary x. And depending on the nature of x, the exact form of $\mathbf{S}_{p,x}$ differs.

With simple algebra, we have

$$\hat{\boldsymbol{\beta}}_p(x) - \boldsymbol{\beta}_p(x) = \mathbf{H}^{-1}\left(\frac{1}{n}\mathbf{X}_h'\mathbf{K}_h\mathbf{X}_h\right)^{-1}\left(\frac{1}{n}\mathbf{X}_h'\mathbf{K}_h(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_p(x))\right),$$

and the following gives a further decomposition of the "numerator".

$$\begin{aligned}
\frac{1}{n}\mathbf{X}_h'\mathbf{K}_h(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_p(x)) &= \frac{1}{n}\sum_i \mathbf{r}_p\left(\frac{x_i - x}{h}\right)\left(\hat{F}(x_i) - \mathbf{r}_p(x_i - x)'\boldsymbol{\beta}_p(x)\right)K_h(x_i - x) \\
&= \frac{1}{n}\sum_i \mathbf{r}_p\left(\frac{x_i - x}{h}\right)\left(F(x_i) - \mathbf{r}_p(x_i - x)'\boldsymbol{\beta}_p(x)\right)K_h(x_i - x) \\
&+ \int_{\frac{x_{\mathrm{L}} - x}{h}}^{\frac{x_{\mathrm{U}} - x}{h}} \mathbf{r}_p(u)\left(\hat{F}(x + hu) - F(x + hu)\right)K(u)g(x + hu)\mathrm{d}u \\
&+ \frac{1}{n}\sum_i \mathbf{r}_p\left(\frac{x_i - x}{h}\right)\left(\hat{F}(x_i) - F(x_i)\right)K_h(x_i - x) - \int_{\frac{x_{\mathrm{L}} - x}{h}}^{\frac{x_{\mathrm{U}} - x}{h}} \mathbf{r}_p(u)\left(\hat{F}(x + hu) - F(x + hu)\right)K(u)g(x + hu)\mathrm{d}u.
\end{aligned}$$

The first part represents the smoothing bias, and the second part can be analyzed as a sample average, which will be given in a lemma. The real difficulty comes from the third term, which can have nonnegligible (first order) contribution. We give it a further decomposition:

$$\begin{aligned}
\frac{1}{n}\sum_i \mathbf{r}_p\left(\frac{x_i - x}{h}\right)\left(\hat{F}(x_i) - F(x_i)\right)K_h(x_i - x) &= \frac{1 + o_{\mathbb{P}}(1)}{n^2}\sum_{i,j}\mathbf{r}_p\left(\frac{x_i - x}{h}\right)w_j\left(\mathbb{1}[x_j \leq x_i] - F(x_i)\right)K_h(x_i - x) \\
= \frac{1 + o_{\mathbb{P}}(1)}{n^2}\sum_i \mathbf{r}_p\left(\frac{x_i - x}{h}\right)w_i\left(1 - F(x_i)\right)K_h(x_i - x) &+ \frac{1 + o_{\mathbb{P}}(1)}{n^2}\sum_{i,j;i\neq j}\mathbf{r}_p\left(\frac{x_i - x}{h}\right)w_j\left(\mathbb{1}[x_j \leq x_i] - F(x_i)\right)K_h(x_i - x),
\end{aligned}$$

hence we have the final decomposition:

$$
\begin{aligned}
&\frac{1}{n}\sum_i \mathbf{r}_p\left(\frac{x_i-x}{h}\right)\Big(\hat{F}(x_i)-\mathbf{r}_p(x_i-x)'\boldsymbol{\beta}_p(x)\Big)K_h(x_i-x)\\
&\quad= \frac{1}{n}\sum_i \mathbf{r}_p\left(\frac{x_i-x}{h}\right)\Big(F(x_i)-\mathbf{r}_p(x_i-x)'\boldsymbol{\beta}_p(x)\Big)K_h(x_i-x) && \text{(smoothing bias } \hat{\mathbf{B}}_{\mathtt{S}})\\
&\quad+ (1+o_{\mathbb{P}}(1))\int_{\frac{x_{\mathtt{L}}-x}{h}}^{\frac{x_{\mathtt{U}}-x}{h}} \mathbf{r}_p(u)\Big(\hat{F}(x+hu)-F(x+hu)\Big)K(u)g(x+hu)\mathrm{d}u && \text{(linear variance } \hat{\mathbf{L}})\\
&\quad+ \frac{1+o_{\mathbb{P}}(1)}{n^2}\sum_i \mathbf{r}_p\left(\frac{x_i-x}{h}\right)w_i\Big(1-F(x_i)\Big)K_h(x_i-x) && \text{(leave-in bias } \hat{\mathbf{B}}_{\mathtt{LI}})\\
&\quad+ \frac{1+o_{\mathbb{P}}(1)}{n^2}\sum_{i,j;i\neq j}\Big\{\mathbf{r}_p\left(\frac{x_i-x}{h}\right)w_j\Big(\mathbb{1}[x_j\le x_i]-F(x_i)\Big)K_h(x_i-x)\\
&\qquad -\mathbb{E}\Big[\mathbf{r}_p\left(\frac{x_i-x}{h}\right)w_j\Big(\mathbb{1}[x_j\le x_i]-F(x_i)\Big)K_h(x_i-x)\Big|x_j,w_j\Big]\Big\}. && \text{(quadratic variance } \hat{\mathbf{R}})
\end{aligned}
$$

Now it becomes clear that for the estimator $\hat{\boldsymbol{\beta}}_p(x)$, it consists the following parts: (i) smoothing bias; (ii) linear influence function; (iii) leave-in bias; (iv) second order degenerate U-statistic. To provide intuition for the previous decomposition, the smoothing bias is a typical feature of nonparametric estimators; leave-in bias occurs since each observation is used twice, in constructing the e.d.f. $\hat{F}$, and as a design point (that is, $\hat{F}$ has to be evaluated at x_i); finally the second order U-statistic shows up since the "dependent variable", $\mathbf{Y}$, is estimated, so that a double sum is involved.

We first handle the two biases.

Lemma 2. *Assume Assumptions 1–3 hold with $\alpha_x \geq p+1$, $\alpha_w \geq p$ and $\alpha_{w,2}\geq 0$. Further $h\to 0$ and $nh\to\infty$. Then*

$$
\hat{\mathbf{B}}_{\mathtt{S}} = h^{p+1}\frac{F^{(p+1)}(x)g(x)}{(p+1)!}\mathbf{c}_{p,x} + o_{\mathbb{P}}(h^{p+1}), \qquad \hat{\mathbf{B}}_{\mathtt{LI}} = O_{\mathbb{P}}\left(n^{-1}\right).
$$

By imposing additional smoothness, it is also possible to characterize the next term in the smoothing bias, which has order h^{p+2}. Since that result is only used for bandwidth selection when the leading bias vanishes, we do not report it here.

Next we consider the "influence function" part, $\hat{\mathbf{L}}$. This term is crucial in the sense that (under suitable conditions such that $\hat{\mathbf{R}}$ is negligible) it determines the asymptotic variance of our estimator, and with correct scaling, it is asymptotically normally distributed.

Lemma 3. *Assume Assumptions 1–3 hold with $\alpha_x \geq 2$, $\alpha_w \geq 1$, $\alpha_{w,2}\geq 0$, and $\mathbb{E}[w_i^4]<\infty$. Further $h\to 0$ and $nh\to\infty$. Define the scaling matrix*

$$
\mathbf{N}_x = \begin{cases} \mathrm{diag}\Big\{1,\quad h^{-1/2},\ h^{-1/2},\ \cdots,\ h^{-1/2}\Big\} & x \textit{ interior},\\ \mathrm{diag}\Big\{h^{-1/2},\ h^{-1/2},\ h^{-1/2},\ \cdots,\ h^{-1/2}\Big\} & x \textit{ boundary}, \end{cases}
$$

then

$$
\sqrt{n}\mathbf{N}_x\Big[g(x)\mathbf{S}_{p,x}\Big]^{-1}\hat{\mathbf{L}} \rightsquigarrow \mathcal{N}(\mathbf{0},\ \mathsf{V}_{p,x}),
$$

with

$$\mathsf{V}_{p,x} = \begin{cases} \Big(H(x) - 2H(x)F(x) + H(x_{\mathtt{U}})F(x)^2\Big)\mathbf{e}_0\mathbf{e}_0' + H^{(1)}(x)(\mathbf{I} - \mathbf{e}_0\mathbf{e}_0')\mathbf{S}_{p,x}^{-1}\mathbf{\Gamma}_{p,x}\mathbf{S}_{p,x}^{-1}(\mathbf{I} - \mathbf{e}_0\mathbf{e}_0') & x \text{ interior} \\ H^{(1)}(x)\left(\mathbf{S}_{p,x}^{-1}\mathbf{\Gamma}_{p,x}\mathbf{S}_{p,x}^{-1} + c\mathbf{e}_0\mathbf{e}_0'\right) & x = x_{\mathtt{L}} + ch \\ H^{(1)}(x)\left(\mathbf{S}_{p,x}^{-1}\mathbf{\Gamma}_{p,x}\mathbf{S}_{p,x}^{-1} + c\mathbf{e}_0\mathbf{e}_0' - (\mathbf{e}_1\mathbf{e}_0' + \mathbf{e}_0\mathbf{e}_1')\right) & x = x_{\mathtt{U}} - ch. \end{cases}$$

$H(u) := \mathbb{E}[w_i^2 \mathbb{1}[x_i \le u]]$.

The scaling matrix depends on whether the evaluation point is located in the interior or boundary, which is a unique feature of our estimator. To see the intuition, consider an interior point x, and recall that the first element of $\hat{\boldsymbol{\beta}}_p(x)$ is the smoothed e.d.f. Since the distribution function is $\sqrt{n}$-estimable, its property is very different from the rest of $\hat{\boldsymbol{\beta}}_p(x)$, which are nonparametric in nature. Indeed, let $w_i \equiv 1$, then $F = G = H$, and the first component of the variance becomes $G(x)(1 - G(x)) = F(x)(1 - F(x))$, which is the variance of the standard e.d.f. Furthermore, the smoothed e.d.f. $\hat{F}_p(x)$ is asymptotically independent of the rest of $\hat{\boldsymbol{\beta}}_p(x)$.

When x is either in the lower or upper boundary region, $\hat{F}_p(x)$ essentially estimates 0 or 1, respectively, hence it is super-consistent in the sense that it converges even faster than $1/\sqrt{n}$. In this case, the leading $1/\sqrt{n}$-variance vanishes, and higher order residual noise dominates, which makes $\hat{F}_p(x)$ no longer independent of other nonparametric estimates, justifying the formula of boundary evaluation points.

It is tempting to estimate the variance $\mathsf{V}_{p,x}$ in a plug-in manner, where unknown objects H, $H^{(1)}$ and F are replaced with estimates. This is feasible, and can be appealing if $w_i \equiv 1$, which forces H to be the distribution function and $H^{(1)}$ the density. In general, however, a plug-in estimator for $\mathsf{V}_{p,x}$ requires estimating the nuisances functions H and $H^{(1)}$ nonparametrically. Later we will propose a fully data-driven and design adaptive estimator, which does not require estimating H and $H^{(1)}$ explicitly.

Finally we consider the second order U-statistic component.

Lemma 4. *Assume Assumptions 1–3 hold with $\alpha_x \ge 1$, $\alpha_w \ge 0$, and $\alpha_{w,2} \ge 0$. Further $h \to 0$ and $nh \to \infty$. Then*

$$\mathbb{V}[\hat{\mathbf{R}}] = \frac{2}{n^2 h} g(x)\left[H(x) - 2H(x)F(x) + H(x_{\mathtt{U}})F(x)^2\right]\mathbf{T}_{p,x} + O(n^{-2}).$$

In particular, when x is in the boundary region, the above has order $O(n^{-2})$.

2.2 Main Results

In this section we provide two main results, one on asymptotic normality, and the other on standard error.

Theorem 1 (Asymptotic Normality). *Assume Assumptions 1–3 hold with $\alpha_x \ge p+1$, $\alpha_w \ge p$, $\alpha_{w,2} \ge 0$ for some integer $p \ge 0$, and $\mathbb{E}[w_i^4] < \infty$. Further $h \to 0$, $nh^2 \to \infty$ and $nh^{2p+1} = O(1)$.*

Then

$$\sqrt{nh^{2v-1}}\Big(\hat{F}_p^{(v)}(x) - F^{(v)}(x) - h^{p+1-v}\mathscr{B}_{p,v,x}\Big) \rightsquigarrow \mathcal{N}\Big(0,\ \mathscr{V}_{p,v,x}\Big), \qquad 1 \le v \le p,$$
$$\sqrt{\frac{n}{\mathscr{V}_{p,0,x}}}\Big(\hat{F}_p(x) - F(x) - h^{p+1}\mathscr{B}_{p,0,x}\Big) \rightsquigarrow \mathcal{N}\Big(0,\ 1\Big).$$

The constants are

$$\mathscr{B}_{p,v,x} = v!\frac{F^{(p+1)}(x)}{(p+1)!}\mathbf{e}_v'\mathbf{S}_{p,x}^{-1}\mathbf{c}_{p,x},$$

and

$$\mathscr{V}_{p,v,x} = \begin{cases} (v!)^2 H^{(1)}(x)\mathbf{e}_v'\mathbf{S}_{p,x}^{-1}\mathbf{\Gamma}_{p,x}\mathbf{S}_{p,x}^{-1}\mathbf{e}_v & 1 \le v \le p \\ H(x) - 2H(x)F(x) + H(x_{\mathtt{U}})F(x)^2 & v = 0,\ x \text{ interior} \\ hH^{(1)}(x)\left(\mathbf{e}_0'\mathbf{S}_{p,x}^{-1}\mathbf{\Gamma}_{p,x}\mathbf{S}_{p,x}^{-1}\mathbf{e}_0 + c\right) & v = 0,\ x = x_{\mathtt{L}} + ch \text{ or } x_{\mathtt{U}} - ch. \end{cases}$$

Remark 3 (On $nh^{2p+1} = O(1)$). This condition ensures that higher order bias, after scaling, is asymptotically negligible.

Remark 4 (On $nh^2 \to \infty$). This condition ensures that the second order U-statistic, $\hat{\mathbf{R}}$, has smaller order compared to $\hat{\mathbf{L}}$. Note that this condition can be dropped for boundary x or when the parameter of interest is $\hat{F}_p$, the smoothed e.d.f.

Remark 5 (On $\mathscr{V}_{p,0,x}$). One might be tempted to conclude that the variance formula has a discontinuity in x for the smoothed e.d.f. (i.e. $v = 0$), when x switches from interior to boundary. This phenomenon, however, is purely an artifact of different asymptotic frameworks. To see this, assume $x_{\mathtt{L}} = 0$ and $x_{\mathtt{U}} = 1$, and for some sample the bandwidth $h = 0.2$ is used. Given our convention, the point $x = 0.3$ is not a boundary point, hence we should consider $\sqrt{n}$ as the correct scaling for $\hat{F}_p(0.3)$.

On the other hand, one can also consider 0.3 as part of the asymptotic sequence $x = 1.5h$, in which case one promises to move the evaluation point closer to the lower boundary as sample size increases. Then despite the fact that such x is not a boundary point, $\hat{F}_p(x)$ is still an estimator of zero, which means it is super consistent and the correct scaling is $\sqrt{n/h}$.

To reconcile, note that the above discussion also applies to the usual e.d.f. $\hat{F}(x)$, and depending on the "promise" one makes, either x is fixed or drifts to boundaries, asymptotic claims change accordingly. Therefore the "discontinuity" of $\mathscr{V}_{p,0,x}$ in x is really the effect of a combination of (i) at boundaries c.d.f. estimators are $\sqrt{n}$-degenerate; and (ii) c.d.f. estimators target at different objectives in different asymptotic frameworks.

Such phenomenon does not occur for other components of $\hat{\boldsymbol{\beta}}_p(x)$, since they have nonparametric nature, and the evaluation point only affects the exact form of multiplicative constants, but not the rate of convergence.

Now we consider the problem of variance estimation. Given the formula in Theorem 1, it is possible to estimate the asymptotic variance by "plug-in" unknown quantities regarding the data generating process. For example consider $\mathscr{V}_{p,1,x}$ for the estimated density. Assume the researcher knows the location of the boundary $x_{\mathtt{L}}$ and $x_{\mathtt{U}}$, the matrices $\mathbf{S}_{p,x}$ and $\mathbf{\Gamma}_{p,x}$ can be constructed with numerical integration, since they are related to features of the kernel function, not the data generating process. The unknown function $H^{(1)}(x)$ can also be estimated, at least when $w_i \equiv 1$.[1]

Another approach is to utilize the decomposition of the estimator, in particular the $\hat{\mathbf{L}}$ term. To introduce our variance estimator, we make the following definitions.

$$\hat{\mathbf{S}}_{p,x} = \frac{1}{n}\mathbf{X}_h\mathbf{K}_h\mathbf{X}_n = \frac{1}{n}\sum_i \mathbf{r}_p\left(\frac{x_i - x}{h}\right)\mathbf{r}_p\left(\frac{x_i - x}{h}\right)' K_h(x_i - x)$$

$$\hat{\mathbf{\Gamma}}_{p,x} = \frac{1}{n^3}\sum_{i,j,k} \mathbf{r}_p\left(\frac{x_j - x}{h}\right)\mathbf{r}_p\left(\frac{x_k - x}{h}\right)' K_h(x_j - x)K_h(x_k - x)w_i^2\Big(\mathbb{1}[x_i \le x_j] - \hat{F}(x_j)\Big)\Big(\mathbb{1}[x_i \le x_k] - \hat{F}(x_k)\Big).$$

Following is the main result regarding variance estimation. It is automatic and fully-adaptive, in the sense that no knowledge about the location of boundaries is needed, neither does it require estimating nuisance parameters (such as H or its derivatives when the weights are not identically 1).

Theorem 2 (Variance Estimation).
Assume Assumptions 1–3 hold with $\alpha_x \ge p+1$, $\alpha_w \ge p$, $\alpha_{w,2} \ge 0$, and $\alpha_{w,4} \ge 0$ for some integer $p \ge 0$. Further $h \to 0$, $nh^2 \to \infty$ and $nh^{2p+1} = O(1)$. Then

$$\hat{\mathscr{V}}_{p,v,x} \equiv (v!)^2\mathbf{e}_v'\mathbf{N}_x\hat{\mathbf{S}}_{p,x}^{-1}\hat{\mathbf{\Gamma}}_{p,x}\hat{\mathbf{S}}_{p,x}^{-1}\mathbf{N}_x\mathbf{e}_v \ \to_{\mathbb{P}}\ \mathscr{V}_{p,v,x}.$$

Define the standard error as

$$\hat{\sigma}_{p,v,x} \equiv (v!)\sqrt{\frac{1}{nh^{2v}}\mathbf{e}_v'\hat{\mathbf{S}}_{p,x}^{-1}\hat{\mathbf{\Gamma}}_{p,x}\hat{\mathbf{S}}_{p,x}^{-1}\mathbf{e}_v},$$

then

$$\hat{\sigma}_{p,v,x}^{-1}\Big(\hat{F}_p^{(v)}(x) - F^{(v)}(x) - h^{p+1-v}\mathscr{B}_{p,v,x}\Big) \rightsquigarrow \mathcal{N}\Big(0,\ 1\Big).$$

Remark 6 ($\hat{\sigma}_{p,v,x}$ being automatic and fully-adaptive). Constructing $\hat{\mathscr{V}}_{p,v,x}$ requires the knowledge of the location of boundaries, since the scaling matrix $\mathbf{N}_x$ depends on whether x is interior or boundary. This is not surprising, since it is used in Theorem 1 to stabilize the estimator.

For statistical inference, it is not necessary to construct the scaling matrix $\mathbf{N}_x$, which is why the location of boundaries is irrelevant for constructing valid standard errors. Indeed, we do not use this information when defining $\hat{\sigma}_{p,v,x}$. Furthermore, despite that we have to split the definition of $\mathscr{V}_{p,v,x}$ according to v and x, $\hat{\sigma}_{p,v,x}$ automatically adapts to different scenarios, hence provides a unified approach for variance estimation.

3 Large Sample Properties with Estimated Weights

In this section we consider the case that the weights w_i are estimated in a previous step. Although intuitive, it is not easy to give general theories encompassing all estimated weights, since how the

[1] In this case $H^{(1)} = F^{(1)} = f = g$, which can be estimated by the consistent estimator $\hat{f}_p$.

weights are estimated may differ in applications, which in turn, is likely to have nontrivial impact on first order asymptotic results. For example in constructing the counterfactual density, the weights are ratios of frequencies of individuals with certain characteristics at two time points, while for Abadie (2003), the weights are constructed by employing binary treatment variable, instrument, and additional covariates. On one hand, we would like to present a framework that is general enough to encompass a wide range of applications, and on the other hand, it should also be tractable so that it is empirically relevant.

Assume the weights take the form $w_i = w(\mathbf{z}_i;\theta_0)$, where $\mathbf{z}_i$ are additional available information besides x_i (note that it is possible to make x_i part of $\mathbf{z}_i$), and θ_0 is some parameter to be estimated. Of course it is possible to let the parameter θ to be vector-valued. This will only make the notation more involved, which we will suppress. Let $\hat{\theta}$ be a consistent estimator of θ_0, the weights used in estimating the distributional properties are $\hat{w}_i = w(\mathbf{z}_i;\hat{\theta})$. To avoid introducing additional notation, let $\hat{F}$ be the e.d.f. except now it is constructed with estimated weights.

