Estimates of derivatives of (log) densities and related objects

Joris Pinkse and Karl Schurter[*]

this version: April 2021

Abstract

We estimate the density and its derivatives using a local polynomial approximation to the logarithm of an unknown density function f. The estimator is guaranteed to be nonnegative and achieves the same optimal rate of convergence in the interior as on the boundary of the support of f. The estimator is therefore well–suited to applications in which nonnegative density estimates are required, such as in semiparametric maximum likelihood estimation. In addition, we show that our estimator compares favorably with other kernel–based methods, both in terms of asymptotic performance and computational ease. Simulation results confirm that our method can perform similarly or better in finite samples compared to these alternative methods when they are used with optimal inputs, i.e. an Epanechnikov kernel and optimally chosen bandwidth sequence. We provide code in several languages.

[*] Corresponding author kschurter@psu.edu

1 Introduction

We propose a new nonparametric estimator for (the logarithm) of a density function and its derivatives that attains the optimal rate of convergence both in the interior and at the boundary of the support. Our density estimator is available in closed form and is guaranteed to be positive unlike several alternatives, which is appealing in some applications and critical in others, such as in semiparametric maximum likelihood estimation (see e.g. Klein and Spady, 1993).[1] The new methodology differs from the previous literature in that it first estimates a function's derivatives, which, if desirable, can then be used to construct an estimate of the function itself. Our general estimation strategy can also be applied to obtain estimates of other quantities of economic interest, including the density in regression discontinuity design models, the (reciprocal) of the propensity score, the inverse bid function in auction models, and any other application in which the density appears inside a logarithm or a denominator.[2]

Specifically, we consider an i.i.d. sequence of random variables $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$ with $\boldsymbol{x}_i$ distributed according to some unknown distribution F with density $f > 0$ on its support $[0, \mathscr{U})$, where $\mathscr{U}$ can be infinite. The standard Rosenblatt–Parzen (RP) kernel density estimator is inconsistent at the boundary and is typically badly biased in finite samples at values of x near the boundary. In contrast, our method employs a local polynomial approximation of $L(x) = \log f(x)$ to obtain asymptotically normal estimates of L and its derivatives, away from, at, or near the boundary.

An advantage of using a polynomial approximation to the log–density instead of the density is that the estimated density can be guaranteed to be positive, which is not true for alternative boundary correction methods that use boundary kernels or a local polynomial approximation of f (Cheng et al., 1997; Zhang and Karunamuni, 1998) of F (Lejeune and Sarda, 1992; Cattaneo et al., 2019).[3] Unlike Loader (1996) and Hjort and Jones (1996), however, the computation of our estimator does not require solving a nonlinear system of equations that involves numerical integration. In fact, the estimator of derivatives of L may be expressed as the solution to a linear system of (weighted) local averages. Thus, our method can be characterized as a local method of moments, similar in spirit to local likelihood density estimation (Loader, 1996; Hjort and Jones, 1996, and much subsequent work) but computationally more similar to local polynomial regression (Lejeune and Sarda, 1992; Cheng et al., 1997; Zhang and Karunamuni, 1998; Cattaneo et al., 2019).[4] We therefore retain the computational ease of a local polynomial regression estimator while eliminating the possibility of

[1] Klein and Spady square their density estimates to ensure positivity.

[2] The object of interest here is nonparametric in nature, i.e. it does not necessarily get averaged out as it does in e.g. Lewbel and Schennach (2007).

[3] An exception is Jones and Foster (1996).

[4] For the convenience of the reader appendix B provides an overview of the alternative approaches with which we will compare our estimator in simulations.

negative density estimates. The estimator of f then obtains in explicit form with estimates of the derivatives of L as inputs.

Apart from its numerical advantages, our estimator for the density has the same first order asymptotic properties when applied with the same bandwidth as the local likelihood estimator. When applied with a larger bandwidth, however, our estimator achieves a smaller asymptotic mean square error. We cannot generally compare the bias of our method with the biases of methods that use a polynomial approximation to f, but our estimator has the same asymptotic variance as traditional methods when they are applied using an optimal kernel and bandwidth sequence. Hence, our local polynomial approximation to L can be expected to outperform alternative estimators for f in finite samples when our bias is smaller, e.g. when the log–density is in fact polynomial.

In large enough samples, an asymptotically unbiased version of our estimator achieves a smaller variance and therefore a smaller mean square error than the optimized alternatives. Importantly, our estimator realizes this improved performance without sacrificing nonnegativity of the estimated density, as would be required to achieve the same asymptotic distribution using alternative methods.[5]

Our estimator requires the choice of several input parameters, more than does RP. This can be seen as a downside and we could have hardwired specific choices of some of the input parameters to reduce the number of input parameters that must be chosen. We have not done so since there can be instances in which the additional flexibility is of value.