Consider the following expansion:

$$
\begin{aligned}
&\frac{1}{n}\sum_i \mathbf{r}_p\left(\frac{x_i-x}{h}\right)\Big(\hat{F}(x_i) - \mathbf{r}_p(x_i-x)'\boldsymbol{\beta}_p(x)\Big)K_h(x_i-x)\\
&\quad= \frac{1}{n}\sum_i \mathbf{r}_p\left(\frac{x_i-x}{h}\right)\Big(F(x_i) - \mathbf{r}_p(x_i-x)'\boldsymbol{\beta}_p(x)\Big)K_h(x_i-x) && (\hat{\mathbf{B}}_{\mathtt{S}})\\
&\quad+ \frac{1+o_{\mathbb{P}}(1)}{n}\sum_i \int_{\frac{x_{\mathtt{L}}-x}{h}}^{\frac{x_{\mathtt{U}}-x}{h}} \mathbf{r}_p(u)\hat{w}_i\Big(\mathbb{1}[x_i\le x+hu] - F(x+hu)\Big)K(u)g(x+hu)\mathrm{d}u && (\hat{\mathbf{L}})\\
&\quad+ \frac{1+o_{\mathbb{P}}(1)}{n^2}\sum_i \mathbf{r}_p\left(\frac{x_i-x}{h}\right)\hat{w}_i\Big(1-F(x_i)\Big)K_h(x_i-x) && (\hat{\mathbf{B}}_{\mathtt{LI}})\\
&\quad+ \frac{1+o_{\mathbb{P}}(1)}{n^2}\sum_{i,j;i\neq j} \hat{w}_j\Big\{\mathbf{r}_p\left(\frac{x_i-x}{h}\right)\Big(\mathbb{1}[x_j\le x_i] - F(x_i)\Big)K_h(x_i-x)\\
&\qquad - \int_{\frac{x_{\mathtt{L}}-x}{h}}^{\frac{x_{\mathtt{U}}-x}{h}} \mathbf{r}_p(u)\Big(\mathbb{1}[x_j\le x+hu] - F(x+hu)\Big)K(u)g(x+hu)\mathrm{d}u\Big\}, && (\hat{\mathbf{R}})
\end{aligned}
$$

provided that $n^{-1}\sum_i \hat{w}_i \to_{\mathbb{P}} 1$. The component representing smoothing bias, i.e. $\hat{\mathbf{B}}_{\mathtt{S}}$ remains the same as before, hence the first half of Lemma 2 remains to apply. For the other terms, we collect some preliminary lemmas in the following subsection.

Assumption 4 (Estimated weights).
(i) $\theta\mapsto w(\cdot;\theta)$ *is twice continuously differentiable, with derivatives denoted by* $\dot{w}$ *and* $\ddot{w}$.
(ii) For some $\delta>0$, $\mathbb{E}[\sup_{|\theta-\theta_0|\le\delta}|w(z_i;\theta)| + |\dot{w}(z_i;\theta)| + |\ddot{w}(z_i;\theta)|] < \infty$.
(iii) $\sqrt{n}(\hat{\theta}-\theta_0) = \sum_i \psi_i/\sqrt{n} + o_{\mathbb{P}}(1)$, *with* ψ_i *having zero mean and finite variance.*

3.1 Preliminary Lemmas

We first consider the leave-in bias.

Lemma 5. *Assume Assumptions 1–4 hold with* $\alpha_x \ge 1$. *Further* $h\to 0$ *and* $nh\to\infty$. *Then* $\hat{\mathbf{B}}_{\mathtt{LI}} = O_{\mathbb{P}}(1/n) = o_{\mathbb{P}}(\sqrt{h/n})$.

The next lemma handles the quadratic variance $\hat{\mathbf{R}}$.

Lemma 6. *Assume Assumptions 1–4 hold with $\alpha_x \geq 1$, $\alpha_w \geq 0$, and $\alpha_{w,2} \geq 0$. Further $h \to 0$ and $nh^2 \to \infty$. Then $\hat{\mathbf{R}} = O_{\mathbb{P}}(1/\sqrt{n^2h} + 1/n) = o_{\mathbb{P}}(\sqrt{h/n})$.*

Note that we emphasize the two terms, $\hat{\mathbf{B}}_{\mathtt{LI}}$ and $\hat{\mathbf{R}}$, having smaller order than $\sqrt{h/n}$, since the latter is the rate of the $\hat{\mathbf{L}}$ term.

Lemma 7. *Assume Assumptions 1–4 hold with $\alpha_x \geq 1$, $\alpha_w \geq 0$, and $\alpha_{w,2} \geq 0$. Further $h \to 0$ and $nh \to \infty$. Then for x in interior,*

$$\sqrt{n}\mathbf{N}_x\mathbf{S}_{p,x}^{-1}\frac{1+o_{\mathbb{P}}(1)}{n}\sum_i \int_{\frac{x_{\mathtt{L}}-x}{h}}^{\frac{x_{\mathtt{U}}-x}{h}} \mathbf{r}_p(u)\Big(\hat{w}_i - w_i\Big)\Big(\mathbb{1}[x_i \leq x+hu] - F(x+hu)\Big)K(u)g(x+hu)\mathrm{d}u$$
$$= g(x)\Big(I(x) - I(x_{\mathtt{U}})F(x)\Big)\mathbf{e}_0\left[\frac{1}{\sqrt{n}}\sum_i \psi_i\right] + o_{\mathbb{P}}(1).$$

And for x in the boundary,

$$\sqrt{n}\mathbf{N}_x\mathbf{S}_{p,x}^{-1}\frac{1+o_{\mathbb{P}}(1)}{n}\sum_i \int_{\frac{x_{\mathtt{L}}-x}{h}}^{\frac{x_{\mathtt{U}}-x}{h}} \mathbf{r}_p(u)\Big(\hat{w}_i - w_i\Big)\Big(\mathbb{1}[x_i \leq x+hu] - F(x+hu)\Big)K(u)g(x+hu)\mathrm{d}u \ = o_{\mathbb{P}}(1).$$

Here $I(x) = \mathbb{E}[\dot{w}_i\mathbb{1}[x_i \leq x]]$.

This lemma has an important implication: estimating the weights will have first order impact only on the smoothed c.d.f., $\mathbf{e}_0'\hat{\boldsymbol{\beta}}_p(x)$ when x is in interior. That is, it does not affect the estimated derivatives, since they are nonparametric objects, compared to which $\hat{\theta}$ has a much faster rate of convergence.

3.2 Main Results

We first give the theorem showing asymptotic normality of $\hat{\boldsymbol{\beta}}_p(x)$ with estimated weights.

Theorem 3 (Asymptotic normality with estimated weights: $\hat{\boldsymbol{\beta}}_p(x)$).
Assume Assumptions 1–4 hold with $\alpha_x \geq p+1$, $\alpha_w \geq p$, $\alpha_{w,2} \geq 0$ for some integer $p \geq 0$, and $\mathbb{E}[w_i^4] < \infty$. Further $h \to 0$, $nh^2 \to \infty$ and $nh^{2p+1} = O(1)$. Then

$$\sqrt{nh^{2v-1}}\Big(\hat{F}_p^{(v)}(x) - F^{(v)}(x) - h^{p+1-v}\mathscr{B}_{p,v,x}\Big) \rightsquigarrow \mathcal{N}\Big(0,\ \mathscr{V}_{p,v,x}\Big), \qquad 1 \leq v \leq p,$$
$$\sqrt{\frac{n}{\mathscr{V}_{p,0,x}}}\Big(\hat{F}_p(x) - F(x) - h^{p+1}\mathscr{B}_{p,0,x}\Big) \rightsquigarrow \mathcal{N}\Big(0,\ 1\Big).$$

The variance $\mathscr{V}_{p,v,x}$ is redefined as

$$\mathscr{V}_{p,v,x} = \begin{cases} (v!)^2H^{(1)}(x)\mathbf{e}_v'\mathbf{S}_{p,x}^{-1}\boldsymbol{\Gamma}_{p,x}\mathbf{S}_{p,x}^{-1}\mathbf{e}_v & 1 \leq v \leq 0 \\ H(x) - 2H(x)F(x) + H(x_{\mathtt{U}})F(x)^2 & \\ \quad +\Big(I(x) - I(x_{\mathtt{U}})F(x)\Big)^2\mathbb{V}[\psi_i] + 2\Big(I(x) - I(x_{\mathtt{U}})F(x)\Big)\mathbb{E}[\psi_i w_i \mathbb{1}[x_i \leq x]] & v = 0,\ x \text{ interior} \\ hH^{(1)}(x)\left(\mathbf{e}_0'\mathbf{S}_{p,x}^{-1}\boldsymbol{\Gamma}_{p,x}\mathbf{S}_{p,x}^{-1}\mathbf{e}_0 + c\right) & v = 0,\ x = x_{\mathtt{L}} + ch \text{ or } x_{\mathtt{U}} - ch. \end{cases}$$

Compared to Theorem 1, the only complication appears when $v = 0$ for interior evaluation point. The reason is simple, with weights estimated at $\sqrt{n}$-rate, the first step estimation will have nontrivial impact on the smoothed e.d.f., since the latter object is also estimated at $\sqrt{n}$-rate. The variance comes from essentially a two-step GMM problem.

The following is a companion result for constructing standard errors.

Theorem 4 (Variance Estimation).
Assume Assumptions 1–4 hold with $\alpha_x \geq p+1$, $\alpha_w \geq p$, $\alpha_{w,2} \geq 0$, and $\alpha_{w,4} \geq 0$ for some integer $p \geq 0$. Further $h \to 0$, $nh^2 \to \infty$ and $nh^{2p+1} = O(1)$. Assume either x is in the boundary regions or $v \geq 1$. Then

$$\hat{\mathscr{V}}_{p,v,x} \equiv (v!)^2 \mathbf{e}_v' \mathbf{N}_x \hat{\mathbf{S}}_{p,x}^{-1} \hat{\mathbf{\Gamma}}_{p,x} \hat{\mathbf{S}}_{p,x}^{-1} \mathbf{N}_x \mathbf{e}_v \;\; \to_{\mathbb{P}} \;\; \mathscr{V}_{p,v,x}.$$

Define the standard error as

$$\hat{\sigma}_{p,v,x} \equiv (v!) \sqrt{\frac{1}{nh^{2v}} \mathbf{e}_v' \hat{\mathbf{S}}_{p,x}^{-1} \hat{\mathbf{\Gamma}}_{p,x} \hat{\mathbf{S}}_{p,x}^{-1} \mathbf{e}_v},$$

then

$$\hat{\sigma}_{p,v,x}^{-1} \Big(\hat{F}_p^{(v)}(x) - F^{(v)}(x) - h^{p+1-v} \mathscr{B}_{p,v,x} \Big) \rightsquigarrow \mathcal{N}\Big(0,\ 1\Big).$$

We excluded the case for $v = 0$ and interior x, since constructing valid standard error requires knowledge of how the weights $\hat{w}_i$ are constructed, and is not captured by our variance estimator. If the object of interest is the smoothed e.d.f., we recommend to construct standard error using standard two-step GMM procedure, or using nonparametric bootstrap.

4 Additional Results

In this section we collect some results that are not essential to our main results, but otherwise will be useful in various applications. In the first part, we briefly illustrate how consistent MSE-optimal bandwidth can be constructed. Then we consider the problem of restricted estimation, when there is a natural way of data splitting. In a third subsection, we illustrate how valid standard errors can be constructed using a jackknife-based method.

4.1 Bandwidth Selection

In this subsection we consider the problem of constructing m.s.e.-optimal bandwidth for our local polynomial regression-based distribution estimators. We focus exclusively on the case $v \geq 1$, hence the object of interest is nonparametric in nature, and will be either the density function or derivatives thereof. Valid bandwidth choice for the distribution function $\hat{F}_p(x)$ is also an interesting topic, but difficulty arises since it is estimated with (at least) parametric rate. We will briefly mention m.s.e. expansion of the estimated c.d.f. at the end.

4.1.1 For Nonparametric Estimates ($v \geq 1$)

Consider some $1 \leq v \leq p$, the following lemma gives finer characterization of the bias.

Lemma 8. *Assume Assumptions 1–4 hold with $\alpha_x \geq p+2$, $\alpha_w \geq p+1$ and $\alpha_{w,2} \geq 0$. Further $h \to 0$ and $nh^3 \to \infty$. Then the leading bias of $\hat{F}_p^{(v)}(x)$ is characterized by*

$$h^{p+1-v}\mathscr{B}_{p,v,x} = h^{p+1-v}\left\{\frac{F^{(p+1)}(x)}{(p+1)!}v!\mathbf{e}_v'\mathbf{S}_{p,x}^{-1}\mathbf{c}_{p,x} + h\left(\frac{F^{(p+2)}(x)}{(p+2)!} + \frac{F^{(p+1)}(x)}{(p+1)!}\frac{G^{(2)}(x)}{G^{(1)}(x)}\right)v!\mathbf{e}_v'\mathbf{S}_{p,x}^{-1}\tilde{\mathbf{c}}_{p,x}\right\}.$$

The previous lemma is a refinement of Lemma 1 and 2, with both leading and higher-order bias explicitly characterized. To see its necessity, we note that when $p-v$ is even and x is in interior, the leading bias is zero, since $\mathbf{e}_v'\mathbf{S}_{p,x}^{-1}\mathbf{c}_{p,x}$ is zero. This is well-documented in the local polynomial regression literature. See Fan and Gijbels (1996) for a discussion. In general (that is, when rare cases such as $F^{(p+1)}(x) = 0$ or $F^{(p+2)}(x) = 0$ are excluded), we have the following:

Order of bias: $h^{p+1-v}\mathscr{B}_{p,v,x} \propto$

	$p-v$ odd	even
x interior	h^{p+1-v}	h^{p+2-v}
boundary	h^{p+1-v}	h^{p+1-v}

Note that for boundary evaluation points, the leading bias never vanishes.

The leading variance is also characterized by Theorem 1, and we reproduce it here:

$$\frac{1}{nh^{2v-1}}\mathscr{V}_{p,v,x} = \frac{1}{nh^{2v-1}}(v!)^2 H^{(1)}(x)\mathbf{e}_v'\mathbf{S}_{p,x}^{-1}\mathbf{\Gamma}_{p,x}\mathbf{S}_{p,x}^{-1}\mathbf{e}_v.$$

The m.s.e.-optimal bandwidth is defined as a minimizer of the following:

$$h_{\mathtt{MSE},p,v,x} = \arg\min_{h>0}\left[\frac{1}{nh^{2v-1}}\mathscr{V}_{p,v,x} + h^{2p+2-2v}\mathscr{B}_{p,v,x}^2\right].$$

Given the discussion we had earlier on the bias, it is easy to see that the MSE-optimal bandwidth has the following asymptotic order:

Order of m.s.e.-optimal bandwidth: $h_{\mathtt{MSE},p,v,x} \propto$

	$p-v$ odd	even
x interior	$n^{-\frac{1}{2p+1}}$	$n^{-\frac{1}{2p+3}}$
boundary	$n^{-\frac{1}{2p+1}}$	$n^{-\frac{1}{2p+1}}$

Again only the case where $p-v$ is even and x is interior needs special attention.

Next we consider the problem of bandwidth estimation/construction. There are two notions of consistency for estimated bandwidth. Let h be some nonstochastic bandwidth sequence, and $\hat{h}$ be the estimated bandwidth (sequence). Then $\hat{h}$ is consistent *in rate* if $\hat{h} \asymp h$ (in most cases it is even true that $\hat{h}/h \to_{\mathbb{P}} C \in (0,\infty)$). And $\hat{h}$ is consistent *in rate and constant* if $\hat{h}/h \to_{\mathbb{P}} 1$.

To construct consistent bandwidth, either rate consistent or consistent in both rate and constant, we need estimates for both the bias and variance. The variance part is easy, since it is demonstrated in Theorem 2 (or Theorem 4 for estimated weights) that the standard error, being completely automatic and adaptive, is consistent:

$$n\ell^{2v-1}\frac{\hat{\sigma}^2_{p,v,x}}{\mathscr{V}_{p,v,x}} \to_{\mathbb{P}} 1,$$

provided conditions specified in those theorems are satisfied. Here ℓ is some preliminary bandwidth used to construct $\hat{\sigma}_{p,v,x}$.

For the bias, there are two approaches. The first one is more common in the literature, where one distinguishes between the boundary and interior case, and provide consistent bias estimators separately. This method is appealing in the sense that the bandwidth constructed will be consistent both in rate and constant. The drawback, however, is that it requires precise knowledge about the location of x relative to the boundaries, which is not always obvious.

We will follow the second approach, where we replace the unknown bias by an estimate which is consistent in rate (but not necessarily in constant). More precisely, our bias estimator will be consistent in rate and constant if either x is boundary or $p-v$ is odd, and will be consistent in rate otherwise. This bias estimator has an appealing feature: it is purely data-driven and no precise knowledge about relative positioning of x to the boundaries is needed, with the price that it (and the bandwidth constructed thereof) is not consistent in constant when x is interior and $p-v$ is even.

To introduce this approach, first assume there are consistent estimators for $F^{(p+1)}(x)$ and $F^{(p+2)}(x)$, denoted by $\hat{F}^{(p+1)}(x)$ and $\hat{F}^{(p+2)}(x)$. We will not be too explicit about how those estimators are constructed. They can be obtained using our local polynomial regression-based approach, or can be constructed with some reference model (such as normal distribution). The critical step is to obtain consistent estimators of the matrices, which are given in the following lemma.

Lemma 9. *Assume Assumptions 1–3 hold with $\alpha_x \geq 1$. Further $\ell \to 0$ and $n\ell \to \infty$. Then*

$$\begin{aligned}\widehat{\mathbf{S}^{-1}_{p,x}\mathbf{c}_{p,x}} &= \left(\frac{1}{n}\sum_i \mathbf{r}_p\left(\frac{x_i-x}{\ell}\right)\mathbf{r}_p\left(\frac{x_i-x}{\ell}\right)' K_\ell(x_i-x)\right)^{-1}\left(\frac{1}{n}\sum_i\left(\frac{x_i-x}{\ell}\right)^{p+1}\mathbf{r}_p\left(\frac{x_i-x}{\ell}\right)K_\ell(x_i-x)\right)\\ &\to_{\mathbb{P}} \mathbf{S}^{-1}_{p,x}\mathbf{c}_{p,x},\end{aligned}$$

and

$$\begin{aligned}\widehat{\mathbf{S}^{-1}_{p,x}\tilde{\mathbf{c}}_{p,x}} &= \left(\frac{1}{n}\sum_i \mathbf{r}_p\left(\frac{x_i-x}{\ell}\right)\mathbf{r}_p\left(\frac{x_i-x}{\ell}\right)' K_\ell(x_i-x)\right)^{-1}\left(\frac{1}{n}\sum_i\left(\frac{x_i-x}{\ell}\right)^{p+2}\mathbf{r}_p\left(\frac{x_i-x}{\ell}\right)K_\ell(x_i-x)\right)\\ &\to_{\mathbb{P}} \mathbf{S}^{-1}_{p,x}\tilde{\mathbf{c}}_{p,x}.\end{aligned}$$

Note that we used different notation, ℓ, for bandwidth.

Now we have enough ingredients for bandwidth selection. Define:

$$h^{p+1-v}\hat{\mathscr{B}}_{p,v,x} = h^{p+1-v}\left\{\frac{\hat{F}^{(p+1)}(x)}{(p+1)!}v!\mathbf{e}_v'\widehat{\mathbf{S}^{-1}_{p,x}\mathbf{c}_{p,x}} + h\frac{\hat{F}^{(p+2)}(x)}{(p+2)!}v!\mathbf{e}_v'\widehat{\mathbf{S}^{-1}_{p,x}\tilde{\mathbf{c}}_{p,x}}\right\},$$

and assume that $\hat{\sigma}_{p,v,x}$ is constructed using the preliminary bandwidth ℓ. Then

$$\hat{h}_{\texttt{MSE},p,v,x} = \arg\min_{h>0} \left[\frac{\ell^{2v-1}}{h^{2v-1}} \hat{\sigma}^2_{p,v,x} + h^{2p+2-2v} \hat{\mathscr{B}}^2_{p,v,x} \right].$$

We make three remarks here.

Remark 7 (Optimization argument h and preliminary bandwidth ℓ). The optimization argument h enters the RHS of the previous display in three places. First it is part of the variance component, by $1/h^{2v-1}$. Second it shows as a multiplicative factor of the bias component, $h^{2p-2v+2}$. Finally within the definition of $\hat{\mathscr{B}}_{p,v,x}$, there is another multiplicative h, in front of the higher order bias.

The preliminary bandwidth ℓ, serves a different role. It is used to estimate the variance and bias components. Of course one can use different preliminary bandwidths for $\hat{\sigma}_{p,v,x}$, $\widehat{\mathbf{S}^{-1}_{p,x}\mathbf{c}_{p,x}}$ and $\widehat{\mathbf{S}^{-1}_{p,x}\tilde{\mathbf{c}}_{p,x}}$, provided they satisfy corresponding regularity conditions.

Remark 8 (Known boundaries). If boundary locations are known, either from *a priori* knowledge or suggested by the data, then it is possible to simplify the problem, and closed-form solution for $\hat{h}_{\texttt{MSE},p,v,x}$ is feasible. To be precise, if it is known that x is a boundary point *or* $p-v$ is odd, one can simply ignore the second component in $\hat{\mathscr{B}}_{p,v,x}$. Similarly, if it is the case that x is interior and $p-v$ is even, then the first component in $\hat{\mathscr{B}}_{p,v,x}$ can be skipped.

The option we opt-for is more flexible in the sense that it adapts to any $p-v$ (odd or even) and any x (interior or boundary).

Remark 9 (Consistent bias estimator). The bias estimator we proposed, $h^{p-v+1}\hat{\mathscr{B}}_{p,v,x}$, is consistent in rate for the true leading bias, but not necessarily in constant. Compare $\hat{\mathscr{B}}_{p,v,x}$ and $\mathscr{B}_{p,v,x}$, it is easily seen that the term involving $F^{(p+1)}(x)G^{(2)}(x)/G^{(1)}(x)$ is not captured. To capture this term, we need two additional nonparametric estimators, one for $G^{(2)}(x)$ and the other for $G^{(1)}(x)$. This is indeed feasible, as one can employ our local polynomial regression-based estimator for this purpose. The complication, however, is that G is a different distribution, hence one needs to construct the estimator from scratch. This leads to additional computational burden which may not be attractive in practice.

There is one case, however, where estimating $G^{(1)}(x)$ and $G^{(2)}(x)$ is almost free – when the weighting satisfies $w_i \equiv 1$. Then $F = G$, and with $p \geq 2$, both are automatically produced hence requires no additional estimation effort.

Theorem 5 (Consistent bandwidth). *Let $1 \leq v \leq p$. Assume the preliminary bandwidth ℓ is chosen such that $nh^{2v-1}\hat{\sigma}^2_{p,v,x}/\mathscr{V}_{p,v,x} \to_{\mathbb{P}} 1$, $\widehat{\mathbf{S}^{-1}_{p,x}\mathbf{c}_{p,x}} \to_{\mathbb{P}} \mathbf{S}^{-1}_{p,x}\mathbf{c}_{p,x}$, and $\widehat{\mathbf{S}^{-1}_{p,x}\tilde{\mathbf{c}}_{p,x}} \to_{\mathbb{P}} \mathbf{S}^{-1}_{p,x}\tilde{\mathbf{c}}_{p,x}$, with other regularity conditions given in Lemma 1 and Theorem 2/4.*

- *If either x is in boundary regions or $p-v$ is odd, let $\hat{F}^{(p+1)}(x)$ be consistent for $F^{(p+1)} \neq 0$. Then*

$$\frac{\hat{h}_{\texttt{MSE},p,v,x}}{h_{\texttt{MSE},p,v,x}} \to_{\mathbb{P}} 1.$$

- *If x is in interior and $p-v$ is even, let $\hat{F}^{(p+2)}(x)$ be consistent for $F^{(p+2)} \neq 0$. Further assume $nh^3 \to 0$ and $h_{\mathtt{MSE},p,v,x}$ is well-defined. Then*

$$\frac{\hat{h}_{\mathtt{MSE},p,v,x}}{h_{\mathtt{MSE},p,v,x}} \to_{\mathbb{P}} C \in (0,\infty).$$

4.1.2 For C.D.F. Estimate ($v = 0$)

In this subsection we mention briefly how to choose bandwidth for the c.d.f. estimate, $\hat{F}_p^{(0)}(x) \equiv \hat{F}_p(x)$. We assume x is in interior. Previous discussions on bias also applies to $\hat{F}_p(x)$:

$$h^{p+1}\mathscr{B}_{p,0,x} = h^{p+1}\left\{\frac{F^{(p+1)}(x)}{(p+1)!}\mathbf{e}_0'\mathbf{S}_{p,x}^{-1}\mathbf{c}_{p,x} + h\left(\frac{F^{(p+2)}(x)}{(p+2)!} + \frac{F^{(p+1)}(x)}{(p+1)!}\frac{G^{(2)}(x)}{G^{(1)}(x)}\right)\mathbf{e}_0'\mathbf{S}_{p,x}^{-1}\tilde{\mathbf{c}}_{p,x}\right\},$$

so that the bias for $\hat{F}_p(x)$ has order h^{p+1} if either x is boundary or p is odd, and h^{p+2} otherwise. Difficulty does arise since the c.d.f. estimator has leading variance of order[2]

$$\mathscr{V}_{p,0,x} \propto \frac{\mathbb{1}[x \text{ interior}] + h}{n},$$

which cannot be used for bandwidth selection, since the leading variance is proportional to the bandwidth, which means there is no bias-variance tradeoff.

The trick is to use a higher order variance term. Recall that the local polynomial regression-based estimator is essentially a second order U-statistic, which is then decomposed into two terms, the linear term $\hat{\mathbf{L}}$, and a quadratic term $\hat{\mathbf{R}}$ which is a degenerate second order U-statistic. The variance of the quadratic term $\hat{\mathbf{R}}$ has been ignored so far, since it is negligible compared to the variance of the linear term. For the c.d.f. estimator, however, it is the variance of this quadratic term that leads to bias-variance trade-off, hence should be used to define m.s.e.-optimal bandwidth. The exact form of this variance is given in Lemma 4/6, and we will not repeat here. With this additional variance term included, we have (with some abuse of notation)

$$\mathscr{V}_{p,0,x} \propto \frac{\mathbb{1}[x \text{ interior}] + h}{n} + \frac{\mathbb{1}[x \text{ interior}] + h}{n^2 h},$$

so that provided x is an interior point, the additional variance term increases as the bandwidth shrinks. Therefore the m.s.e.-optimal bandwidth for $\hat{F}_p(x)$ is well-defined. And estimating this bandwidth is also straightforward, simply by replacing unknown quantities with their estimates. The following table summarizes the order of the m.s.e.-optimal bandwidth for the estimated c.d.f.

Order of m.s.e.-optimal bandwidth: $h_{\mathtt{MSE},p,0,x} \propto$

	$p-v$ odd	even
x interior	$n^{-\frac{2}{2p+3}}$	$n^{-\frac{2}{2p+5}}$
boundary	undefined	undefined

[2]More precisely, the leading variance depends on the asymptotic framework used – whether x is regarded as a fixed point in the interior, or it is a drifting sequence to boundaries.

What if x is in boundary region? Then the m.s.e.-optimal bandwidth for $\hat{F}_p(x)$ is not well defined. The leading variance now takes the form $h/n+1/n^2$, which is proportional to the bandwidth – this is not surprising, for boundary x, the c.d.f. is known, hence a very small bandwidth (as long as one still has enough observations to construct the estimator numerically) gives a super-consistent estimator, although not an interesting one, as it estimates either 0 or 1. However, we would like to mention that, although m.s.e.-optimal bandwidth for $\hat{F}_p(x)$ is not well-defined for boundary x, it is still feasible to minimize the empirical MSE. To see how this works, one first estimate the bias term and variance term with some preliminary bandwidth ℓ, leading to $\hat{\mathscr{B}}_{p,0,x}$ and $\hat{\mathscr{V}}_{p,0,x}$. Then the m.s.e.-optimal bandwidth can be constructed by minimizing the empirical m.s.e.. Under regularity conditions, $\hat{\mathscr{B}}_{p,0,x}$ will converge to some nonzero constant, while, if x is boundary, $\hat{\mathscr{V}}_{p,0,x}$ has order ℓ, the same as the preliminary bandwidth. Then the MSE-optimal bandwidth constructed in this way will have the following order:

Order of estimated m.s.e.-optimal bandwidth: $\hat{h}_{\mathtt{MSE},p,0,x} \propto$

	$p-v$ odd	even
x interior	$n^{-\frac{2}{2p+3}}$	$n^{-\frac{2}{2p+5}}$
boundary	$(n^2/\ell)^{-\frac{1}{2p+3}}$	$(n^2/\ell)^{-\frac{1}{2p+5}}$

Note that the preliminary bandwidth enters the rate of $\hat{h}_{\mathtt{MSE},p,0,x}$ for boundary x, since it determines the rate at which the variance estimator $\hat{\mathscr{V}}_{p,0,x}$ vanishes. Although this estimated bandwidth is not consistent for any well-defined object, it can be useful in practice, and it does reflect the fact that for boundary x it is appropriate to use bandwidth shrinks fast when the object of interest is the c.d.f.

4.2 Imposing Restrictions with Joint Estimation

We devote this subsection to estimation problems where it can be desirable to have joint estimation and/or impose restrictions. To illustrate the idea, we will discuss in the context of density discontinuity (manipulation) tests in regression discontinuity designs.

Assume there is a natural (and known) partition of the support $\mathcal{X} = [x_{\mathtt{L}}, x_{\mathtt{U}}] = [x_{\mathtt{L}}, \bar{x}) \cup [\bar{x}, x_{\mathtt{U}}] = \mathcal{X}_- \cup \mathcal{X}_+$, and the regularity conditions we imposed so far are satisfied on each of the partitions, $\mathcal{X}_-$ and $\mathcal{X}_+$, but not necessarily the union. To be more precise, assume the distribution F is continuously differentiable to a certain order on each of the partitions, but the derivatives are not necessarily continuous across the cutoff $\bar{x}$. In this case consistent estimates of the densities and derivatives thereof require fitting local polynomials separately on each sides of $\bar{x}$, with corresponding subsamples. Alternatively, one can use the joint estimation framework introduced below.

In problems with joint estimation and/or restrictions, notation tends to be cumbersome. For the ease of exposition, we will assume $w_i \equiv 1$ throughout this subsection. Corresponding results with nontrivial weighting scheme or even estimated weights can be obtained with some additional efforts. Also we fix the evaluation point $\bar{x}$, and drop the corresponding subscript whenever possible.

4.2.1 Unrestricted Model

By an unrestricted model with cutoff $\bar{x}$, we consider the following polynomial basis $\mathbf{r}_p$

$$\mathbf{r}_p(u) = \Big[\mathbf{1}_{\{u<0\}} \quad u\mathbf{1}_{\{u<0\}} \quad \cdots \quad u^p\mathbf{1}_{\{u<0\}} \quad \Big| \quad \mathbf{1}_{\{u\geq 0\}} \quad u\mathbf{1}_{\{u\geq 0\}} \quad \cdots \quad u^p\mathbf{1}_{\{u\geq 0\}}\Big]' \in \mathbb{R}^{2p+2}.$$

The following two vectors will arise later, which we give the definition here:

$$\begin{aligned}
\mathbf{r}_{-,p}(u) &= \Big[1 \quad u \quad \cdots \quad u^p \quad 0 \quad \cdots \quad 0\Big]' \\
\mathbf{r}_{+,p}(u) &= \Big[0 \quad 0 \quad \cdots \quad 0 \quad 1 \quad \cdots \quad u^p\Big]'.
\end{aligned}$$

Also we define the vectors to extract the corresponding derivatives

$$\mathbf{I}_{2p+2} = \Big[\mathbf{e}_{0,-} \quad \mathbf{e}_{1,-} \quad \cdots \quad \mathbf{e}_{p,-} \quad \mathbf{e}_{0,+} \quad \mathbf{e}_{1,+} \quad \cdots \quad \mathbf{e}_{p,+}\Big].$$

With the above definition, the estimator at the cutoff is[3]

$$\hat{\boldsymbol{\beta}}_p(\bar{x}) = \arg\min_{\mathbf{b}\in\mathbb{R}^{2p+2}} \sum_i \Big(\hat{F}(x_i) - \mathbf{r}_p(x_i-\bar{x})'\mathbf{b}\Big)^2 K_h(x_i-\bar{x}).$$

Other notations (for example $\mathbf{X}$ and $\mathbf{X}_h$) are redefined similarly, with the scaling matrix $\mathbf{H}$ adjusted so that $\mathbf{H}^{-1}\mathbf{r}_p(u) = \mathbf{r}_p(h^{-1}u)$ is always true.

Note that the above is equivalent to fitting local polynomials separately on each side, while the joint estimation framework is more systematic which we will keep using. To see the connection between the joint estimation and estimating separately on each side, we observe the following result, which can be easily seen using least squares algebra:

Relation between joint and separate estimations.

	Joint estimation	Separate estimation
$\hat{F}_p(\bar{x}-)$	$\mathbf{e}_{0,-}'\hat{\boldsymbol{\beta}}_p(\bar{x}) = \mathbf{e}_{0,-}'\hat{\boldsymbol{\beta}}_{p,-}(\bar{x})$	"joint" $\times \frac{n}{n_-}$
$\hat{F}_p(\bar{x}+)$	$\mathbf{e}_{0,+}'\hat{\boldsymbol{\beta}}_p(\bar{x}) = \mathbf{e}_{0,+}'\hat{\boldsymbol{\beta}}_{p,+}(\bar{x})$	"joint" $\times \frac{n}{n_+} - \frac{n_-}{n_+}$
$\hat{F}_p^{(v)}(\bar{x}-)$	$\mathbf{e}_{v,-}'\hat{\boldsymbol{\beta}}_p(\bar{x}) = \mathbf{e}_{v,-}'\hat{\boldsymbol{\beta}}_{p,-}(\bar{x})$	"joint" $\times \frac{n}{n_-}$
$\hat{F}_p^{(v)}(\bar{x}+)$	$\mathbf{e}_{v,+}'\hat{\boldsymbol{\beta}}_p(\bar{x}) = \mathbf{e}_{v,+}'\hat{\boldsymbol{\beta}}_{p,+}(\bar{x})$	"joint" $\times \frac{n}{n_+}$

and the difference comes from the fact that by separate estimation, one obtains estimates of the conditional c.d.f. and the derivatives. Here n_- and n_+ are the sample sizes in the two regions, $\mathcal{X}_-$ and $\mathcal{X}_+$, respectively.

In the following lemmas, we will give asymptotic results of the joint estimation problem. Proofs are omitted.

Lemma 10. *Let Assumptions of Lemma 1 hold separately on $\mathcal{X}_-$ and $\mathcal{X}_+$, then*

$$\begin{aligned}
\frac{1}{n}\mathbf{X}_h'\mathbf{K}_h\mathbf{X}_h &= f(\bar{x}-)\mathbf{S}_{-,p} + f(\bar{x}+)\mathbf{S}_{+,p} + O\Big(h\Big) + O_{\mathbb{P}}\Big(1/\sqrt{nh}\Big) \\
&= \mathbf{S}_{f,p} + O(h) + O_{\mathbb{P}}(1/\sqrt{nh}),
\end{aligned}$$

[3] The e.d.f. is defined with the whole sample as before: $\hat{F}(u) = n^{-1}\sum_i \mathbb{1}[x_i \leq u]$.

where

$$\mathbf{S}_{-,p} = \int_{-1}^{0} \mathbf{r}_{-,p}(u)\mathbf{r}_{-,p}(u)'K(u)\mathrm{d}u, \qquad \mathbf{S}_{+,p} = \int_{0}^{1} \mathbf{r}_{+,p}(u)\mathbf{r}_{+,p}(u)'K(u)\mathrm{d}u.$$

Again we decompose the estimator into four terms, namely $\hat{\mathbf{B}}_{\mathtt{LI}}$, $\hat{\mathbf{B}}_{\mathtt{S}}$, $\hat{\mathbf{L}}$ and $\hat{\mathbf{R}}$.

Lemma 11. *Let Assumptions of Lemma 2 hold separately on $\mathcal{X}_-$ and $\mathcal{X}_+$, then*

$$\hat{\mathbf{B}}_{\mathtt{S}} = h^{p+1}\left\{\frac{F^{(p+1)}(\bar{x}-)f(\bar{x}-)}{(p+1)!}\mathbf{c}_{-,p} + \frac{F^{(p+1)}(\bar{x}+)f(x+)}{(p+1)!}\mathbf{c}_{+,p}\right\} + o_{\mathbb{P}}(h^{p+1}), \qquad \hat{\mathbf{B}}_{\mathtt{LI}} = O_{\mathbb{P}}\left(\frac{1}{n}\right),$$

where

$$\mathbf{c}_{-,p} = \int_{-1}^{0} u^{p+1}\mathbf{r}_{-,p}(u)K(u)\mathrm{d}u, \qquad \mathbf{c}_{+,p} = \int_{0}^{1} u^{p+1}\mathbf{r}_{+,p}(u)K(u)\mathrm{d}u.$$

Lemma 12. *Let Assumptions of Lemma 3 hold separately on $\mathcal{X}_-$ and $\mathcal{X}_+$, then*

$$\begin{aligned}
&\mathbb{V}\left[\hat{\mathbf{L}}\right] \\
&= F(\bar{x})\Big(1-F(\bar{x})\Big)\mathbf{S}_{f,p}\Big(\mathbf{e}_{0,-}+\mathbf{e}_{0,+}\Big)\Big(\mathbf{e}_{0,-}+\mathbf{e}_{0,+}\Big)'\mathbf{S}_{f,p} \\
&+ h\Big\{F(\bar{x})\Big(1-F(\bar{x})\Big)\frac{f'(\bar{x}-)}{f(\bar{x}-)} - F(\bar{x})f(\bar{x}-) + f(\bar{x}-)\Big\}\mathbf{S}_{f,p}\Big(\mathbf{e}_{0,-}\mathbf{e}_{1,-}' + \mathbf{e}_{1,-}\mathbf{e}_{0,-}'\Big)\mathbf{S}_{f,p} \\
&+ h\Big\{F(\bar{x})\Big(1-F(\bar{x})\Big)\frac{f'(\bar{x}+)}{f(\bar{x}+)} - F(\bar{x})f(\bar{x}+)\Big\}\mathbf{S}_{f,p}\Big(\mathbf{e}_{0,+}\mathbf{e}_{1,+}' + \mathbf{e}_{1,+}\mathbf{e}_{0,+}'\Big)\mathbf{S}_{f,p} \\
&+ h\Big\{F(\bar{x})\Big(1-F(\bar{x})\Big)\frac{f'(\bar{x}-)}{f(\bar{x}-)} - F(\bar{x})f(\bar{x}-) + f(\bar{x}-)\Big\}\mathbf{S}_{f,p}\Big(\mathbf{e}_{1,-}\mathbf{e}_{0,+}' + \mathbf{e}_{0,+}\mathbf{e}_{1,-}'\Big)\mathbf{S}_{f,p} \\
&+ h\Big\{F(\bar{x})\Big(1-F(\bar{x})\Big)\frac{f'(\bar{x}+)}{f(\bar{x}+)} - F(\bar{x})f(\bar{x}+)\Big\}\mathbf{S}_{f,p}\Big(\mathbf{e}_{0,-}\mathbf{e}_{1,+}' + \mathbf{e}_{1,+}\mathbf{e}_{0,-}'\Big)\mathbf{S}_{f,p} \\
&+ h\Big\{f(\bar{x}-)^3\mathbf{\Psi}\mathbf{\Gamma}_{+,p}\mathbf{\Psi} + f(\bar{x}+)^3\mathbf{\Gamma}_{+,p}\Big\} + O(h^2),
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{\Gamma}_{-,p} &= \iint_{[-1,0]^2}(u\wedge v)\mathbf{r}_{-,p}(u)\mathbf{r}_{-,p}(v)'K(u)K(v)\ \mathrm{d}u\mathrm{d}v, \\
\mathbf{\Gamma}_{+,p} &= \iint_{[0,1]^2}(u\wedge v)\mathbf{r}_{+,p}(u)\mathbf{r}_{+,p}(v)'K(u)K(v)\ \mathrm{d}u\mathrm{d}v,
\end{aligned}$$

and

$$\mathbf{\Psi} = \begin{bmatrix} & & & & (-1)^0 & & & \\ & & & & & (-1)^1 & & \\ & & & & & & \ddots & \\ & & & & & & & (-1)^p \\ (-1)^0 & & & & & & & \\ & (-1)^1 & & & & & & \\ & & \ddots & & & & & \\ & & & (-1)^p & & & & \end{bmatrix}_{(2p+2)\times(2p+2)}.$$

We would like to consider the asymptotic variance (hence asymptotic distribution) after proper scaling. Here we define the scaling matrix by

$$\mathbf{N} = \text{diag}\left\{\sqrt{n}, \sqrt{\frac{n}{h}}, \cdots, \sqrt{\frac{n}{h}}, \sqrt{n}, \sqrt{\frac{n}{h}}, \cdots, \sqrt{\frac{n}{h}}\right\}_{(2p+2)\times(2p+2)}$$
$$= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}_{2\times 2} \otimes \ \text{diag}\left\{\sqrt{n}, \sqrt{\frac{n}{h}}, \cdots, \sqrt{\frac{n}{h}}\right\}_{(p+1)\times(p+1)},$$

then

$$\begin{aligned}\mathbb{V}\left[\mathbf{N}\mathbf{S}_{f,p}^{-1}\hat{\mathbf{L}}\right] &= F(\bar{x})\Big(1-F(\bar{x})\Big)\Big(\mathbf{e}_{0,-}+\mathbf{e}_{0,+}\Big)\Big(\mathbf{e}_{0,-}+\mathbf{e}_{0,+}\Big)' \\ &+ \Big(\mathbf{I}-\Big(\mathbf{e}_{0,-}+\mathbf{e}_{0,+}\Big)\Big(\mathbf{e}_{0,-}+\mathbf{e}_{0,+}\Big)'\Big)\mathbf{S}_{f,p}^{-1}\Big(f(\bar{x}-)^3\mathbf{\Psi}\mathbf{\Gamma}_{+,p}\mathbf{\Psi}+f(\bar{x}+)^3\mathbf{\Gamma}_{+,p}\Big) \\ &\qquad \mathbf{S}_{f,p}^{-1}\Big(\mathbf{I}-\Big(\mathbf{e}_{0,-}+\mathbf{e}_{0,+}\Big)\Big(\mathbf{e}_{0,-}+\mathbf{e}_{0,+}\Big)'\Big) \\ &+ O(\sqrt{h}),\end{aligned}$$

where $O(\sqrt{h})$ represent the order of the covariances between the c.d.f. (the parametric part) and the derivatives (nonparametric part). By using the notation $\mathbf{S}_{f,p}^{-1} = \frac{1}{f(\bar{x}-)}\mathbf{S}_{-,p}^{-1} + \frac{1}{f(\bar{x}+)}\mathbf{S}_{+,p}^{-1}$, we have[4]

$$\begin{aligned}\mathbb{V}\left[\mathbf{N}\mathbf{S}_{f,p}^{-1}\hat{\mathbf{L}}\right] &= F(\bar{x})\Big(1-F(\bar{x})\Big)\Big(\mathbf{e}_{0,-}+\mathbf{e}_{0,+}\Big)\Big(\mathbf{e}_{0,-}+\mathbf{e}_{0,+}\Big)' \\ &+ f(\bar{x}-)\Big(\mathbf{I}-\mathbf{e}_{0,-}\mathbf{e}_{0,-}'\Big)\mathbf{\Psi}\mathbf{S}_{+,p}^{-1}\mathbf{\Gamma}_{+,p}\mathbf{S}_{+,p}^{-1}\mathbf{\Psi}\Big(\mathbf{I}-\mathbf{e}_{0,-}\mathbf{e}_{0,-}'\Big) \\ &+ f(\bar{x}+)\Big(\mathbf{I}-\mathbf{e}_{0,+}\mathbf{e}_{0,+}'\Big)\mathbf{S}_{+,p}^{-1}\mathbf{\Gamma}_{+,p}\mathbf{S}_{+,p}^{-1}\Big(\mathbf{I}-\mathbf{e}_{0,+}\mathbf{e}_{0,+}'\Big) \\ &+ o(1)\end{aligned}$$

$$= \left[\begin{array}{cc|cc} F(\bar{x})\Big(1-F(\bar{x})\Big) & \mathbf{0} & F(\bar{x})\Big(1-F(\bar{x})\Big) & \mathbf{0} \\ \mathbf{0} & \Big\{f(\bar{x}-)\mathbf{\Psi}\mathbf{S}_{+,p}^{-1}\mathbf{\Gamma}_{+,p}\mathbf{S}_{+,p}^{-1}\mathbf{\Psi}\Big\}_{(2:p+1)} & \mathbf{0} & \mathbf{0} \\ \hline F(\bar{x})\Big(1-F(\bar{x})\Big) & \mathbf{0} & F(\bar{x})\Big(1-F(\bar{x})\Big) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \Big\{f(\bar{x}+)\mathbf{S}_{+,p}^{-1}\mathbf{\Gamma}_{+,p}\mathbf{S}_{+,p}^{-1}\Big\}_{(p+3:2p+2)} \end{array}\right],$$

where the operator $\{\cdot\}_{(2:p+1)}$ indicates keeping only the second to $(p+1)$-th rows and columns. Therefore asymptotically

1. the c.d.f. (parametric part) and the derivatives (nonparametric part) are independent;

2. the two c.d.f. estimators (on each sides) have correlation 1 (not surprising, since we assume the DGP does not have point mass);

3. the derivatives (nonparametric part) on the two sides are independent.

Finally the order of $\hat{\mathbf{R}}$ can also be established.