We also note that the log–density or its derivatives may be of direct interest to the researcher, in which case our method may be an attractive alternative to transforming estimates of f and its derivatives to obtain the desired estimates. For instance, the generalized reflection method of Karunamuni and Alberts (2005) and Karunamuni and Zhang (2008) requires an estimate of $L'(0)$ which they obtain using a finite difference approximation. Estimates of L can moreover be used as an input into other objects. One case that has already been mentioned is semiparametric maximum likelihood estimation, of which Klein and Spady (1993) is a classical example in which the likelihood objective can be written as a function of the log–density. But there are other important examples. For instance, in regression discontinuity design, estimation of the density at the discontinuity point can be of interest (see e.g. Cattaneo et al., 2019). A second example would be the estimation of propensity scores which are of importance in the estimation of treatment effects. A final example is that of the estimation of auction models for which a version of our estimator can be used to obtain direct estimates of the inverse strategy function; see e.g. Hickman and Hubbard (2015); Pinkse and Schurter

[5] One could replace negative density estimates produced by alternative methods by zero, but that is both clunky and would not help in cases in which the density must be positive.

(2019). These examples and more are discussed in section 5. Finally, we provide code in several languages at `https://github.com/kschurter/logdensity`.

2 Estimator

We now discuss our main estimator, postponing the discussion of applications and variants to section 5. Let $L(y) = \log f(y)$ denote the log–density function and assume it to possess at least $S + 1 \geq 2$ derivatives at x, the point at which we wish to estimate L. Our estimator will be in the kernel family of estimators and we denote our bandwidth by h.[6] To specifically allow for x approximating the boundary, we introduce the notation $z = \min(x/h, 1)$.

In the first step of our estimation procedure, we estimate derivatives of L, which are subsequently used to construct an estimate of L itself. Our estimator of derivatives of L is based on the fact that for any differentiable function $g : \mathbb{R} \to \mathbb{R}^S$ with support $[-z, 1]$ and for which $g(-z) = g(1) = 0 \in \mathbb{R}^S$, we have

$$\begin{aligned}\frac{1}{h^2}\int_{x-zh}^{x+h} g'\Big(\frac{y-x}{h}\Big) f(y)\,\mathrm{d}y &= -\frac{1}{h}\int_{x-zh}^{x+h} g\Big(\frac{y-x}{h}\Big) L'(y) f(y)\,\mathrm{d}y \\ &= -\frac{1}{h}\int_{x-zh}^{x+h} g\Big(\frac{y-x}{h}\Big)\sum_{s=1}^{S+1}\frac{L^{(s)}(x)(y-x)^{s-1}}{(s-1)!} f(y)\,\mathrm{d}y + o\big(h^S\big), \quad (1)\end{aligned}$$

where $L^{(s)}$ denotes the s–th derivative. The above follows from integration by parts under the assumption that f is bounded on the domain of integration and a Taylor expansion of L' around x. We define $\beta_s = L^{(s)}(x)$ and gather these coefficients into a vector $\beta = [\beta_1 \dots \beta_S]^\intercal$. We then estimate integrals on the right and left sides of (1) by their sample analogs and estimate β by solving

$$\frac{1}{nh^2}\sum_{i=1}^{n} g'\Big(\frac{\boldsymbol{x}_i - x}{h}\Big) = -\sum_{s=1}^{S}\hat{\boldsymbol{\beta}}_s \frac{1}{nh}\sum_{i=1}^{n} g\Big(\frac{\boldsymbol{x}_i - x}{h}\Big)\frac{(\boldsymbol{x}_i - x)^{s-1}}{(s-1)!}\,. \quad (2)$$

Since (2) is linear in $\hat{\boldsymbol{\beta}}$, the solution will generally be unique.[7] Moreover, we can choose g such that our estimator $\hat{\boldsymbol{\beta}}$ of β is in closed form.

There are many functions g that satisfy the desiderata outlined above. For the purpose of providing

[6] By kernel family of estimators, we mean that the estimator is based on local weighted averages, though our weighting function is not a kernel. We adopt the standard definition of a kernel as a function k that integrates to one.