Lemma 13. *Let Assumptions of Lemma 4 hold separately on $\mathcal{X}_-$ and $\mathcal{X}_+$, then*

$$\hat{\mathbf{R}} = O_{\mathbb{P}}\left(\sqrt{\frac{1}{n^2 h}}\right).$$

[4] $\mathbf{S}_{-,p}$ and $\mathbf{S}_{+,p}$ are not invertible. Here $\mathbf{S}_{-,p}^{-1}$ and $\mathbf{S}_{+,p}^{-1}$ are obtained by inverting the corresponding nonzero blocks. More precisely, they are Moore-Penrose inverse.

We note that it is also possible to give exact form of the variance of $\hat{\mathbf{R}}$.

In what follows we will consider the bias and the asymptotic distribution of $\hat{f}_p(\bar{x}+) - \hat{f}_p(\bar{x}-)$, which is the object of interest for density discontinuity tests.

Theorem 6. *Let Assumptions of Theorem 1 hold separately on $\mathcal{X}_-$ and $\mathcal{X}_+$, then*

$$\sqrt{nh_n}\,\frac{\hat{f}_p(\bar{x}+) - \hat{f}_p(\bar{x}-) - \Big(f(\bar{x}+) - f(\bar{x}-)\Big) - h^p\mathscr{B}_{p,1}}{\sqrt{\mathscr{V}_{p,1}}} \rightsquigarrow \mathcal{N}(0,1),$$

where

$$\mathscr{B}_{p,1} = \left\{\frac{F^{(p+1)}(\bar{x}+)}{(p+1)!}\mathbf{e}_{1,+}'\mathbf{S}_{+,p}^{-1}\mathbf{c}_{+,p} - \frac{F^{(p+1)}(\bar{x}-)}{(p+1)!}\mathbf{e}_{1,-}'\mathbf{S}_{-,p}^{-1}\mathbf{c}_{-,p}\right\}$$
$$\mathscr{V}_{p,1} = \Big(f(\bar{x}+) + f(-)\Big)\mathbf{e}_{1,+}'\mathbf{S}_{+,p}^{-1}\mathbf{\Gamma}_{+,p}\mathbf{S}_{+,p}^{-1}\mathbf{e}_{1,+}.$$

Remark 10. We make two remarks here.
(1) The asymptotic variance takes additive form.
(2) The standard error proposed earlier remains valid. Note that by the specific structure of $\mathbf{r}_{-,p}$ and $\mathbf{r}_{+,p}$, it is equivalent to apply the method on each side with the corresponding subsample.

4.2.2 Restricted Model

In previous discussion, we gave a test procedure on the discontinuity of the density by estimating on the two sides of the cutoff separately. This procedure is flexible and requires minimum assumptions. There are ways, however, to improve the power of the test when the densities are estimated with additional assumptions on the smoothness of the c.d.f.

In a restricted model, the polynomial basis is re-defined as

$$\mathbf{r}_p(u) = \begin{bmatrix}1 & u\mathbf{1}(u<0) & u\mathbf{1}(u\geq 0) & u^2 & u^3 & \cdots & u^p\end{bmatrix}' \in \mathbb{R}^{p+2},$$

and the estimator in the fully restricted model is

$$\hat{\boldsymbol{\beta}}_p(\bar{x}) = \begin{bmatrix}\hat{F}_p(\bar{x}) & \hat{f}_p(\bar{x}-) & \hat{f}_p(\bar{x}+) & \frac{1}{2}\hat{F}_p^{(2)}(\bar{x}) & \cdots & \frac{1}{p!}\hat{F}_p^{(p)}(\bar{x})\end{bmatrix}'$$
$$= \arg\max_{\mathbf{b}\in\mathbb{R}^{p+2}} \sum_i \Big(\hat{F}(x_i) - \mathbf{r}_p(x_i - \bar{x})'\mathbf{b}\Big)^2 K_h(x_i - \bar{x}).$$

Again the notations (for example $\mathbf{X}$ and $\mathbf{X}_h$) are redefined similarly, with the scaling matrix $\mathbf{H}$ adjusted to make sure $\mathbf{H}^{-1}\mathbf{r}_p(u) = \mathbf{r}_p(h^{-1}u)$. Here $\hat{F}_p(\bar{x})$ is the estimated c.d.f. and $\frac{1}{2}\hat{F}_p^{(2)}(\bar{x})$, $\cdots$, $\frac{1}{p!}\hat{F}_p^{(p)}(\bar{x})$ are the estimated higher order derivatives, which we assume are all continuous at $\bar{x}$, while $\hat{f}_p(\bar{x}-)$ and $\hat{f}_p(\bar{x}+)$ are the estimated densities on the two sides of $\bar{x}$. Therefore we call the above model restricted, since it only allows discontinuity of the first derivative of F (i.e. the density) but not the other derivatives.

With the modification of the polynomial basis, all other matrices in the previous subsection are redefined similarly, and

$$\mathbf{I}_{p+2} = \begin{bmatrix}\mathbf{e}_0 & \mathbf{e}_{1,-} & \mathbf{e}_{1,+} & \mathbf{e}_2 & \cdots & \mathbf{e}_p\end{bmatrix}_{(p+2)\times(p+2)}.$$

where the subscripts indicate the corresponding derivatives to extract. Moreover

$$\mathbf{r}_{-,p}(u) = \begin{bmatrix}1 & u & 0 & u^2 & \cdots & u^p\end{bmatrix}, \qquad \mathbf{r}_{+,p}(u) = \begin{bmatrix}1 & 0 & u & u^2 & \cdots & u^p\end{bmatrix}.$$

Lemma 14. *Let Assumptions of Lemma 1 hold with the exception that f may be discontinuous across $\bar{x}$, then*

$$\begin{aligned}\frac{1}{n}\mathbf{X}_h'\mathbf{K}_h\mathbf{X}_h &= \{f(\bar{x}-)\mathbf{S}_{-,p} + f(\bar{x}+)\mathbf{S}_{+,p}\} + O(h) + O_{\mathbb{P}}(1/\sqrt{nh}) \\ &= \mathbf{S}_{f,p} + O(h) + O_{\mathbb{P}}(1/\sqrt{nh}),\end{aligned}$$

where

$$\mathbf{S}_{-,p} = \int_{-1}^{0} \mathbf{r}_{-,p}(u)\mathbf{r}_{-,p}(u)'K(u)\mathrm{d}u, \qquad \mathbf{S}_{+,p} = \int_{0}^{1} \mathbf{r}_{+,p}(u)\mathbf{r}_{+,p}(u)'K(u)\mathrm{d}u.$$

Again we decompose the estimator into four terms, $\hat{\mathbf{B}}_{\mathtt{LI}}$, $\hat{\mathbf{B}}_{\mathtt{S}}$, $\hat{\mathbf{L}}$ and $\hat{\mathbf{R}}$, which correspond to leave-in bias, smoothing bias, linear variance and quadratic variance, respectively.

Lemma 15. *Let Assumptions of Lemma 2 hold with the exception that f may be discontinuous across $\bar{x}$, then*

$$\hat{\mathbf{B}}_{\mathtt{S}} = h^{p+1}\left\{\frac{F^{(p+1)}(\bar{x}-)f(\bar{x}-)}{(p+1)!}\mathbf{c}_{-,p} + \frac{F^{(p+1)}(\bar{x}+)f(\bar{x}+)}{(p+1)!}\mathbf{c}_{+,p}\right\} + o_{\mathbb{P}}(h^{p+1}), \qquad \hat{\mathbf{B}}_{\mathtt{LI}} = O_{\mathbb{P}}\left(\frac{1}{n}\right), \tag{1}$$

where

$$\mathbf{c}_{-,p} = \int_{-1}^{0} u^{p+1}\mathbf{r}_{-,p}(u)K(u)\mathrm{d}u, \qquad \mathbf{c}_{+,p} = \int_{0}^{1} u^{p+1}\mathbf{r}_{+,p}(u)K(u)\mathrm{d}u.$$

Lemma 16. *Let Assumptions of Lemma 3 hold with the exception that f may be discontinuous across $\bar{x}$, then*

$$\begin{aligned}&\mathbb{V}\left[\mathbf{L}_h(x_i)\right] \\ &= F(\bar{x})\Big(1-F(\bar{x})\Big)\mathbf{S}_{f,p}\mathbf{e}_0\mathbf{e}_0'\mathbf{S}_{f,p} \\ &+ h\Big\{F(\bar{x})\Big(1-F(\bar{x})\Big)\frac{f'(\bar{x}-)}{f(\bar{x}-)} - F(\bar{x})f(\bar{x}-) + f(\bar{x}-)\Big\}\mathbf{S}_{f,p}\mathbf{e}_0\mathbf{e}_{1,-}'\mathbf{S}_{f,p} \\ &+ h\Big\{F(\bar{x})\Big(1-F(\bar{x})\Big)\frac{f'(\bar{x}-)}{f(\bar{x}-)} - F(\bar{x})f(\bar{x}-) + f(\bar{x}-)\Big\}\mathbf{S}_{f,p}\mathbf{e}_{1,-}\mathbf{e}_0'\mathbf{S}_{f,p} \\ &+ h\Big\{F(\bar{x})\Big(1-F(\bar{x})\Big)\frac{f'(\bar{x}+)}{f(\bar{x}+)} - F(\bar{x})f(\bar{x}+)\Big\}\mathbf{S}_{f,p}\mathbf{e}_0\mathbf{e}_{1,+}'\mathbf{S}_{f,p} \\ &+ h\Big\{F(\bar{x})\Big(1-F(\bar{x})\Big)\frac{f'(\bar{x}+)}{f(\bar{x}+)} - F(\bar{x})f(\bar{x}+)\Big\}\mathbf{S}_{f,p}\mathbf{e}_{1,+}\mathbf{e}_0'\mathbf{S}_{f,p} \\ &+ hf(\bar{x}-)^3\mathbf{\Psi}\mathbf{\Gamma}_{+,p}\mathbf{\Psi} + hf(\bar{x}+)^3\mathbf{\Gamma}_{+,p} + O(h^2)\end{aligned} \tag{2}$$

where

$$\begin{aligned}\mathbf{\Gamma}_{-,p} &= \iint_{[-1,0]^2}(u\wedge v)\mathbf{r}_{-,p}(u)\mathbf{r}_{-,p}(v)'K(u)K(v)\ \mathrm{d}u\mathrm{d}v, \\ \mathbf{\Gamma}_{+,p} &= \iint_{[0,1]^2}(u\wedge v)\mathbf{r}_{+,p}(u)\mathbf{r}_{+,p}(v)'K(u)K(v)\ \mathrm{d}u\mathrm{d}v,\end{aligned}$$

and

$$\mathbf{\Psi} = \begin{bmatrix} (-1)^0 & & & & & & \\ & & (-1)^1 & & & & \\ & (-1)^1 & & & & & \\ & & & (-1)^2 & & & \\ & & & & (-1)^3 & & \\ & & & & & \ddots & \\ & & & & & & (-1)^p \end{bmatrix}.$$

Remark 11. Now we consider the asymptotic variance after proper scaling. The proper scaling matrix is

$$\mathbf{N} = \mathrm{diag}\left\{\sqrt{n}, \sqrt{\frac{n}{h}}, \cdots, \sqrt{\frac{n}{h}}\right\}_{(p+2)\times(p+2)}$$

then

$$\begin{aligned} \mathbb{V}\left[\mathbf{N}\mathbf{S}_{f,p}^{-1}\hat{\mathbf{L}}\right] &= F(\bar{x})\Big(1 - F(\bar{x})\Big)\mathbf{e}_0\mathbf{e}_0' \\ &\quad + \Big(\mathbf{I} - \mathbf{e}_0\mathbf{e}_0'\Big)\mathbf{S}_{f,p}^{-1}\Big(f(\bar{x}-)^3\mathbf{\Psi}\mathbf{\Gamma}_{+,p}\mathbf{\Psi} + f(\bar{x}+)^3\mathbf{\Gamma}_{+,p}\Big)\mathbf{S}_{f,p}^{-1}\Big(\mathbf{I} - \mathbf{e}_0\mathbf{e}_0'\Big) \\ &\quad + O(\sqrt{h}). \end{aligned}$$

$$= \begin{bmatrix} F(\bar{x})\Big(1 - F(\bar{x})\Big) & \mathbf{0} \\ \mathbf{0} & \Big\{\mathbf{S}_{f,p}^{-1}\Big(f(\bar{x}-)^3\mathbf{\Psi}\mathbf{\Gamma}_{+,p}\mathbf{\Psi} + f(\bar{x}+)^3\mathbf{\Gamma}_{+,p}\Big)\mathbf{S}_{f,p}^{-1}\Big\}_{(2:p+2)} \end{bmatrix},$$

where the operator $\{\cdot\}_{(2:p+2)}$ excludes the first row and column. Therefore asymptotically

1. the c.d.f. (parametric part) and the derivatives (nonparametric part) remain to be independent;

2. the derivatives (nonparametric part) on the two sides are *not* independent.

Again we can show that the quadratic part is negligible.

Lemma 17. *Let Assumptions of Lemma 4 hold with the exception that f may not be continuous across $\bar{x}$, then*

$$\hat{\mathbf{R}} = O_{\mathbb{P}}\left(\sqrt{\frac{1}{n^2 h}}\right).$$

Theorem 7. *Let Assumptions of Theorem 1 hold with the exception that f may not be continuous across $\bar{x}$, then*

$$\sqrt{nh_n}\,\frac{\hat{f}_p(\bar{x}+) - \hat{f}_p(\bar{x}-) - \Big(f(\bar{x}+) - f(\bar{x}-)\Big) - h^p\mathscr{B}_{p,1}}{\sqrt{\mathscr{V}_{p,1}}} \rightsquigarrow \mathcal{N}(0,1),$$

where

$$\mathscr{B}_{p,1} = \frac{1}{(p+1)!}\Big(\mathbf{e}_{1,+} - \mathbf{e}_{1,-}\Big)'\mathbf{S}_{f,p}^{-1}\Big(F^{(p+1)}(\bar{x}+)f(\bar{x}+)\mathbf{c}_{+,p} + F^{(p+1)}(\bar{x}-)f(\bar{x}-)\mathbf{c}_{-,p}\Big),$$
$$\mathscr{V}_{p,1} = \Big(\mathbf{e}_{1,+} - \mathbf{e}_{1,-}\Big)'\mathbf{S}_{f,p}^{-1}\Big(f(\bar{x}-)^3\boldsymbol{\Psi}\boldsymbol{\Gamma}_{+,p}\boldsymbol{\Psi} + f(\bar{x}+)^3\boldsymbol{\Gamma}_{+,p}\Big)\mathbf{S}_{f,p}^{-1}\Big(\mathbf{e}_{1,+} - \mathbf{e}_{1,-}\Big).$$

Remark 12.
(1) Now the matrix $\mathbf{S}_{f,p}$ is no longer block diagonal, which indicates $\hat{f}_p(\bar{x}+)$ and $\hat{f}_p(\bar{x}-)$ have nonzero covariance. Therefore the asymptotic variance does not take an additive form.
(2) The standard error estimator remains valid.

4.3 Plug-in and Jackknife-based Standrd Errors

The standard error $\hat{\sigma}_{p,v,x}$ (see Theorem 2/4) is fully automatic and adapts to both interior and boundary regions. In this section we consider two other ways to construct standard errors.

4.3.1 Plug-in Standard Error

Take $v \geq 1$. Then the asymptotic variance of $\hat{F}_p^{(v)}(x)$ takes the following form:

$$\mathscr{V}_{p,v,x} = (v!)^2 H^{(1)}(x)\mathbf{e}_v'\mathbf{S}_{p,x}^{-1}\boldsymbol{\Gamma}_{p,x}\mathbf{S}_{p,x}^{-1}\mathbf{e}_v.$$

One way of constructing estimate of the above quantity is to plug-in a consistent estimator of $H^{(1)}(x)$. This may not be appealing, since $H^{(1)}(x)$ is a nonparametric object. There is, however, one case in which $H^{(1)}(x)$ is automatically available. Assume the weights are identically 1, i.e. $w_i \equiv 1$, then $H^{(1)}(x) = F^{(1)}(x) = G^{(1)}(x)$, which is simply the estimated density. Hence we can use

$$\hat{\mathscr{V}}_{p,v,x} = (v!)^2 \hat{F}_p^{(1)}(x)\mathbf{e}_v'\mathbf{S}_{p,x}^{-1}\boldsymbol{\Gamma}_{p,x}\mathbf{S}_{p,x}^{-1}\mathbf{e}_v.$$

The next question is how $\mathbf{S}_{p,x}$ and $\boldsymbol{\Gamma}_{p,x}$ should be constructed. Note that they are related to the kernel, evaluation point x and the bandwidth h, but *not* the data generating process. Therefore the three matrices can be constructed by either analytical integration or numerical method.

Back to the original case. How to estimate $H^{(1)}(x)$ with nontrivial weighting scheme? Recall that $H(x) = \mathbb{E}[w_i^2\mathbb{1}[x_i \leq x]]$. Then $\hat{H}(x) = \sum_i w_i^2\mathbb{1}[x_i \leq x]/n$ is unbiased and $\sqrt{n}$-consistent. On the other hand, it has the same problem as the empirical distribution function: it is not differentiable. Local polynomial smoothing can be applied here, the same as how it is applied to smooth out the empirical distribution function to obtain estimates of derivatives. More precisely, one can replace $\hat{F}(x_i)$ by $\hat{H}(x_i)$ in $\hat{\boldsymbol{\beta}}_p(x)$, and the slope coefficient will be consistent for $H^{(1)}(x)$, under very mild regularity conditions.

4.3.2 Jackknife-based Standard Error

The standard error $\hat{\sigma}_{p,v,x}$ is obtained by inspecting the asymptotic linear representation. It is fully automatic and adapts to both interior and boundaries. In this part, we present another standard error which resembles $\hat{\sigma}_{p,v,x}$, albeit with a different motivation.

Recall that $\hat{\boldsymbol{\beta}}_p(x)$ is essentially a second order U-statistic, and the following expansion is justified:

$$
\begin{aligned}
&\frac{1}{n}\mathbf{X}_h'\mathbf{K}_h\left(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}_p(x)\right)\\
&\quad=\frac{1}{n}\sum_i \mathbf{r}_p\left(\frac{x_i-x}{h}\right)\left(\hat{F}(x_i)-\mathbf{r}_p(x_i-x)'\boldsymbol{\beta}_p(x)\right)K_h(x_i-x)\\
&\quad=\frac{1}{n}\sum_i \mathbf{r}_p\left(\frac{x_i-x}{h}\right)\left(\frac{1}{n-1}\sum_{j;j\neq i} w_j\left(\mathbb{1}(x_j\le x_i)-\mathbf{r}_p(x_i-x)'\boldsymbol{\beta}_p(x)\right)\right)K_h(x_i-x)+O_{\mathbb{P}}\left(\frac{1}{n}\right)\\
&\quad=\frac{1}{n(n-1)}\sum_{i,j;i\neq j}\mathbf{r}_p\left(\frac{x_i-x}{h}\right)w_j\left(\mathbb{1}(x_j\le x_i)-\mathbf{r}_p(x_i-x)'\boldsymbol{\beta}_p(x)\right)K_h(x_i-x)+O_{\mathbb{P}}\left(\frac{1}{n}\right),
\end{aligned}
$$

where the remainder represents leave-in bias. Note that the above could be written as a U-statistic, and to apply the Hoeffding decomposition, define

$$
\begin{aligned}
\mathbf{U}(x_i,w_i,x_j,w_j)&=\mathbf{r}_p\left(\frac{x_i-x}{h}\right)w_j\left(\mathbb{1}(x_j\le x_i)-\mathbf{r}_p(x_i-x)'\boldsymbol{\beta}_p(x)\right)K_h(x_i-x)\\
&\quad+\mathbf{r}_p\left(\frac{x_j-x}{h}\right)w_i\left(\mathbb{1}(x_i\le x_j)-\mathbf{r}_p(x_j-x)'\boldsymbol{\beta}_p(x)\right)K_h(x_j-x),
\end{aligned}
$$

which is symmetric in its two arguments. Then (with estimated weights one has to make further expansion, but the main idea is the same)

$$
\begin{aligned}
\frac{1}{n}\mathbf{X}_h'\mathbf{K}_h\left(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}_p(x)\right)&=\mathbb{E}\left[\mathbf{U}(x_i,w_i,x_j,w_j)\right]\\
&\quad+\frac{1}{n}\sum_i\Big(\mathbf{U}_1(x_i,w_i)-\mathbb{E}\left[\mathbf{U}(x_i,w_i,x_j,w_j)\right]\Big)\\
&\quad+\binom{n}{2}^{-1}\sum_{i,j;i<j}\Big(\mathbf{U}(x_i,w_i,x_j,w_j)-\mathbf{U}_1(x_i,w_i)-\mathbf{U}_1(x_j,w_j)+\mathbb{E}\left[\mathbf{U}(x_i,w_i,x_j,w_j)\right]\Big).
\end{aligned}
$$

Here $\mathbf{U}_1(x_i)=\mathbb{E}\left[\mathbf{U}(x_i,w_i,x_j,w_j)\middle|\,x_i,w_i\right]$. The second line in the above display is the analogue of $\hat{\mathbf{L}}$, which contributes to the leading variance, and the third line is negligible. The new standard error, we call the jackknife-based standard error, is given by the following:

$$
\hat{\sigma}_{p,v,x}^{(\mathtt{JK})}\equiv(v!)\sqrt{\frac{1}{nh^{2v}}\mathbf{e}_v'\hat{\mathbf{S}}_{p,x}^{-1}\hat{\boldsymbol{\Gamma}}_{p,x}^{\mathtt{JK}}\hat{\mathbf{S}}_{p,x}^{-1}\mathbf{e}_v},
$$

with

$$
\begin{aligned}
\hat{\boldsymbol{\Gamma}}_{p,x}^{\mathtt{JK}}&=\frac{1}{n}\sum_i\left(\frac{1}{n-1}\sum_{j;j\neq i}\hat{\mathbf{U}}(x_i,w_i,x_j,w_j)\right)\left(\frac{1}{n-1}\sum_{j;j\neq i}\hat{\mathbf{U}}(x_i,w_i,x_j,w_j)\right)'\\
&\quad-\left(\binom{n}{2}^{-1}\sum_{i,j;i\neq j}\hat{\mathbf{U}}(x_i,w_i,x_j,w_j)\right)\left(\binom{n}{2}^{-1}\sum_{i,j;i\neq j}\hat{\mathbf{U}}(x_i,w_i,x_j,w_j)\right)',
\end{aligned}
$$

and

$$
\begin{aligned}
\hat{\mathbf{U}}(x_i,w_i,x_j,w_j)&=\mathbf{r}_p\left(\frac{x_i-x}{h}\right)w_j\left(\mathbb{1}(x_j\le x_i)-\mathbf{r}_p(x_i-x)'\hat{\boldsymbol{\beta}}_p(x)\right)K_h(x_i-x)\\
&\quad+\mathbf{r}_p\left(\frac{x_j-x}{h}\right)w_i\left(\mathbb{1}(x_i\le x_j)-\mathbf{r}_p(x_j-x)'\hat{\boldsymbol{\beta}}_p(x)\right)K_h(x_j-x).
\end{aligned}
$$

The name jackknife comes from the fact that we use leave-one-out "estimator" for $\mathbf{U}_1(x_i,w_i)$: with

x_i and w_i fixed,

$$\text{``}\frac{1}{n-1}\sum_{j;j\neq i}\hat{\mathbf{U}}(x_i,w_i,x_j,w_j)\to_{\mathbb{P}}\mathbf{U}_1(x_i,w_i)\text{''}.$$

Under the same conditions specified in Theorem 2, one can show that the jackknife-based standard error is consistent. For estimated weights, regularity conditions specified in Theorem 4 suffice.

5 Simulation Study

5.1 DGP 1: Truncated Normal Distribution

In this subsection, we conduct simulation study based on truncated normal distribution. To be more specific, the underlying distribution of x_i is the standard normal distribution truncated below at -0.8. We do not incorporate extra weighting, hence

$$G(x)=F(x)=\frac{\Phi(x)-\Phi(-0.8)}{1-\Phi(-0.8)},\qquad x\geq -0.8,$$

and zero otherwise. Equivalently, x_i has Lebesgue density $\Phi^{(1)}(x)/(1-\Phi(-0.8))$ on $[-0.8,\infty]$.

In this simulation study, the target parameter is the density function evaluated at various points. Note that both the variance and the bias of our estimator depend on the evaluation point, and in particular, the magnitude of the bias depends on higher order derivatives of the distribution function.

1. *Evaluation point.* We estimate the density at $x\in\{-0.8,\ -0.5,\ 0.5,\ 1.5\}$. Note that -0.8 is the boundary point, where classical density estimators such as the kernel density estimator has high bias. The point -0.5, given our bandwidth choice, is fairly close to the boundary, hence should be understood as in the lower boundary region. The two points 0.5 and 1.5 are interior, but the curvature of the normal density is quite different at those two points, and we expect to see the estimators having different bias behaviors.

2. *Polynomial order.* We consider $p\in\{2,3\}$. For density estimation using our estimators, $p=2$ should be the default choice, since it corresponds to estimating conditional mean with local linear regression. Such choice is also recommended by Fan and Gijbels (1996), according to which one should always choose $p-s=2-1=1$ to be an odd number. We include $p=3$ for completeness.

3. *Kernel function.* For local polynomial regression, the choice of kernel function is usually not very important. We use the triangular kernel $k(u)=(1-|u|)\vee 0$.

4. *Sample size.* The sample size used consists of $n\in\{1000,\ 2000\}$. For most empirical studies employing nonparametric density estimation, the sample size is well above 1000, hence $n=2000$ is more representative.

Overall, we have $4 \times 2 \times 2 = 16$ designs, and for each design, we conduct 5000 Monte Carlo repetitions.

We consider a grid of bandwidth choices, which correspond to multiples of the MSE-optimal bandwidth, ranging from $0.1h_{\texttt{MSE}}$ to $2h_{\texttt{MSE}}$. We also consider the estimated bandwidth. The MSE-optimal bandwidth, $h_{\texttt{MSE}}$, is chosen by minimizing the asymptotic mean squared error, using the true underlying distribution.

For each design, we report the empirical bias of the estimator, $\mathbb{E}[\hat{f}_p(x) - f(x)]$, under bias. And empirical standard deviations, $\mathbb{V}^{1/2}[\hat{f}_p(x)]$, and empirical root-m.s.e., under sd and $\sqrt{\text{mse}}$, respectively. For the standard errors constructed from the variance estimators, we report their empirical average under mean, which should be compared to sd. We also report the empirical rejection rate of t-statistics at 5% nominal level, under size. The t-statistic is $(\hat{f}_p(x) - \mathbb{E}\hat{f}_p(x))/\text{se}$, which is exactly centered, hence rejection rate thereof is a measure of accuracy of normal approximation.

5.2 DGP 2: Exponential Distribution

In this subsection, we conduct simulation study based on exponential distribution. To be more specific, the underlying distribution of x_i is $F(x) = 1 - e^{-x}$. We do not incorporate extra weighting. Equivalently, x_i has Lebesgue density e^{-x} for $x \geq 0$.

In this simulation study, the target parameter is the density function evaluated at various points. Note that both the variance and the bias of our estimator depend on the evaluation point, and in particular, the magnitude of the bias depends on higher order derivatives of the distribution function.

1. *Evaluation point.* We estimate the density at $x \in \{0,\ 1,\ 1.5\}$. Note that 0 is the boundary point, where classical density estimators such as the kernel density estimator has high bias. The two points 1 and 1.5 are interior.

2. *Polynomial order.* We consider $p \in \{2, 3\}$. For density estimation using our estimators, $p = 2$ should be the default choice, since it corresponds to estimating conditional mean with local linear regression. Such choice is also recommended by Fan and Gijbels (1996), according to which one should always choose $p - s = 2 - 1 = 1$ to be an odd number. We include $p = 3$ for completeness.

3. *Kernel function.* For local polynomial regression, the choice of kernel function is usually not very important. We use the triangular kernel $k(u) = (1 - |u|) \vee 0$.

4. *Sample size.* The sample size used consists of $n \in \{1000,\ 2000\}$. For most empirical studies employing nonparametric density estimation, the sample size is well above 1000, hence $n = 2000$ is more representative.

Overall, we have $3 \times 2 \times 2 = 12$ designs, and for each design, we conduct 5000 Monte Carlo repetitions.

We consider a grid of bandwidth choices, which correspond to multiples of the MSE-optimal bandwidth, ranging from $0.1h_{\mathtt{MSE}}$ to $2h_{\mathtt{MSE}}$. We also consider the estimated bandwidth. The MSE-optimal bandwidth, $h_{\mathtt{MSE}}$, is chosen by minimizing the asymptotic mean squared error, using the true underlying distribution.

For each design, we report the empirical bias of the estimator, $\mathbb{E}[\hat{f}_p(x) - f(x)]$, under bias. And empirical standard deviations, $\mathbb{V}^{1/2}[\hat{f}_p(x)]$, and empirical root-m.s.e., under sd and $\sqrt{\text{mse}}$, respectively. For the standard errors constructed from the variance estimators, we report their empirical average under mean, which should be compared to sd. We also report the empirical rejection rate of t-statistics at 5% nominal level, under size. The t-statistic is $(\hat{f}_p(x) - \mathbb{E}\hat{f}_p(x))/\text{se}$, which is exactly centered, hence rejection rate thereof is a measure of accuracy of normal approximation.

References

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

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

6 Proof

6.1 Proof of Lemma 1

Proof. A generic element of the matrix $\frac{1}{n}\mathbf{X}_h'\mathbf{K}_h\mathbf{X}_h$ takes the form:

$$\frac{1}{n}\sum_i \frac{1}{h}\left(\frac{x_i - x}{h}\right)^s K\left(\frac{x_i - x}{h}\right), \qquad 0 \le s \le 2p.$$

Then we compute the expectation:

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{n}\sum_i \frac{1}{h}\left(\frac{x_i - x}{h}\right)^s K\left(\frac{x_i - x}{h}\right)\right] = \mathbb{E}\left[\frac{1}{h}\left(\frac{x_i - x}{h}\right)^s K\left(\frac{x_i - x}{h}\right)\right] \\
&= \int_{x_\mathtt{L}}^{x_\mathtt{U}} \frac{1}{h}\left(\frac{u - x}{h}\right)^s K\left(\frac{u - x}{h}\right) g(u)\mathrm{d}u \\
&= \int_{\frac{x_\mathtt{L}-x}{h}}^{\frac{x_\mathtt{U}-x}{h}} v^s K(v)\, g(x + vh)\mathrm{d}v = \int_{\frac{x_\mathtt{L}-x}{h}}^{\frac{x_\mathtt{U}-x}{h}} v^s K(v)\, g(x + vh)\mathrm{d}v,
\end{aligned}$$

hence for x in the interior,

$$\mathbb{E}\left[\frac{1}{n}\sum_i \frac{1}{h}\left(\frac{x_i - x}{h}\right)^s K\left(\frac{x_i - x}{h}\right)\right] = g(x)\int_{\mathbb{R}} \mathbf{r}_p(v)\mathbf{r}_p(v)' K(v)\mathrm{d}v + o(1),$$

and for $x = x_\mathtt{L} + ch$ with $c \in [0,1]$,

$$\mathbb{E}\left[\frac{1}{n}\sum_i \frac{1}{h}\left(\frac{x_i - x}{h}\right)^s K\left(\frac{x_i - x}{h}\right)\right] = g(x_\mathtt{L})\int_{-c}^{\infty} \mathbf{r}_p(v)\mathbf{r}_p(v)' K(v)\mathrm{d}v + o(1),$$

and for $x = x_\mathtt{U} - ch$ with $c \in [0,1]$,

$$\mathbb{E}\left[\frac{1}{n}\sum_i \frac{1}{h}\left(\frac{x_i - x}{h}\right)^s K\left(\frac{x_i - x}{h}\right)\right] = g(x_\mathtt{U})\int_{-\infty}^{c} \mathbf{r}_p(v)\mathbf{r}_p(v)' K(v)\mathrm{d}v + o(1),$$

provided that $G \in \mathcal{C}^1$.

The variance satisfies

$$\begin{aligned}
&\mathbb{V}\left[\frac{1}{n}\sum_i \frac{1}{h}\left(\frac{x_i - x}{h}\right)^s K\left(\frac{x_i - x}{h}\right)\right] = \frac{1}{n}\mathbb{V}\left[\frac{1}{h}\left(\frac{x_i - x}{h}\right)^s K\left(\frac{x_i - x}{h}\right)\right] \\
&\le \frac{1}{n}\mathbb{E}\left[\frac{1}{h^2}\left(\frac{x_i - x}{h}\right)^{2s} K\left(\frac{x_i - x}{h}\right)^2\right] = O\left(\frac{1}{nh}\right),
\end{aligned}$$

provided that $G \in \mathcal{C}^1$. ■

6.2 Proof of Lemma 2

Proof. First consider the smoothing bias. The leading term can be easily obtain by taking expectation together with Taylor expansion of F to power $p+1$. The variance of this term has order $n^{-1}h^{-1}h^{2p+2}$, which gives the residual estimate $o_\mathbb{P}(h^{p+1})$ since it is assumed that $nh \to \infty$.

Next for the leave-in bias, note that it has expectation of order n^{-1}, and variance of order $n^{-3}h^{-1}$, hence overall this term of order $O_\mathbb{P}(n^{-1})$. ■

6.3 Proof of Lemma 3

Proof. We first compute the variance. Note that

$$\begin{aligned}
&\int_{\frac{x_{\mathtt{L}}-x}{h}}^{\frac{x_{\mathtt{U}}-x}{h}} \mathbf{r}_p(u)\Big(\hat{F}(x+hu)-F(x+hu)\Big)K(u)g(x+hu)\mathrm{d}u \\
&= \frac{1+o_{\mathbb{P}}(1)}{n}\int_{\frac{x_{\mathtt{L}}-x}{h}}^{\frac{x_{\mathtt{U}}-x}{h}} \mathbf{r}_p(u)\,w_i\Big(\mathbb{1}[x_i\le x+hu]-F(x+hu)\Big)K(u)g(x+hu)\mathrm{d}u,
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{V}\left[\int_{\frac{x_{\mathtt{L}}-x}{h}}^{\frac{x_{\mathtt{U}}-x}{h}} \mathbf{r}_p(u)\,w_i\Big(\mathbb{1}[x_i\le x+hu]-F(x+hu)\Big)K(u)g(x+hu)\mathrm{d}u\right] \\
&= \iint_{\frac{x_{\mathtt{L}}-x}{h}}^{\frac{x_{\mathtt{U}}-x}{h}} \mathbf{r}_p(u)\,\mathbf{r}_p(v)'\,K(u)K(v)g(x+hu)g(x+hv) \\
&\qquad\left[\int_{\mathbb{R}} w_2(t)\left(\mathbb{1}[t\le x+hu]-F(x+hu)\right)\left(\mathbb{1}[t\le x+hv]-F(x+hv)\right)g(t)\mathrm{d}t\right]\mathrm{d}u\mathrm{d}v. \qquad \text{(I)}
\end{aligned}$$

For notational simplicity, let

$$H(u)=\mathbb{E}[w_i^2\mathbb{1}[x_i\le u]]=\int_{x_{\mathtt{L}}}^{u} w_2(t)g(t)\mathrm{d}t.$$

Then

$$\begin{aligned}
\text{(I)} &= \iint_{\frac{x_{\mathtt{L}}-x}{h}}^{\frac{x_{\mathtt{U}}-x}{h}} \mathbf{r}_p(u)\,\mathbf{r}_p(v)'\,K(u)K(v)g(x+hu)g(x+hv) \\
&\Big(H(x+h(u\wedge v))-H(x+hu)F(x+hv)-F(x+hu)H(x+hv)+H(x_{\mathtt{U}})F(x+hu)F(x+hv)\Big)\mathrm{d}u\mathrm{d}v.
\end{aligned}$$

We first consider the interior case, where the above reduces to:

$$\begin{aligned}
&\text{(I)}_{\text{interior}} \\
&= \iint_{\mathbb{R}} \mathbf{r}_p(u)\,\mathbf{r}_p(v)'\,K(u)K(v)g(x)^2\Big(H(x)-2H(x)F(x)+H(x_{\mathtt{U}})F(x)^2\Big)\mathrm{d}u\mathrm{d}v \\
&\quad + h\iint_{\mathbb{R}}(u\wedge v)\mathbf{r}_p(u)\,\mathbf{r}_p(v)'\,K(u)K(v)g(x)^2H^{(1)}(x)\mathrm{d}u\mathrm{d}v \\
&\quad - h\iint_{\mathbb{R}}(u+v)\mathbf{r}_p(u)\,\mathbf{r}_p(v)'\,K(u)K(v)g(x)^2\Big(H^{(1)}(x)F(x)+H(x)F^{(1)}(x)\Big)\mathrm{d}u\mathrm{d}v \\
&\quad + h\iint_{\mathbb{R}}(u+v)\mathbf{r}_p(u)\,\mathbf{r}_p(v)'\,K(u)K(v)g(x)^2H(x_{\mathtt{U}})F^{(1)}(x)F(x)\mathrm{d}u\mathrm{d}v \\
&\quad + h\iint_{\mathbb{R}}(u+v)\mathbf{r}_p(u)\,\mathbf{r}_p(v)'\,K(u)K(v)G^{(2)}(x)^2\Big(H(x)-2H(x)F(x)+H(x_{\mathtt{U}})F(x)^2\Big)\mathrm{d}u\mathrm{d}v+o(h) \\
&= g(x)^2\Big(H(x)-2H(x)F(x)+H(x_{\mathtt{U}})F(x)^2\Big)\mathbf{S}_{p,x}\mathbf{e}_0\mathbf{e}_0'\mathbf{S}_{p,x} \\
&\quad - hg(x)^2\Big(H^{(1)}(x)F(x)+H(x)F^{(1)}(x)\Big)\mathbf{S}_{p,x}(\mathbf{e}_1\mathbf{e}_0'+\mathbf{e}_0\mathbf{e}_1')\mathbf{S}_{p,x} \\
&\quad + hg(x)^2H(x_{\mathtt{U}})F^{(1)}(x)F(x)\mathbf{S}_{p,x}(\mathbf{e}_1\mathbf{e}_0'+\mathbf{e}_0\mathbf{e}_1')\mathbf{S}_{p,x} \\
&\quad + hG^{(2)}(x)^2\Big(H(x)-2H(x)F(x)+H(x_{\mathtt{U}})F(x)^2\Big)\mathbf{S}_{p,x}(\mathbf{e}_1\mathbf{e}_0'+\mathbf{e}_0\mathbf{e}_1')\mathbf{S}_{p,x} \\
&\quad + hg(x)^2H^{(1)}(x)\mathbf{\Gamma}_{p,x}+o(h).
\end{aligned}$$

For $x = x_{\mathtt{L}} + hc$ with $c \in [0, 1)$ in the lower boundary region,

$$
\begin{aligned}
&(\mathrm{I})_{\text{lower boundary}}\\
&= \iint_{\mathbb{R}} \mathbf{r}_p(u)\mathbf{r}_p(v)' K(u)K(v)g(x_{\mathtt{L}})^2\Big(H(x_{\mathtt{L}}) - 2H(x_{\mathtt{L}})F(x_{\mathtt{L}}) + H(x_{\mathtt{U}})F(x_{\mathtt{L}})^2\Big)\mathrm{d}u\mathrm{d}v\\
&\quad + h\iint_{\mathbb{R}} (u \wedge v + c)\mathbf{r}_p(u)\mathbf{r}_p(v)' K(u)K(v)g(x_{\mathtt{L}})^2 H^{(1)}(x_{\mathtt{L}})\mathrm{d}u\mathrm{d}v\\
&\quad - h\iint_{\mathbb{R}} (u + v + 2c)\mathbf{r}_p(u)\mathbf{r}_p(v)' K(u)K(v)g(x_{\mathtt{L}})^2 \Big(H^{(1)}(x_{\mathtt{L}})F(x_{\mathtt{L}}) + H(x_{\mathtt{L}})F^{(1)}(x_{\mathtt{L}})\Big)\mathrm{d}u\mathrm{d}v\\
&\quad + h\iint_{\mathbb{R}} (u + v + 2c)\mathbf{r}_p(u)\mathbf{r}_p(v)' K(u)K(v)g(x_{\mathtt{L}})^2 H(x_{\mathtt{U}})F^{(1)}(x_{\mathtt{L}})F(x_{\mathtt{L}})\mathrm{d}u\mathrm{d}v\\
&\quad + h\iint_{\mathbb{R}} (u + v + 2c)\mathbf{r}_p(u)\mathbf{r}_p(v)' K(u)K(v)G^{(2)}(x_{\mathtt{L}})^2\Big(H(x_{\mathtt{L}}) - 2H(x_{\mathtt{L}})F(x_{\mathtt{L}}) + H(x_{\mathtt{U}})F(x_{\mathtt{L}})^2\Big)\mathrm{d}u\mathrm{d}v + o(h)\\
&= hg(x_{\mathtt{L}})^2 H^{(1)}(x_{\mathtt{L}})\left(\mathbf{\Gamma}_{p,x} + c\mathbf{S}_{p,x}\mathbf{e}_0\mathbf{e}_0'\mathbf{S}_{p,x}\right) + o(h).
\end{aligned}
$$

Finally, we have

$$
\begin{aligned}
&(\mathrm{I})_{\text{upper boundary}}\\
&= \iint_{\mathbb{R}} \mathbf{r}_p(u)\mathbf{r}_p(v)' K(u)K(v)g(x_{\mathtt{U}})^2\Big(H(x_{\mathtt{U}}) - 2H(x_{\mathtt{U}})F(x_{\mathtt{U}}) + H(x_{\mathtt{U}})F(x_{\mathtt{U}})^2\Big)\mathrm{d}u\mathrm{d}v\\
&\quad + h\iint_{\mathbb{R}} (u \wedge v - c)\mathbf{r}_p(u)\mathbf{r}_p(v)' K(u)K(v)g(x_{\mathtt{U}})^2 H^{(1)}(x_{\mathtt{U}})\mathrm{d}u\mathrm{d}v\\
&\quad - h\iint_{\mathbb{R}} (u + v - 2c)\mathbf{r}_p(u)\mathbf{r}_p(v)' K(u)K(v)g(x_{\mathtt{U}})^2 \Big(H^{(1)}(x_{\mathtt{U}})F(x_{\mathtt{U}}) + H(x_{\mathtt{U}})F^{(1)}(x_{\mathtt{U}})\Big)\mathrm{d}u\mathrm{d}v\\
&\quad + h\iint_{\mathbb{R}} (u + v - 2c)\mathbf{r}_p(u)\mathbf{r}_p(v)' K(u)K(v)g(x_{\mathtt{U}})^2 H(x_{\mathtt{U}})F^{(1)}(x_{\mathtt{U}})F(x_{\mathtt{U}})\mathrm{d}u\mathrm{d}v\\
&\quad + h\iint_{\mathbb{R}} (u + v - 2c)\mathbf{r}_p(u)\mathbf{r}_p(v)' K(u)K(v)G^{(2)}(x_{\mathtt{U}})^2\Big(H(x_{\mathtt{U}}) - 2H(x_{\mathtt{U}})F(x_{\mathtt{U}}) + H(x_{\mathtt{U}})F(x_{\mathtt{U}})^2\Big)\mathrm{d}u\mathrm{d}v + o(h)\\
&= hg(x_{\mathtt{U}})^2 H^{(1)}(x_{\mathtt{U}})\left(\mathbf{\Gamma}_{p,x} + c\mathbf{S}_{p,x}\mathbf{e}_0\mathbf{e}_0'\mathbf{S}_{p,x} - \mathbf{S}_{p,x}(\mathbf{e}_1\mathbf{e}_0' + \mathbf{e}_0\mathbf{e}_1')\mathbf{S}_{p,x}\right) + o(h).
\end{aligned}
$$

With the above results, it is easy to varify the variance formula, provided that we can show the asymptotic normality.

We first consider the interior case, and verify the Lindeberg condition on the fourth moment. Let $\boldsymbol{\alpha} \in \mathbb{R}^{p+1}$ be an arbitrary nonzero vector, then

$$
\begin{aligned}
&\sum_i \mathbb{E}\left(\frac{1}{\sqrt{n}}\boldsymbol{\alpha}'\mathbf{N}_x(g(x)\mathbf{S}_{p,x})^{-1}\int_{\frac{x_{\mathtt{L}}-x}{h}}^{\frac{x_{\mathtt{U}}-x}{h}} \mathbf{r}_p(u)\, w_i\Big(\mathbb{1}[x_i \le x + hu] - F(x + hu)\Big)K(u)g(x + hu)\mathrm{d}u\right)^4\\
&= \frac{1}{n}\mathbb{E}\left(\boldsymbol{\alpha}'\mathbf{N}_x(g(x)\mathbf{S}_{p,x})^{-1}\int_{\frac{x_{\mathtt{L}}-x}{h}}^{\frac{x_{\mathtt{U}}-x}{h}} \mathbf{r}_p(u)\, w_i\Big(\mathbb{1}[x_i \le x + hu] - F(x + hu)\Big)K(u)g(x + hu)\mathrm{d}u\right)^4\\
&= \frac{1}{n}\iiiint_{\mathcal{A}} \prod_{j=1,2,3,4}\Big(\boldsymbol{\alpha}'\mathbf{N}_x(g(x)\mathbf{S}_{p,x})^{-1}\mathbf{r}_p(u_j)\, K(u_j)\Big)g(x + hu_j)\\
&\qquad \left[\int_{\mathbb{R}} w_4(t)\prod_{j=1,2,3,4}\Big(\mathbb{1}[t \le x + hu_j] - F(x + hu_j)\Big)g(t)\mathrm{d}t\right]\mathrm{d}u_1\mathrm{d}u_2\mathrm{d}u_3\mathrm{d}u_4\\
&\le \frac{C}{n}\cdot\iiiint_{\mathcal{A}} \prod_{j=1,2,3,4}\Big(\boldsymbol{\alpha}'\mathbf{N}_x(g(x)\mathbf{S}_{p,x})^{-1}\mathbf{r}_p(u_j)\, K(u_j)\Big)g(x)\mathrm{d}u_1\mathrm{d}u_2\mathrm{d}u_3\mathrm{d}u_4 + O\left(\frac{1}{nh}\right),
\end{aligned}
$$

where $\mathcal{A} = [\frac{x_{\mathtt{L}}-x}{h}, \frac{x_{\mathtt{U}}-x}{h}]^4 \subset \mathbb{R}^4$. The first term in the above display is asymptotically negligible, since it is takes the form $C \cdot (\boldsymbol{\alpha}'\mathbf{N}_x\mathbf{e}_0)^4/n$ where the constant C depends on the DGP, and is finite since we assumed $\mathbb{E}[w_i^4] < \infty$. The order of the next term is $1/(nh)$, which comes from multiplying n^{-1}, h^{-2} (from the scaling matrix $\mathbf{N}_x$), and h (from linearization), hence is also negligible.

Under the assumption that $nh \to \infty$, the Lindeberg condition is verified for interior case. The same logic applies to the boundary case, whose proof is easier than the interior case, since the leading term in the calculation is identically zero for x in either the lower or upper boundary. ■

6.4 Proof of Lemma 4

Proof. For $\hat{\mathbf{R}}$, we rewrite it as a second order degenerate U-statistic:

$$\hat{\mathbf{R}} = \frac{1+o_{\mathbb{P}}(1)}{n^2}\sum_{i,j;i<j}\hat{\mathbf{U}}_{ij},$$

where

$$\begin{aligned}\hat{\mathbf{U}}_{ij} &= \mathbf{r}_p\left(\frac{x_i-x}{h}\right)w_j\Big(\mathbb{1}[x_j\le x_i]-F(x_i)\Big)K_h(x_i-x)\\ &+\mathbf{r}_p\left(\frac{x_j-x}{h}\right)w_i\Big(\mathbb{1}[x_i\le x_j]-F(x_j)\Big)K_h(x_j-x)\\ &-\mathbb{E}\left[\mathbf{r}_p\left(\frac{x_i-x}{h}\right)w_j\Big(\mathbb{1}[x_j\le x_i]-F(x_i)\Big)K_h(x_i-x)\Big|x_j,w_j\right]\\ &-\mathbb{E}\left[\mathbf{r}_p\left(\frac{x_j-x}{h}\right)w_i\Big(\mathbb{1}[x_i\le x_j]-F(x_j)\Big)K_h(x_j-x)\Big|x_i,w_i\right].\end{aligned}$$

To compute the leading term, it suffices to consider

$$\begin{aligned}&2\mathbb{E}\left[\mathbf{r}_p\left(\frac{x_i-x}{h}\right)\mathbf{r}_p\left(\frac{x_i-x}{h}\right)'w_j^2\left(\mathbb{1}[x_j\le x_i]-F(x_i)\right)^2K_h(x_i-x)^2\right]\\ &=2\mathbb{E}\left[\mathbf{r}_p\left(\frac{x_i-x}{h}\right)\mathbf{r}_p\left(\frac{x_i-x}{h}\right)'\Big(H(x_i)-2H(x_i)F(x_i)+H(x_{\mathtt{U}})F(x_i)^2\Big)K_h(x_i-x)^2\right]\\ &=\frac{2}{h}\int_{\frac{x_{\mathtt{L}}-x}{h}}^{\frac{x_{\mathtt{U}}-x}{h}}\mathbf{r}_p(v)\mathbf{r}_p(v)'\Big(H(x+hv)-2H(x+hv)F(x+hv)+H(x_{\mathtt{U}})F(x+hv)^2\Big)K(v)^2g(x+hv)\mathrm{d}v\\ &=\frac{2}{h}\int_{\frac{x_{\mathtt{L}}-x}{h}}^{\frac{x_{\mathtt{U}}-x}{h}}\mathbf{r}_p(v)\mathbf{r}_p(v)'\Big(H(x)-2H(x)F(x)+H(x_{\mathtt{U}})F(x)^2\Big)K(v)^2g(x)\mathrm{d}v+O(1)\\ &=_{\text{interior}}\frac{2}{h}g(x)\left[H(x)-2H(x)F(x)+H(x_{\mathtt{U}})F(x)^2\right]\mathbf{T}_{p,x}+O(1),\\ &=_{\text{boundary}}O(1),\end{aligned}$$

which closes the proof. ■

6.5 Proof of Theorem 1

Proof. This follows from previous lemmas. ■

6.6 Proof of Theorem 2

Proof. First we note that the second half of the theorem follows from the first half and the asymptotic normality result of Theorem 1, hence it suffices to prove the first half, i.e. the consistency of $\hat{\mathscr{V}}_{p,v,x}$.

The analysis of this estimator is quite involved, since it takes the form of a third order V-statistic. Moreover, since the empirical d.f. $\hat{F}$ is involved in the formula, a full expansion leads to a fifth order V-statistic. However, some simple tricks will greatly simplify the problem.

We first split $\hat{\mathbf{\Gamma}}_{p,x}$ into four terms, respectively

$$\hat{\mathbf{\Sigma}}_{p,x,1}=\frac{1}{n^3}\sum_{i,j,k}\mathbf{r}_p\left(\frac{x_j-x}{h}\right)\mathbf{r}_p\left(\frac{x_k-x}{h}\right)'K_h(x_j-x)K_h(x_k-x)w_i^2\Big(\mathbb{1}[x_i\le x_j]-F(x_j)\Big)\Big(\mathbb{1}[x_i\le x_k]-F(x_k)\Big)$$

$$\hat{\mathbf{\Sigma}}_{p,x,2}=\frac{1}{n^3}\sum_{i,j,k}\mathbf{r}_p\left(\frac{x_j-x}{h}\right)\mathbf{r}_p\left(\frac{x_k-x}{h}\right)'K_h(x_j-x)K_h(x_k-x)w_i^2\Big(F(x_j)-\hat{F}(x_j)\Big)\Big(\mathbb{1}[x_i\le x_k]-\hat{F}(x_k)\Big)$$

$$\hat{\mathbf{\Sigma}}_{p,x,3}=\frac{1}{n^3}\sum_{i,j,k}\mathbf{r}_p\left(\frac{x_j-x}{h}\right)\mathbf{r}_p\left(\frac{x_k-x}{h}\right)'K_h(x_j-x)K_h(x_k-x)w_i^2\Big(\mathbb{1}[x_i\le x_j]-\hat{F}(x_j)\Big)\Big(F(x_k)-\hat{F}(x_k)\Big)$$

$$\hat{\mathbf{\Sigma}}_{p,x,4}=\frac{1}{n^3}\sum_{i,j,k}\mathbf{r}_p\left(\frac{x_j-x}{h}\right)\mathbf{r}_p\left(\frac{x_k-x}{h}\right)'K_h(x_j-x)K_h(x_k-x)w_i^2\Big(F(x_j)-\hat{F}(x_j)\Big)\Big(F(x_k)-\hat{F}(x_k)\Big).$$

Leaving $\hat{\boldsymbol{\Sigma}}_{p,x,1}$ for a while, since it is the key component in this variance estimator. We first consider $\mathbf{N}_x\hat{\mathbf{S}}_{p,x}^{-1}\hat{\boldsymbol{\Sigma}}_{p,x,4}\hat{\mathbf{S}}_{p,x}^{-1}\mathbf{N}_x$. By the uniform consistency of the empirical d.f., it can be shown easily that

$$\mathbf{N}_x\hat{\mathbf{S}}_{p,x}^{-1}\hat{\boldsymbol{\Sigma}}_{p,x,4}\hat{\mathbf{S}}_{p,x}^{-1}\mathbf{N}_x = O_{\mathbb{P}}\left((nh)^{-1}\right).$$

Note that the extra h^{-1} comes from the scaling matrix $\mathbf{N}_x$, but not the kernel function K_h. Next we consider $\mathbf{N}_x\hat{\mathbf{S}}_{p,x}^{-1}\hat{\boldsymbol{\Sigma}}_{p,x,2}\hat{\mathbf{S}}_{p,x}^{-1}\mathbf{N}_x$, which takes the following form (up to the negligible smoothing bias):

$$\begin{aligned}&\mathbf{N}_x\hat{\mathbf{S}}_{p,x}^{-1}\hat{\boldsymbol{\Sigma}}_{p,x,2}\hat{\mathbf{S}}_{p,x}^{-1}\mathbf{N}_x\\ =&\mathbf{N}_x\mathbf{H}(\boldsymbol{\beta}_p(x)-\hat{\boldsymbol{\beta}}_p(x))\left(\frac{1}{n^2}\sum_{i,k}\mathbf{r}_p\left(\frac{x_k-x}{h}\right)'K_h(x_k-x)w_i^2\Big(\mathbb{1}[x_i\le x_k]-\hat{F}(x_k)\Big)\right)\hat{\mathbf{S}}_{p,x}^{-1}\mathbf{N}_x\\ =&O_{\mathbb{P}}((nh)^{-1/2})=o_{\mathbb{P}}(1),\end{aligned}$$

where the last line uses the asymptotic normality of $\hat{\boldsymbol{\beta}}_p(x)$. For $\hat{\boldsymbol{\Sigma}}_{p,x,1}$, we make the observation that it is possible to ignore all "diagonal" terms, meaning that

$$\hat{\boldsymbol{\Sigma}}_{p,x,1}=\frac{1}{n^3}\sum_{\substack{i,j,k\\ \text{distinct}}}\mathbf{r}_p\left(\frac{x_j-x}{h}\right)\mathbf{r}_p\left(\frac{x_k-x}{h}\right)'K_h(x_j-x)K_h(x_k-x)w_i^2\Big(\mathbb{1}[x_i\le x_j]-F(x_j)\Big)\Big(\mathbb{1}[x_i\le x_k]-F(x_k)\Big)+o_{\mathbb{P}}(h),$$

under the assumption that $nh^2\to\infty$. As a surrogate, define

$$\mathbf{U}_{i,j,k}=\mathbf{r}_p\left(\frac{x_j-x}{h}\right)\mathbf{r}_p\left(\frac{x_k-x}{h}\right)'K_h(x_j-x)K_h(x_k-x)w_i^2\Big(\mathbb{1}[x_i\le x_j]-F(x_j)\Big)\Big(\mathbb{1}[x_i\le x_k]-F(x_k)\Big),$$

which means

$$\hat{\boldsymbol{\Sigma}}_{p,x,1}=\frac{1}{n^3}\sum_{\substack{i,j,k\\ \text{distinct}}}\mathbf{U}_{i,j,k}.$$

The critical step is to further decompose the above into

$$\begin{aligned}\hat{\boldsymbol{\Sigma}}_{p,x,1}=&\frac{1}{n^3}\sum_{\substack{i,j,k\\ \text{distinct}}}\mathbb{E}[\mathbf{U}_{i,j,k}|x_i] &\text{(I)}\\ &+\frac{1}{n^3}\sum_{\substack{i,j,k\\ \text{distinct}}}\Big(\mathbf{U}_{i,j,k}-\mathbb{E}[\mathbf{U}_{i,j,k}|x_i,x_j]\Big) &\text{(II)}\\ &+\frac{1}{n^3}\sum_{\substack{i,j,k\\ \text{distinct}}}\Big(\mathbb{E}[\mathbf{U}_{i,j,k}|x_i,x_j]-\mathbb{E}[\mathbf{U}_{i,j,k}|x_i]\Big). &\text{(III)}\end{aligned}$$

We already investigated the properties of term (I) in Lemma 3, hence it remains to show that both (II) and (III) are $o(h)$, hence does not affect the estimation of asymptotic variance. We consider (II) as an example, and the analysis of (III) is similar. Since (II) has zero expectation, we consider its variance (for simplicity treat $\mathbf{U}$ as a scaler):

$$\mathbb{V}[\text{(II)}]=\mathbb{E}\left[\frac{1}{n^6}\sum_{\substack{i,j,k\\ \text{distinct}}}\sum_{\substack{i',j',k'\\ \text{distinct}}}\Big(\mathbf{U}_{i,j,k}-\mathbb{E}[\mathbf{U}_{i,j,k}|x_i,x_j]\Big)\Big(\mathbf{U}_{i,j,k}-\mathbb{E}[\mathbf{U}_{i',j',k'}|x_{i'},x_{j'}]\Big)\right].$$

The expectation will be zero if the six indices are all distinct. Similarly, when there are only two indices among the six are equal, the expectation will be zero *unless* $k = k'$, hence

$$\mathbb{V}[(\mathrm{II})] = \mathbb{E}\left[\frac{1}{n^6}\sum_{\substack{i,j,k\\ \text{distinct}}}\sum_{\substack{i',j',k'\\ \text{distinct}}}\Big(\mathbf{U}_{i,j,k} - \mathbb{E}[\mathbf{U}_{i,j,k}|x_i,x_j]\Big)\Big(\mathbf{U}_{i,j,k} - \mathbb{E}[\mathbf{U}_{i',j',k'}|x_{i'},x_{j'}]\Big)\right]$$
$$= \mathbb{E}\left[\frac{1}{n^6}\sum_{\substack{i,j,k,i'j'\\ \text{distinct}}}\Big(\mathbf{U}_{i,j,k} - \mathbb{E}[\mathbf{U}_{i,j,k}|x_i,x_j]\Big)\Big(\mathbf{U}_{i,j,k} - \mathbb{E}[\mathbf{U}_{i',j',k}|x_{i'},x_{j'}]\Big)\right]$$
$$+ \cdots,$$

where $\cdots$ represent cases where more than two indices among the six are equal. We can easily compute the order from the above as

$$\mathbb{V}[(\mathrm{II})] = O(n^{-1}) + O((nh)^{-2}),$$

which shows that

$$(\mathrm{II}) = O_{\mathbb{P}}(n^{-1/2} + (nh)^{-1}) = o_{\mathbb{P}}(h),$$

which closes the proof. ■

6.7 Proof of Lemma 5

Proof. First replace $\hat{w}_i$ by w_i, then it reduces to the leave-in bias, which has been shown in Lemma 2 to have order $O_{\mathbb{P}}(n^{-1})$. Then consider the remaining piece:

$$\frac{1+o_{\mathbb{P}}(1)}{n^2}\sum_i \mathbf{r}_p\left(\frac{x_i-x}{h}\right)(\hat{w}_i - w_i)\Big(1-F(x_i)\Big)K_h(x_i-x) \tag{I}$$
$$\le (\hat{\theta}-\theta_0)\cdot\frac{1}{n^2}\sum_i \sup_{|\theta-\theta_0|\le\delta}|\dot{w}(z_i,\theta)|\cdot\left|\mathbf{r}_p\left(\frac{x_i-x}{h}\right)\right|\Big(1-F(x_i)\Big)K_h(x_i-x)$$
$$\le \frac{1}{\sqrt{n}n}\left[\frac{1}{n}\sum_i \sup_{|\theta-\theta_0|\le\delta}|\dot{w}(z_i,\theta)|\cdot\left|\mathbf{r}_p\left(\frac{x_i-x}{h}\right)\right|\Big(1-F(x_i)\Big)K_h(x_i-x)\right],$$

where the remaining term can be further bounded by expectation calculation:

$$\mathbb{E}\left[\sup_{|\theta-\theta_0|\le\delta}|\dot{w}(z_i,\theta)|\cdot\left|\mathbf{r}_p\left(\frac{x_i-x}{h}\right)\right|\Big(1-F(x_i)\Big)K_h(x_i-x)\right]$$
$$= \mathbb{E}\left[a(x_i)\cdot\left|\mathbf{r}_p\left(\frac{x_i-x}{h}\right)\right|\Big(1-F(x_i)\Big)K_h(x_i-x)\right], \qquad a(x_i) = \mathbb{E}\left[\sup_{|\theta-\theta_0|\le\delta}|\dot{w}(z_i,\theta)|\,\Big|x_i\right]$$
$$= \int a(x+hv)\cdot|\mathbf{r}_p(v)|\Big(1-F(x+hv)\Big)K(v)g(x+hv)\mathrm{d}v$$
$$\le C\int_{-1}^{1} a(x+hv)g(x+hv)\mathrm{d}v$$
$$\le C\mathbb{E}\left[\sup_{|\theta-\theta_0|\le\delta}|\dot{w}(z_i,\theta)\right] < \infty.$$

Therefore (I) has order $O_{\mathbb{P}}(n^{-3/2})$, hence is asymptotically negligible. ■

6.8 Proof of Lemma 6

Proof. First replace $\hat{w}_i$ by w_i, then Lemma 4 shows that it has order $O_{\mathbb{P}}(\sqrt{n^{-2}h^{-1}}) = o_{\mathbb{P}}(\sqrt{n^{-1}h})$, provided that $nh^2 \to \infty$. Then we consider the difference, which can be written as (ignoring the $o_{\mathbb{P}}(1)$ term in $\hat{\mathbf{R}}$)

$$\begin{aligned}
&\frac{1}{n^2}\sum_{i,j;i\neq j}(\hat{w}_j - w_j)\Big\{\mathbf{r}_p\left(\frac{x_i-x}{h}\right)\Big(\mathbb{1}[x_j \le x_i] - F(x_i)\Big)K_h(x_i - x) \\
&\qquad - \int_{\frac{x_{\mathrm{L}}-x}{h}}^{\frac{x_{\mathrm{U}}-x}{h}} \mathbf{r}_p(u)\Big(\mathbb{1}[x_j \le x+hu] - F(x+hu)\Big)K(u)g(x+hu)\mathrm{d}u\Big\} \\
&= \Big(\hat{\theta}-\theta_0\Big)\frac{1}{n^2}\sum_{i,j;i\neq j}\dot{w}_j\Big\{\mathbf{r}_p\left(\frac{x_i-x}{h}\right)\Big(\mathbb{1}[x_j \le x_i] - F(x_i)\Big)K_h(x_i - x) && \text{(I)}\\
&\qquad - \int_{\frac{x_{\mathrm{L}}-x}{h}}^{\frac{x_{\mathrm{U}}-x}{h}} \mathbf{r}_p(u)\Big(\mathbb{1}[x_j \le x+hu] - F(x+hu)\Big)K(u)g(x+hu)\mathrm{d}u\Big\} \\
&\quad + \Big(\hat{\theta}-\theta_0\Big)\frac{1}{n^2}\sum_{i,j;i\neq j}(\dot{w}(\mathbf{z}_j,\tilde{\theta}) - \dot{w}_j)\Big\{\mathbf{r}_p\left(\frac{x_i-x}{h}\right)\Big(\mathbb{1}[x_j \le x_i] - F(x_i)\Big)K_h(x_i - x) && \text{(II)}\\
&\qquad - \int_{\frac{x_{\mathrm{L}}-x}{h}}^{\frac{x_{\mathrm{U}}-x}{h}} \mathbf{r}_p(u)\Big(\mathbb{1}[x_j \le x+hu] - F(x+hu)\Big)K(u)g(x+hu)\mathrm{d}u\Big\}.
\end{aligned}$$

Term (I) remains to be a U-statistic with zero expectation, but not necessarily degenerate. Its order can easily be seen to be, with standard variance calculation:

$$\text{(I)} = O_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{n^2h}}\right)\right) = O_{\mathbb{P}}\left(\frac{1}{n} + \frac{1}{n\sqrt{nh}}\right),$$

which has the same order as the leave-in bias, hence can be ignored (provided that $\dot{w}_i$ has finite variance).

For (II), we observe:

$$\begin{aligned}
|\text{(II)}| &\le |\hat{\theta}-\theta_0|^2 \cdot \frac{1}{n^2}\sum_{i,j;i\neq j}\sup_{|\theta-\theta_0|\le\delta}|\ddot{w}(z_i,\theta)| \cdot \Big|\mathbf{r}_p\left(\frac{x_i-x}{h}\right)\Big(\mathbb{1}[x_j \le x_i] - F(x_i)\Big)K_h(x_i - x) \\
&\qquad - \int_{\frac{x_{\mathrm{L}}-x}{h}}^{\frac{x_{\mathrm{U}}-x}{h}} \mathbf{r}_p(u)\Big(\mathbb{1}[x_j \le x+hu] - F(x+hu)\Big)K(u)g(x+hu)\mathrm{d}u\Big| \\
&= O_{\mathbb{P}}\left(\frac{1}{n}\right),
\end{aligned}$$

by direct expectation calculation. ∎

6.9 Proof of Lemma 7

Proof. First note that

$$\begin{aligned}
&\frac{1+o_{\mathbb{P}}(1)}{n}\sum_i \int_{\frac{x_{\mathrm{L}}-x}{h}}^{\frac{x_{\mathrm{U}}-x}{h}} \mathbf{r}_p(u)\Big(\hat{w}_i - w_i\Big)\Big(\mathbb{1}[x_i \le x+hu] - F(x+hu)\Big)K(u)g(x+hu)\mathrm{d}u \\
&= \frac{1+o_{\mathbb{P}}(1)}{n}\left[\sum_i \int_{\frac{x_{\mathrm{L}}-x}{h}}^{\frac{x_{\mathrm{U}}-x}{h}} \mathbf{r}_p(u)\,\dot{w}_i\Big(\mathbb{1}[x_i \le x+hu] - F(x+hu)\Big)K(u)g(x+hu)\mathrm{d}u\right]\Big(\hat{\theta}-\theta_0\Big) && \text{(I)}\\
&+ \frac{1+o_{\mathbb{P}}(1)}{n}\left[\sum_i \int_{\frac{x_{\mathrm{L}}-x}{h}}^{\frac{x_{\mathrm{U}}-x}{h}} \mathbf{r}_p(u)\,(\dot{w}(\mathbf{z}_i,\tilde{\theta}) - \dot{w}_i)\Big(\mathbb{1}[x_i \le x+hu] - F(x+hu)\Big)K(u)g(x+hu)\mathrm{d}u\right]\Big(\hat{\theta}-\theta_0\Big). && \text{(II)}
\end{aligned}$$

We first consider the interior case, with the following expectation calculation for (I):

$$\begin{aligned}
&\mathbb{E}\left[\mathbf{S}_{p,x}^{-1}\frac{1}{n}\sum_i\int_{\frac{x_{\mathtt{L}}-x}{h}}^{\frac{x_{\mathtt{U}}-x}{h}}\mathbf{r}_p\left(u\right)\dot{w}_i\Big(\mathbb{1}[x_i\leq x+hu]-F(x+hu)\Big)K(u)g(x+hu)\mathrm{d}u\right]\\
&=\mathbf{S}_{p,x}^{-1}\int_{\frac{x_{\mathtt{L}}-x}{h}}^{\frac{x_{\mathtt{U}}-x}{h}}\mathbf{r}_p\left(u\right)\Big(I(x+hu)-I(x_{\mathtt{U}})F(x+hu)\Big)K(u)g(x+hu)\mathrm{d}u\\
&=g(x)(I(x)-I(x_{\mathtt{U}})F(x))\mathbf{e}_0+O(h).
\end{aligned}$$

Since the variance of the above quantity has order $1/n$, we have, when x is in interior, that

$$\mathbf{N}_x\mathbf{S}_{p,x}^{-1}\frac{1}{n}\sum_i\int_{\frac{x_{\mathtt{L}}-x}{h}}^{\frac{x_{\mathtt{U}}-x}{h}}\mathbf{r}_p\left(u\right)\dot{w}_i\Big(\mathbb{1}[x_i\leq x+hu]-F(x+hu)\Big)K(u)g(x+hu)\mathrm{d}u=g(x)(I(x)-I(x_{\mathtt{U}})F(x))\mathbf{e}_0+O_{\mathbb{P}}\left(\sqrt{h}+\frac{1}{\sqrt{nh}}\right).$$

When x is in either the lower boundary or the upper boundary region,

$$\mathbf{N}_x\mathbf{S}_{p,x}^{-1}\frac{1}{n}\sum_i\int_{\frac{x_{\mathtt{L}}-x}{h}}^{\frac{x_{\mathtt{U}}-x}{h}}\mathbf{r}_p\left(u\right)\dot{w}_i\Big(\mathbb{1}[x_i\leq x+hu]-F(x+hu)\Big)K(u)\mathrm{d}u=O_{\mathbb{P}}\left(\sqrt{h}+\frac{1}{\sqrt{nh}}\right),$$

since the leading constant term vanishes.

As for (II), it is bounded through the following quantity:

$$\begin{aligned}
\text{(II)}&\asymp_{\mathbb{P}}\left(\hat{\theta}-\theta_0\right)\frac{1}{n}\left[\sum_i\int_{\frac{x_{\mathtt{L}}-x}{h}}^{\frac{x_{\mathtt{U}}-x}{h}}\mathbf{r}_p\left(u\right)\left(\dot{w}(\mathbf{z}_i,\tilde{\theta})-\dot{w}_i\right)\Big(\mathbb{1}[x_i\leq x+hu]-F(x+hu)\Big)K(u)g(x+hu)\mathrm{d}u\right]\\
&\leq|\hat{\theta}-\theta_0|^2\cdot\frac{1}{n}\left[\sum_i\int_{\frac{x_{\mathtt{L}}-x}{h}}^{\frac{x_{\mathtt{U}}-x}{h}}\sup_{|\theta-\theta_0|\leq\delta}|\ddot{w}(\mathbf{z}_i,\theta)|\left|\mathbf{r}_p\left(u\right)\Big(\mathbb{1}[x_i\leq x+hu]-F(x+hu)\Big)K(u)g(x+hu)\right|\mathrm{d}u\right],
\end{aligned}$$

which has the order $O_{\mathbb{P}}(n^{-1})$ by expectation calculation. ■

6.10 Proof of Theorem 3

Proof. This follows from previous lemmas. ■

6.11 Proof of Theorem 4

Proof. The proof resembles that of Theorem 2 with minor changes. ■

6.12 Proof of Lemma 8

Proof. We rely on Lemma 1 and 2 (note that whether the weights are estimated is irrelevant here), hence will not repeat arguments already established there. Instead, extra care will be given to ensure the characterization of higher order bias.

Consider the case where with enough smoothness on G, then the bias is characterized by

$$\begin{aligned}
&h^{-v}v!\mathbf{e}_v'\left[G^{(1)}(x)\mathbf{S}_{p,x}+hG^{(2)}(x)\tilde{\mathbf{S}}_{p,x}+o(h)+O_{\mathbb{P}}(1/\sqrt{nh})\right]^{-1}\\
&\qquad\left[h^{p+1}\frac{F^{(p+1)}(x)}{(p+1)!}G^{(1)}(x)\mathbf{c}_{p,x}+h^{p+2}\left[\frac{F^{(p+2)}(x)}{(p+2)!}G^{(1)}(x)+\frac{F^{(p+1)}(x)}{(p+1)!}G^{(2)}(x)\right]\tilde{\mathbf{c}}_{p,x}+o(h^{p+2})\right]\\
&=h^{-v}v!\mathbf{e}_v'\left[\frac{1}{G^{(1)}(x)}\mathbf{S}_{p,x}^{-1}-h\frac{G^{(2)}(x)}{[G^{(1)}(x)]^2}\mathbf{S}_{p,x}^{-1}\tilde{\mathbf{S}}_{p,x}\mathbf{S}_{p,x}^{-1}+O_{\mathbb{P}}\left(1/\sqrt{nh}\right)\right]\\
&\qquad\left[h^{p+1}\frac{F^{(p+1)}(x)}{(p+1)!}G^{(1)}(x)\mathbf{c}_{p,x}+h^{p+2}\left[\frac{F^{(p+2)}(x)}{(p+2)!}G^{(1)}(x)+\frac{F^{(p+1)}(x)}{(p+1)!}G^{(2)}(x)\right]\tilde{\mathbf{c}}_{p,x}+o(h^{p+2})\right]\{1+o_{\mathbb{P}}(1)\},
\end{aligned}$$

which gives the desired result. Here $\tilde{\mathbf{S}}_{p,x}=\int_{\frac{x_{\mathtt{L}}-x}{h}}^{\frac{x_{\mathtt{U}}-x}{h}}u\mathbf{r}_p(u)\mathbf{r}_p(u)'k(u)\mathrm{d}u$. And for the last line to hold, one needs the extra condition $nh^3\to\infty$ so that $O_{\mathbb{P}}\left(1/\sqrt{nh}\right)=o_{\mathbb{P}}(h)$. See Fan and Gijbels (1996) (Theorem 3.1, pp. 62). ■

6.13 Proof of Lemma 9

Proof. The proof resembles that of Lemma 1, and is omitted here. ■

6.14 Proof of Theorem 5

Proof. The proof splits into two cases. We sketch one of them. Assume either x is boundary or $p - v$ is odd, the MSE-optimal bandwidth is asymptotically equivalent to the following:

$$\frac{\tilde{h}_{\mathtt{MSE},p,v,x}}{h_{\mathtt{MSE},p,v,x}} \to 1, \qquad \tilde{h}_{\mathtt{MSE},p,v,x} = \left(\frac{1}{n} \frac{(2v-1)H^{(1)}(x)\mathbf{e}_v'\mathbf{S}_{p,x}^{-1}\mathbf{\Gamma}_{p,x}\mathbf{S}_{p,x}^{-1}\mathbf{e}_v}{(2p-2v+2)(\frac{F^{(p+1)}(x)}{(p+1)!}\mathbf{e}_v'\mathbf{S}_{p,x}^{-1}\mathbf{c}_{p,x})^2} \right)^{\frac{1}{2p+1}},$$

which is obtained by optimizing MSE ignoring the higher order bias term. With consistency of the preliminary estimates, it can be shown that

$$\hat{h}_{\mathtt{MSE},p,v,x} = \left(\frac{1}{n} \frac{(2v-1)\hat{\sigma}_{p,v,x}^2 n\ell^{2v-1}}{(2p-2v+2)(v!\frac{\hat{F}^{(p+1)}(x)}{(p+1)!}\mathbf{e}_v'\mathbf{S}_{p,x}^{-1}\mathbf{c}_{p,x})^2} \right)^{\frac{1}{2p+1}} \{1 + o_{\mathbb{P}}(1)\}.$$

Apply the consistency assumption of the preliminary estimates again, one can easily show that $\hat{h}_{\mathtt{MSE},p,v,x}$ is consistent both in rate and constant.

A similar argument can be made for the other case, and is omitted here. ■

Table 1. Simulation (truncated Normal). $x = -0.8$, $p = 2$, triangular kernel.

(a) $n = 1000$

	$\hat{f}_p$			**SE**	
	bias	sd	$\sqrt{\text{mse}}$	mean	size
$h_{\texttt{MSE}} \times$					
0.1	0.008	0.170	0.170	0.168	6.00
0.3	0.003	0.097	0.097	0.094	5.98
0.5	0.002	0.074	0.074	0.074	4.92
0.7	0.006	0.062	0.062	0.063	4.76
0.9	0.013	0.054	0.056	0.056	4.40
1	0.017	0.052	0.054	0.053	4.46
1.1	0.021	0.050	0.054	0.051	4.74
1.3	0.031	0.046	0.056	0.047	5.06
1.5	0.044	0.043	0.061	0.044	4.64
1.7	0.058	0.040	0.071	0.041	4.60
1.9	0.074	0.038	0.083	0.039	4.88
$\hat{h}$					
	0.013	0.070	0.071	0.066	8.24

	Quantile				
	0.10	0.25	0.50	0.75	0.90
$\hat{h}/h_{\texttt{MSE}}$					
	0.4	0.475	0.607	0.861	1.277

(b) $n = 2000$

	$\hat{f}_p$			**SE**	
	bias	sd	$\sqrt{\text{mse}}$	mean	size
$h_{\texttt{MSE}} \times$					
0.1	0.003	0.121	0.121	0.122	5.14
0.3	0.002	0.071	0.071	0.070	5.54
0.5	0.002	0.055	0.055	0.055	5.32
0.7	0.004	0.046	0.046	0.047	4.84
0.9	0.009	0.040	0.042	0.041	4.22
1	0.012	0.038	0.040	0.039	4.36
1.1	0.016	0.037	0.040	0.038	4.46
1.3	0.024	0.034	0.042	0.035	4.56
1.5	0.034	0.032	0.046	0.033	4.56
1.7	0.045	0.030	0.054	0.031	4.62
1.9	0.058	0.028	0.065	0.029	4.62
$\hat{h}$					
	0.012	0.054	0.055	0.048	9.50

	Quantile				
	0.10	0.25	0.50	0.75	0.90
$\hat{h}/h_{\texttt{MSE}}$					
	0.407	0.492	0.638	0.906	1.36

Note. (i) bias: empirical bias of the estimators; (ii) sd: empirical standard deviation of the estimators; (iii) $\sqrt{\text{mse}}$: empirical m.s.e. of the estimators; (iv) mean: empirical average of the estimated standard errors; (v) size: empirical size of testing the hypothesis at nominal 5% level, the test statistic is centered at $\mathbb{E}\hat{f}_p$.

Table 2. Simulation (truncated Normal). $x = -0.8$, $p = 3$, triangular kernel.

(a) $n = 1000$

	$\hat{f}_p$			**SE**	
	bias	sd	$\sqrt{\text{mse}}$	mean	size
$h_{\texttt{MSE}}\times$					
0.1	0.015	0.234	0.234	0.229	4.94
0.3	0.005	0.129	0.129	0.127	5.50
0.5	−0.001	0.100	0.100	0.100	4.96
0.7	−0.004	0.085	0.085	0.085	4.58
0.9	−0.004	0.075	0.075	0.076	4.58
1	−0.005	0.071	0.071	0.072	4.50
1.1	−0.006	0.068	0.069	0.069	4.86
1.3	−0.007	0.064	0.064	0.064	5.12
1.5	−0.007	0.059	0.060	0.060	4.92
1.7	−0.005	0.056	0.056	0.057	4.64
1.9	0.000	0.053	0.053	0.054	4.92
$\hat{h}$					
	0.001	0.114	0.114	0.110	5.06

	Quantile				
	0.10	0.25	0.50	0.75	0.90
$\hat{h}/h_{\texttt{MSE}}$					
	0.317	0.344	0.387	0.462	0.59

(b) $n = 2000$

	$\hat{f}_p$			**SE**	
	bias	sd	$\sqrt{\text{mse}}$	mean	size
$h_{\texttt{MSE}}\times$					
0.1	0.006	0.161	0.161	0.161	4.46
0.3	0.003	0.094	0.094	0.093	5.36
0.5	−0.001	0.074	0.074	0.073	5.70
0.7	−0.003	0.062	0.062	0.062	5.24
0.9	−0.004	0.055	0.055	0.056	4.52
1	−0.005	0.052	0.053	0.053	4.54
1.1	−0.006	0.050	0.051	0.051	4.50
1.3	−0.008	0.047	0.047	0.047	4.60
1.5	−0.008	0.043	0.044	0.044	4.44
1.7	−0.007	0.041	0.042	0.042	4.54
1.9	−0.004	0.039	0.039	0.040	4.78
$\hat{h}$					
	0.000	0.081	0.081	0.078	5.58

	Quantile				
	0.10	0.25	0.50	0.75	0.90
$\hat{h}/h_{\texttt{MSE}}$					
	0.332	0.359	0.402	0.483	0.628

Note. (i) bias: empirical bias of the estimators; (ii) sd: empirical standard deviation of the estimators; (iii) $\sqrt{\text{mse}}$: empirical m.s.e. of the estimators; (iv) mean: empirical average of the estimated standard errors; (v) size: empirical size of testing the hypothesis at nominal 5% level, the test statistic is centered at $\mathbb{E}\hat{f}_p$.

Table 3. Simulation (truncated Normal). $x = -0.5$, $p = 2$, triangular kernel.

(a) $n = 1000$

	$\hat{f}_p$			SE	
	bias	sd	$\sqrt{\text{mse}}$	mean	size
$h_{\texttt{MSE}}\times$					
0.1	0.003	0.067	0.067	0.068	5.34
0.3	0.001	0.037	0.037	0.037	5.14
0.5	−0.002	0.028	0.028	0.028	5.26
0.7	−0.003	0.026	0.026	0.026	4.86
0.9	−0.003	0.026	0.026	0.026	4.90
1	−0.002	0.026	0.026	0.026	4.86
1.1	−0.001	0.026	0.026	0.026	4.86
1.3	0.001	0.026	0.026	0.026	4.76
1.5	0.005	0.026	0.026	0.026	4.88
1.7	0.010	0.025	0.027	0.026	5.02
1.9	0.016	0.025	0.030	0.025	4.68
$\hat{h}$					
	0.004	0.031	0.031	0.026	9.78

	Quantile				
	0.10	0.25	0.50	0.75	0.90
$\hat{h}/h_{\texttt{MSE}}$					
	0.716	0.828	1.004	1.299	1.861

(b) $n = 2000$

	$\hat{f}_p$			SE	
	bias	sd	$\sqrt{\text{mse}}$	mean	size
$h_{\texttt{MSE}}\times$					
0.1	0.002	0.052	0.052	0.051	5.46
0.3	0.000	0.028	0.028	0.029	5.28
0.5	−0.002	0.021	0.021	0.021	5.76
0.7	−0.003	0.018	0.019	0.019	5.48
0.9	−0.003	0.018	0.018	0.018	5.46
1	−0.003	0.018	0.018	0.018	5.30
1.1	−0.003	0.018	0.018	0.018	5.28
1.3	−0.001	0.018	0.018	0.018	5.12
1.5	0.001	0.018	0.018	0.018	4.68
1.7	0.005	0.018	0.019	0.018	4.58
1.9	0.009	0.018	0.020	0.018	4.50
$\hat{h}$					
	0.005	0.027	0.027	0.018	12.86

	Quantile				
	0.10	0.25	0.50	0.75	0.90
$\hat{h}/h_{\texttt{MSE}}$					
	0.728	0.843	1.044	1.433	2.223

Note. (i) bias: empirical bias of the estimators; (ii) sd: empirical standard deviation of the estimators; (iii) $\sqrt{\text{mse}}$: empirical m.s.e. of the estimators; (iv) mean: empirical average of the estimated standard errors; (v) size: empirical size of testing the hypothesis at nominal 5% level, the test statistic is centered at $\mathbb{E}\hat{f}_p$.

Table 4. Simulation (truncated Normal). $x = -0.5$, $p = 3$, triangular kernel.

(a) $n = 1000$

	$\hat{f}_p$			**SE**	
	bias	sd	$\sqrt{\text{mse}}$	mean	size
$h_{\texttt{MSE}} \times$					
0.1	0.001	0.061	0.061	0.061	5.08
0.3	0.001	0.036	0.036	0.035	5.46
0.5	0.001	0.027	0.027	0.027	4.94
0.7	0.001	0.025	0.025	0.025	4.90
0.9	0.001	0.025	0.025	0.025	5.00
1	0.002	0.026	0.026	0.026	5.16
1.1	0.002	0.026	0.026	0.026	5.18
1.3	0.005	0.026	0.027	0.026	4.80
1.5	0.010	0.026	0.028	0.026	4.74
1.7	0.017	0.026	0.031	0.026	4.70
1.9	0.025	0.026	0.036	0.026	4.66
$\hat{h}$					
	0.001	0.031	0.031	0.031	4.56

	Quantile				
	0.10	0.25	0.50	0.75	0.90
$\hat{h}/h_{\texttt{MSE}}$					
	0.289	0.314	0.361	0.437	0.587

(b) $n = 2000$

	$\hat{f}_p$			**SE**	
	bias	sd	$\sqrt{\text{mse}}$	mean	size
$h_{\texttt{MSE}} \times$					
0.1	0.001	0.046	0.046	0.045	5.12
0.3	0.001	0.026	0.026	0.026	5.54
0.5	0.001	0.019	0.019	0.020	4.90
0.7	0.001	0.017	0.017	0.018	4.84
0.9	0.001	0.018	0.018	0.018	4.86
1	0.001	0.018	0.018	0.018	4.84
1.1	0.002	0.018	0.018	0.018	4.74
1.3	0.003	0.018	0.019	0.018	4.72
1.5	0.007	0.018	0.020	0.019	4.60
1.7	0.013	0.018	0.022	0.019	4.38
1.9	0.019	0.018	0.027	0.018	4.32
$\hat{h}$					
	0.001	0.022	0.022	0.023	4.74

	Quantile				
	0.10	0.25	0.50	0.75	0.90
$\hat{h}/h_{\texttt{MSE}}$					
	0.3	0.327	0.374	0.457	0.619

Note. (i) bias: empirical bias of the estimators; (ii) sd: empirical standard deviation of the estimators; (iii) $\sqrt{\text{mse}}$: empirical m.s.e. of the estimators; (iv) mean: empirical average of the estimated standard errors; (v) size: empirical size of testing the hypothesis at nominal 5% level, the test statistic is centered at $\mathbb{E}\hat{f}_p$.

Table 5. Simulation (truncated Normal). $x = 0.5$, $p = 2$, triangular kernel.

(a) $n = 1000$

	$\hat{f}_p$			SE	
	bias	sd	$\sqrt{\text{mse}}$	mean	size
$h_{\texttt{MSE}}\times$					
0.1	0.004	0.068	0.068	0.068	5.26
0.3	0.000	0.037	0.037	0.037	5.04
0.5	−0.002	0.027	0.027	0.028	4.84
0.7	−0.004	0.022	0.022	0.022	4.64
0.9	−0.008	0.018	0.020	0.018	4.68
1	−0.010	0.017	0.020	0.017	4.92
1.1	−0.012	0.016	0.020	0.016	4.92
1.3	−0.016	0.014	0.021	0.014	5.20
1.5	−0.021	0.012	0.024	0.012	5.72
1.7	−0.026	0.011	0.028	0.011	6.04
1.9	−0.031	0.010	0.032	0.009	6.44
$\hat{h}$					
	−0.009	0.023	0.025	0.017	19.00

	Quantile				
	0.10	0.25	0.50	0.75	0.90
$\hat{h}/h_{\texttt{MSE}}$					
	0.748	0.829	0.971	1.256	1.785

(b) $n = 2000$

	$\hat{f}_p$			SE	
	bias	sd	$\sqrt{\text{mse}}$	mean	size
$h_{\texttt{MSE}}\times$					
0.1	0.002	0.051	0.051	0.051	5.10
0.3	0.000	0.028	0.028	0.028	4.56
0.5	−0.001	0.021	0.021	0.021	4.62
0.7	−0.003	0.017	0.017	0.017	4.86
0.9	−0.006	0.014	0.016	0.014	4.90
1	−0.007	0.013	0.015	0.013	4.98
1.1	−0.009	0.013	0.015	0.012	5.08
1.3	−0.013	0.011	0.017	0.011	5.00
1.5	−0.016	0.010	0.019	0.010	5.16
1.7	−0.020	0.009	0.022	0.009	5.24
1.9	−0.025	0.008	0.026	0.008	5.46
$\hat{h}$					
	−0.007	0.018	0.020	0.013	17.38

	Quantile				
	0.10	0.25	0.50	0.75	0.90
$\hat{h}/h_{\texttt{MSE}}$					
	0.772	0.849	0.976	1.214	1.703

Note. (i) bias: empirical bias of the estimators; (ii) sd: empirical standard deviation of the estimators; (iii) $\sqrt{\text{mse}}$: empirical m.s.e. of the estimators; (iv) mean: empirical average of the estimated standard errors; (v) size: empirical size of testing the hypothesis at nominal 5% level, the test statistic is centered at $\mathbb{E}\hat{f}_p$.

Table 6. Simulation (truncated Normal). $x = 0.5$, $p = 3$, triangular kernel.

(a) $n = 1000$

	$\hat{f}_p$			**SE**	
	bias	sd	$\sqrt{\text{mse}}$	mean	size
$h_{\texttt{MSE}}\times$					
0.1	0.001	0.061	0.061	0.061	5.04
0.3	0.001	0.033	0.033	0.034	4.88
0.5	0.001	0.024	0.024	0.025	4.68
0.7	0.001	0.019	0.019	0.020	4.70
0.9	0.000	0.017	0.017	0.017	5.06
1	−0.002	0.016	0.016	0.016	5.18
1.1	−0.004	0.015	0.016	0.015	5.30
1.3	−0.009	0.014	0.017	0.014	5.64
1.5	−0.014	0.013	0.019	0.013	5.60
1.7	−0.018	0.012	0.022	0.012	5.80
1.9	−0.021	0.012	0.024	0.011	6.10
$\hat{h}$					
	−0.001	0.020	0.020	0.018	7.92

	Quantile				
	0.10	0.25	0.50	0.75	0.90
$\hat{h}/h_{\texttt{MSE}}$					
	0.641	0.697	0.784	0.926	1.156

(b) $n = 2000$

	$\hat{f}_p$			**SE**	
	bias	sd	$\sqrt{\text{mse}}$	mean	size
$h_{\texttt{MSE}}\times$					
0.1	0.001	0.045	0.045	0.045	4.88
0.3	0.000	0.024	0.024	0.025	4.54
0.5	0.001	0.018	0.018	0.018	4.72
0.7	0.001	0.015	0.015	0.015	5.18
0.9	0.001	0.012	0.012	0.012	5.10
1	−0.001	0.012	0.012	0.012	5.06
1.1	−0.002	0.011	0.011	0.011	5.04
1.3	−0.007	0.010	0.012	0.010	5.10
1.5	−0.011	0.010	0.015	0.009	5.32
1.7	−0.015	0.009	0.018	0.009	5.42
1.9	−0.019	0.009	0.021	0.008	5.88
$\hat{h}$					
	−0.001	0.015	0.015	0.013	8.98

	Quantile				
	0.10	0.25	0.50	0.75	0.90
$\hat{h}/h_{\texttt{MSE}}$					
	0.664	0.721	0.81	0.96	1.176

Note. (i) bias: empirical bias of the estimators; (ii) sd: empirical standard deviation of the estimators; (iii) $\sqrt{\text{mse}}$: empirical m.s.e. of the estimators; (iv) mean: empirical average of the estimated standard errors; (v) size: empirical size of testing the hypothesis at nominal 5% level, the test statistic is centered at $\mathbb{E}\hat{f}_p$.

Table 7. Simulation (truncated Normal). $x = 1.5$, $p = 2$, triangular kernel.

(a) $n = 1000$

	$\hat{f}_p$			SE	
	bias	sd	$\sqrt{\text{mse}}$	mean	size
$h_{\texttt{MSE}}\times$					
0.1	0.005	0.042	0.042	0.043	6.48
0.3	0.002	0.024	0.024	0.024	5.12
0.5	0.003	0.018	0.018	0.018	4.68
0.7	0.004	0.015	0.016	0.015	4.24
0.9	0.005	0.013	0.014	0.013	4.36
1	0.006	0.012	0.014	0.013	4.22
1.1	0.007	0.012	0.014	0.012	4.20
1.3	0.009	0.011	0.014	0.011	4.26
1.5	0.011	0.010	0.015	0.011	4.16
1.7	0.013	0.010	0.016	0.010	3.96
1.9	0.014	0.010	0.017	0.010	4.02
$\hat{h}$					
	0.006	0.016	0.017	0.012	12.76

	Quantile				
	0.10	0.25	0.50	0.75	0.90
$\hat{h}/h_{\texttt{MSE}}$					
	0.726	0.837	1.035	1.388	2.029

(b) $n = 2000$

	$\hat{f}_p$			SE	
	bias	sd	$\sqrt{\text{mse}}$	mean	size
$h_{\texttt{MSE}}\times$					
0.1	0.002	0.032	0.032	0.032	5.74
0.3	0.001	0.018	0.018	0.018	5.08
0.5	0.002	0.014	0.014	0.014	5.18
0.7	0.003	0.012	0.012	0.012	5.04
0.9	0.004	0.010	0.011	0.010	5.08
1	0.005	0.010	0.011	0.010	4.90
1.1	0.006	0.009	0.011	0.009	4.76
1.3	0.007	0.008	0.011	0.008	4.68
1.5	0.009	0.008	0.012	0.008	4.60
1.7	0.011	0.007	0.013	0.008	4.54
1.9	0.012	0.007	0.014	0.007	4.44
$\hat{h}$					
	0.005	0.013	0.014	0.009	14.70

	Quantile				
	0.10	0.25	0.50	0.75	0.90
$\hat{h}/h_{\texttt{MSE}}$					
	0.758	0.863	1.042	1.381	1.96

Note. (i) bias: empirical bias of the estimators; (ii) sd: empirical standard deviation of the estimators; (iii) $\sqrt{\text{mse}}$: empirical m.s.e. of the estimators; (iv) mean: empirical average of the estimated standard errors; (v) size: empirical size of testing the hypothesis at nominal 5% level, the test statistic is centered at $\mathbb{E}\hat{f}_p$.

Table 8. Simulation (truncated Normal). $x = 1.5$, $p = 3$, triangular kernel.

(a) $n = 1000$

	$\hat{f}_p$			SE	
	bias	sd	$\sqrt{\text{mse}}$	mean	size
$h_{\texttt{MSE}}\times$					
0.1	0.002	0.045	0.045	0.045	6.12
0.3	0.001	0.026	0.026	0.026	5.34
0.5	0.001	0.020	0.020	0.020	4.90
0.7	0.002	0.017	0.017	0.017	4.34
0.9	0.003	0.015	0.016	0.016	4.38
1	0.004	0.015	0.015	0.015	4.12
1.1	0.006	0.014	0.016	0.015	4.30
1.3	0.010	0.013	0.017	0.014	4.38
1.5	0.017	0.013	0.021	0.013	4.36
1.7	0.024	0.012	0.027	0.012	4.34
1.9	0.032	0.011	0.034	0.012	4.70
$\hat{h}$					
	0.006	0.018	0.019	0.015	10.70

	Quantile				
	0.10	0.25	0.50	0.75	0.90
$\hat{h}/h_{\texttt{MSE}}$					
	0.79	0.874	0.993	1.18	1.487

(b) $n = 2000$

	$\hat{f}_p$			SE	
	bias	sd	$\sqrt{\text{mse}}$	mean	size
$h_{\texttt{MSE}}\times$					
0.1	0.001	0.033	0.033	0.033	5.82
0.3	0.000	0.019	0.019	0.019	5.08
0.5	0.000	0.015	0.015	0.015	5.10
0.7	0.001	0.013	0.013	0.013	5.10
0.9	0.002	0.011	0.011	0.011	4.80
1	0.003	0.011	0.011	0.011	4.54
1.1	0.004	0.010	0.011	0.011	4.60
1.3	0.008	0.010	0.013	0.010	4.82
1.5	0.013	0.009	0.016	0.009	4.86
1.7	0.019	0.009	0.021	0.009	4.78
1.9	0.026	0.008	0.028	0.009	4.38
$\hat{h}$					
	0.005	0.014	0.015	0.011	12.12

	Quantile				
	0.10	0.25	0.50	0.75	0.90
$\hat{h}/h_{\texttt{MSE}}$					
	0.827	0.91	1.033	1.235	1.533

Note. (i) bias: empirical bias of the estimators; (ii) sd: empirical standard deviation of the estimators; (iii) $\sqrt{\text{mse}}$: empirical m.s.e. of the estimators; (iv) mean: empirical average of the estimated standard errors; (v) size: empirical size of testing the hypothesis at nominal 5% level, the test statistic is centered at $\mathbb{E}\hat{f}_p$.

Table 9. Simulation (Exponential). $x = 0$, $p = 2$, triangular kernel.

(a) $n = 1000$

	$\hat{f}_p$			SE	
	bias	sd	$\sqrt{\text{mse}}$	mean	size
$h_{\mathtt{MSE}}\times$					
0.1	0.005	0.259	0.259	0.254	5.90
0.3	−0.002	0.142	0.142	0.140	5.46
0.5	−0.009	0.105	0.106	0.104	5.16
0.7	−0.020	0.085	0.087	0.085	4.74
0.9	−0.032	0.072	0.079	0.072	5.04
1	−0.038	0.068	0.077	0.067	5.08
1.1	−0.044	0.063	0.077	0.063	5.58
1.3	−0.058	0.057	0.081	0.056	5.62
1.5	−0.072	0.051	0.088	0.050	5.98
1.7	−0.086	0.046	0.098	0.046	6.14
1.9	−0.100	0.043	0.109	0.042	5.64
$\hat{h}$					
	−0.033	0.094	0.100	0.089	7.74

	Quantile				
	0.10	0.25	0.50	0.75	0.90
$\hat{h}/h_{\mathtt{MSE}}$					
	0.749	0.836	0.99	1.274	1.94

(b) $n = 2000$

	$\hat{f}_p$			SE	
	bias	sd	$\sqrt{\text{mse}}$	mean	size
$h_{\mathtt{MSE}}\times$					
0.1	0.000	0.187	0.187	0.185	5.36
0.3	0.000	0.103	0.103	0.103	5.54
0.5	−0.007	0.078	0.078	0.077	5.08
0.7	−0.017	0.063	0.065	0.063	4.90
0.9	−0.028	0.054	0.061	0.053	4.92
1	−0.034	0.051	0.061	0.050	4.88
1.1	−0.039	0.048	0.062	0.047	4.90
1.3	−0.052	0.042	0.067	0.041	5.12
1.5	−0.065	0.038	0.075	0.037	5.36
1.7	−0.078	0.035	0.085	0.034	5.86
1.9	−0.091	0.032	0.097	0.031	5.88
$\hat{h}$					
	−0.031	0.073	0.079	0.064	8.70

	Quantile				
	0.10	0.25	0.50	0.75	0.90
$\hat{h}/h_{\mathtt{MSE}}$					
	0.815	0.916	1.089	1.438	2.155

Note. (i) bias: empirical bias of the estimators; (ii) sd: empirical standard deviation of the estimators; (iii) $\sqrt{\text{mse}}$: empirical m.s.e. of the estimators; (iv) mean: empirical average of the estimated standard errors; (v) size: empirical size of testing the hypothesis at nominal 5% level, the test statistic is centered at $\mathbb{E}\hat{f}_p$.

Table 10. Simulation (Exponential). $x = 0$, $p = 3$, triangular kernel.

(a) $n = 1000$

	$\hat{f}_p$			SE	
	bias	sd	$\sqrt{\text{mse}}$	mean	size
$h_{\texttt{MSE}}\times$					
0.1	0.008	0.332	0.332	0.326	5.30
0.3	0.003	0.182	0.182	0.179	5.24
0.5	0.002	0.135	0.135	0.134	5.26
0.7	−0.003	0.108	0.108	0.109	4.52
0.9	−0.007	0.092	0.093	0.093	4.58
1	−0.009	0.086	0.087	0.087	5.02
1.1	−0.012	0.081	0.082	0.081	4.90
1.3	−0.018	0.073	0.075	0.073	5.60
1.5	−0.024	0.067	0.071	0.066	5.58
1.7	−0.031	0.061	0.068	0.060	6.16
1.9	−0.039	0.056	0.068	0.055	5.76
$\hat{h}$					
	0.000	0.135	0.135	0.128	5.96

	Quantile				
	0.10	0.25	0.50	0.75	0.90
$\hat{h}/h_{\texttt{MSE}}$					
	0.47	0.511	0.574	0.678	0.877

(b) $n = 2000$

	$\hat{f}_p$			SE	
	bias	sd	$\sqrt{\text{mse}}$	mean	size
$h_{\texttt{MSE}}\times$					
0.1	0.000	0.239	0.239	0.237	5.52
0.3	0.004	0.132	0.132	0.132	5.26
0.5	0.003	0.099	0.099	0.098	4.84
0.7	−0.001	0.080	0.080	0.080	4.78
0.9	−0.005	0.069	0.069	0.069	5.02
1	−0.007	0.065	0.065	0.064	4.80
1.1	−0.010	0.061	0.062	0.060	4.88
1.3	−0.015	0.055	0.057	0.054	5.06
1.5	−0.021	0.050	0.054	0.049	5.42
1.7	−0.027	0.046	0.053	0.045	5.50
1.9	−0.034	0.042	0.054	0.041	5.60
$\hat{h}$					
	0.000	0.097	0.097	0.092	5.78

	Quantile				
	0.10	0.25	0.50	0.75	0.90
$\hat{h}/h_{\texttt{MSE}}$					
	0.49	0.531	0.594	0.705	0.9

Note. (i) bias: empirical bias of the estimators; (ii) sd: empirical standard deviation of the estimators; (iii) $\sqrt{\text{mse}}$: empirical m.s.e. of the estimators; (iv) mean: empirical average of the estimated standard errors; (v) size: empirical size of testing the hypothesis at nominal 5% level, the test statistic is centered at $\mathbb{E}\hat{f}_p$.

Table 11. Simulation (Exponential). $x = 1$, $p = 2$, triangular kernel.

(a) $n = 1000$

	$\hat{f}_p$			SE	
	bias	sd	$\sqrt{\text{mse}}$	mean	size
$h_{\texttt{MSE}}\times$					
0.1	0.006	0.065	0.065	0.065	5.88
0.3	0.003	0.036	0.036	0.036	5.30
0.5	0.004	0.027	0.027	0.027	5.32
0.7	0.006	0.022	0.023	0.022	5.22
0.9	0.009	0.019	0.021	0.018	5.00
1	0.011	0.017	0.020	0.017	5.10
1.1	0.013	0.016	0.020	0.016	4.90
1.3	0.017	0.014	0.022	0.014	4.74
1.5	0.023	0.012	0.026	0.012	4.66
1.7	0.028	0.011	0.030	0.011	4.42
1.9	0.033	0.010	0.034	0.010	4.10
$\hat{h}$					
	0.008	0.021	0.022	0.017	11.84

	Quantile				
	0.10	0.25	0.50	0.75	0.90
$\hat{h}/h_{\texttt{MSE}}$					
	0.783	0.846	0.934	1.065	1.269

(b) $n = 2000$

	$\hat{f}_p$			SE	
	bias	sd	$\sqrt{\text{mse}}$	mean	size
$h_{\texttt{MSE}}\times$					
0.1	0.004	0.049	0.049	0.049	5.26
0.3	0.002	0.027	0.028	0.027	5.16
0.5	0.003	0.021	0.021	0.021	5.34
0.7	0.005	0.017	0.018	0.017	5.36
0.9	0.007	0.014	0.016	0.014	5.20
1	0.008	0.013	0.016	0.013	5.32
1.1	0.010	0.013	0.016	0.013	5.28
1.3	0.013	0.011	0.017	0.011	5.02
1.5	0.017	0.010	0.020	0.010	4.92
1.7	0.022	0.009	0.024	0.009	4.76
1.9	0.027	0.008	0.028	0.008	4.50
$\hat{h}$					
	0.007	0.015	0.017	0.014	9.64

	Quantile				
	0.10	0.25	0.50	0.75	0.90
$\hat{h}/h_{\texttt{MSE}}$					
	0.813	0.869	0.943	1.043	1.173

Note. (i) bias: empirical bias of the estimators; (ii) sd: empirical standard deviation of the estimators; (iii) $\sqrt{\text{mse}}$: empirical m.s.e. of the estimators; (iv) mean: empirical average of the estimated standard errors; (v) size: empirical size of testing the hypothesis at nominal 5% level, the test statistic is centered at $\mathbb{E}\hat{f}_p$.

Table 12. Simulation (Exponential). $x = 1$, $p = 3$, triangular kernel.

(a) $n = 1000$

	$\hat{f}_p$			**SE**	
	bias	sd	$\sqrt{\text{mse}}$	mean	size
$h_{\texttt{MSE}}\times$					
0.1	0.003	0.064	0.064	0.064	5.72
0.3	0.001	0.036	0.036	0.036	5.04
0.5	0.000	0.027	0.027	0.027	5.04
0.7	−0.001	0.023	0.023	0.022	5.08
0.9	−0.001	0.020	0.020	0.020	5.52
1	−0.001	0.019	0.019	0.019	5.44
1.1	0.000	0.018	0.018	0.018	5.32
1.3	0.002	0.017	0.017	0.016	4.96
1.5	0.005	0.015	0.015	0.015	4.76
1.7	0.008	0.013	0.015	0.013	4.78
1.9	0.011	0.012	0.016	0.012	4.66
$\hat{h}$					
	0.000	0.022	0.022	0.020	7.46

	Quantile				
	0.10	0.25	0.50	0.75	0.90
$\hat{h}/h_{\texttt{MSE}}$					
	0.708	0.765	0.862	1.021	1.271

(b) $n = 2000$

	$\hat{f}_p$			**SE**	
	bias	sd	$\sqrt{\text{mse}}$	mean	size
$h_{\texttt{MSE}}\times$					
0.1	0.002	0.047	0.047	0.047	5.62
0.3	0.001	0.026	0.026	0.026	5.20
0.5	0.000	0.020	0.020	0.020	5.38
0.7	−0.001	0.017	0.017	0.016	5.16
0.9	−0.002	0.015	0.015	0.015	5.18
1	−0.002	0.014	0.014	0.014	5.20
1.1	−0.001	0.014	0.014	0.013	5.16
1.3	0.001	0.012	0.012	0.012	5.16
1.5	0.003	0.011	0.012	0.011	5.28
1.7	0.006	0.010	0.011	0.010	5.24
1.9	0.008	0.009	0.012	0.009	5.16
$\hat{h}$					
	0.000	0.016	0.016	0.014	7.30

	Quantile				
	0.10	0.25	0.50	0.75	0.90
$\hat{h}/h_{\texttt{MSE}}$					
	0.76	0.823	0.927	1.096	1.35

Note. (i) bias: empirical bias of the estimators; (ii) sd: empirical standard deviation of the estimators; (iii) $\sqrt{\text{mse}}$: empirical m.s.e. of the estimators; (iv) mean: empirical average of the estimated standard errors; (v) size: empirical size of testing the hypothesis at nominal 5% level, the test statistic is centered at $\mathbb{E}\hat{f}_p$.

Table 13. Simulation (Exponential). $x = 1.5$, $p = 2$, triangular kernel.

(a) $n = 1000$

	$\hat{f}_p$			SE	
	bias	sd	$\sqrt{\text{mse}}$	mean	size
$h_{\texttt{MSE}} \times$					
0.1	0.003	0.048	0.049	0.048	6.04
0.3	0.002	0.027	0.027	0.027	5.30
0.5	0.002	0.020	0.020	0.020	4.76
0.7	0.004	0.017	0.017	0.017	4.34
0.9	0.006	0.014	0.016	0.015	4.30
1	0.008	0.013	0.015	0.014	4.38
1.1	0.009	0.013	0.016	0.013	4.52
1.3	0.013	0.011	0.017	0.012	4.46
1.5	0.016	0.010	0.019	0.011	4.38
1.7	0.021	0.009	0.023	0.010	4.02
1.9	0.025	0.009	0.026	0.009	4.12
$\hat{h}$					
	0.006	0.016	0.017	0.014	8.74

	Quantile				
	0.10	0.25	0.50	0.75	0.90
$\hat{h}/h_{\texttt{MSE}}$					
	0.803	0.863	0.947	1.065	1.21

(b) $n = 2000$

	$\hat{f}_p$			SE	
	bias	sd	$\sqrt{\text{mse}}$	mean	size
$h_{\texttt{MSE}} \times$					
0.1	0.001	0.037	0.037	0.036	6.12
0.3	0.001	0.021	0.021	0.020	5.96
0.5	0.002	0.016	0.016	0.016	5.62
0.7	0.003	0.013	0.013	0.013	5.14
0.9	0.005	0.011	0.012	0.011	4.92
1	0.006	0.011	0.012	0.011	5.04
1.1	0.007	0.010	0.012	0.010	5.00
1.3	0.010	0.009	0.013	0.009	4.94
1.5	0.013	0.008	0.015	0.008	4.90
1.7	0.016	0.008	0.018	0.008	4.54
1.9	0.020	0.007	0.021	0.007	4.60
$\hat{h}$					
	0.005	0.012	0.013	0.011	8.24

	Quantile				
	0.10	0.25	0.50	0.75	0.90
$\hat{h}/h_{\texttt{MSE}}$					
	0.83	0.88	0.952	1.046	1.157

Note. (i) bias: empirical bias of the estimators; (ii) sd: empirical standard deviation of the estimators; (iii) $\sqrt{\text{mse}}$: empirical m.s.e. of the estimators; (iv) mean: empirical average of the estimated standard errors; (v) size: empirical size of testing the hypothesis at nominal 5% level, the test statistic is centered at $\mathbb{E}\hat{f}_p$.

Table 14. Simulation (Exponential). $x = 1.5$, $p = 3$, triangular kernel.

(a) $n = 1000$

	$\hat{f}_p$			SE	
	bias	sd	$\sqrt{\text{mse}}$	mean	size
$h_{\texttt{MSE}}\times$					
0.1	0.000	0.049	0.049	0.048	6.14
0.3	0.000	0.027	0.027	0.027	4.88
0.5	0.000	0.021	0.021	0.021	4.34
0.7	−0.001	0.018	0.018	0.018	4.58
0.9	−0.003	0.016	0.016	0.016	4.62
1	−0.004	0.015	0.016	0.015	4.76
1.1	−0.006	0.015	0.016	0.015	4.68
1.3	−0.007	0.014	0.016	0.014	4.78
1.5	−0.006	0.014	0.015	0.014	4.78
1.7	−0.004	0.013	0.014	0.013	4.86
1.9	−0.002	0.012	0.012	0.012	4.92
$\hat{h}$					
	−0.003	0.015	0.016	0.014	7.32

	Quantile				
	0.10	0.25	0.50	0.75	0.90
$\hat{h}/h_{\texttt{MSE}}$					
	0.962	1.035	1.17	1.393	1.761

(b) $n = 2000$

	$\hat{f}_p$			SE	
	bias	sd	$\sqrt{\text{mse}}$	mean	size
$h_{\texttt{MSE}}\times$					
0.1	0.000	0.036	0.036	0.035	6.02
0.3	0.000	0.020	0.020	0.020	6.04
0.5	0.000	0.016	0.016	0.015	5.42
0.7	−0.001	0.013	0.013	0.013	5.08
0.9	−0.002	0.012	0.012	0.012	5.14
1	−0.003	0.011	0.012	0.011	5.22
1.1	−0.005	0.011	0.012	0.011	4.96
1.3	−0.007	0.011	0.013	0.010	5.38
1.5	−0.007	0.010	0.012	0.010	5.24
1.7	−0.006	0.010	0.011	0.010	5.22
1.9	−0.004	0.009	0.010	0.009	5.16
$\hat{h}$					
	−0.003	0.011	0.012	0.010	8.32

	Quantile				
	0.10	0.25	0.50	0.75	0.90
$\hat{h}/h_{\texttt{MSE}}$					
	1.015	1.091	1.225	1.454	1.91

Note. (i) bias: empirical bias of the estimators; (ii) sd: empirical standard deviation of the estimators; (iii) $\sqrt{\text{mse}}$: empirical m.s.e. of the estimators; (iv) mean: empirical average of the estimated standard errors; (v) size: empirical size of testing the hypothesis at nominal 5% level, the test statistic is centered at $\mathbb{E}\hat{f}_p$.