[7] Recall that g is an S-dimensional vector, which can be chosen to ensure that its Jacobian is invertible. Our results will show that $\hat{\boldsymbol{\beta}}$ is unique with probability approaching one. For intuition, consider what happens if $S = 1$. Our discussion in this section assumes that L is continuously differentiable at x, which implies that $f(x) > 0$. Write the right hand side in (2) as $\hat{\boldsymbol{\beta}}_1\hat{\Sigma}_n$. Then by the Hoeffding inequality, $\Pr(|\hat{\Sigma}_n - \mathbb{E}\hat{\Sigma}_n| > \epsilon)$ goes for any fixed $\epsilon > 0$ to zero at an exponential rate and $\mathbb{E}\hat{\Sigma}_n$ converges by a standard nonparametric kernel bias expansion to a nonzero limit. The case $f(x) = 0$ will be discussed in section 3.3.

examples, we choose g to be a vector whose j–th element is

$$g_j(t) = (t+z)^j(1-t)\mathbb{1}(-z \le t \le 1).$$

In section 4 we describe the sense in which this choice of g may in fact be optimal.

Example 1. *If $S = 1$ then $\hat{\boldsymbol{\beta}}_1$ is simply the derivative of the logarithm of the RP estimator with kernel $6\mathbb{1}(-z \le t \le 1)\{z + t(1-z) - t^2\}/(1+z)^3$, which simplifies to an Epanechnikov kernel if $z = 1$.* □

Example 2. *If $z = 1$ and $g_j(u) = k(u)u^{j-1}$ for some symmetric, nonnegative kernel function k then (1) represents the first–order condition for the minimizer of the least squares criterion in a local polynomial regression of $L'(x_i)$ on x_i, which would be infeasible because $L'(x_i)$ is not observed.*

Example 2 illustrates that the integration by parts in (1) can be viewed as a device for obtaining a feasible set of local moment conditions from an infeasible set of moment conditions involving L'. Thus, g fulfills a role similar to a kernel, but the restrictions we impose are different. Indeed, we require $g(-z) = g(1) = 0$ so that $g(1)f(x+h) - g(-z)f(x-zh)$ is zero after integration by parts.

Now, once we have an estimator $\hat{\boldsymbol{\beta}}$ of the derivatives of L at x, we can use it to construct estimators of $f(x)$ and $L(x)$. Indeed, substituting our approximation for the log–density in $\int_{-z}^{1} m_z(t)f(x+th)\,\mathrm{d}t$ and rearranging suggests an estimator for β_0 similar to that of Loader (1996):

$$\hat{\boldsymbol{f}}(x) = \frac{\hat{\boldsymbol{f}}_m(x)}{\displaystyle\int_{-z}^{1} m_z(t)\exp\Big(\sum_{s=1}^{S}\frac{\hat{\boldsymbol{\beta}}_s t^s h^s}{s!}\Big)\,\mathrm{d}t}, \tag{3}$$

where $\hat{\boldsymbol{f}}_m$ is a RP estimator using a nonnegative kernel m_z with support $[-z, 1]$. It should be apparent that $\hat{\boldsymbol{f}}(x)$ cannot be negative and the numerator is zero only if there are no data on the interval $[x - zh, x + h]$.

If $z < 1$ then m_z can be thought of as a traditional boundary kernel[8] such that the bias in the numerator of (3) is $O(h)$. The role of the denominator is that it is an (asymptotically) biased estimator of the number one. Indeed, the denominator bias compensates for the numerator bias such that for $S = 1$, the bias of $\hat{\boldsymbol{f}}(x)$ is again $O(h^2)$. We define $\hat{\boldsymbol{L}}(x) = \log \hat{\boldsymbol{f}}(x)$ and show that its asymptotic bias is $\beta_{S+1}h^{S+1}$ times a constant that is independent of both n and fixed x.[9]

Computing $\hat{\boldsymbol{f}}$ is relatively simple because $\hat{\boldsymbol{\beta}}$ is simply a local least squares statistic and $\hat{\boldsymbol{f}}$ is a ratio. Although $\hat{\boldsymbol{f}}$'s denominator contains an integral, for values of $S \le 2$, which will be the most common scenario,

[8] $k/\int_{\max(x/h-1,0)} k$; see e.g. Gasser and Müller (1979).

[9] For the case $x = zh$, the asymptotic bias does depend on z.

the denominator in (3) obtains in closed form if m_z is a truncated Epanechnikov; for $S = 1$ this is demonstrated in example 3 below.[10] For other kernels and greater values of S, an asymptotically equivalent closed form expression can be obtained by expanding the denominator in (3) in terms of exponential Bell polynomials (Bell, 1927).[11] Standard numerical integration methods can also be applied to this integral, in which case an advantage of our method is that this integral only needs to be computed once rather than at each iterate of a maximization routine, as in a local maximum likelihood approach.

The following examples compare the asymptotic behavior of our estimator and traditional approaches to kernel density estimation at and away from the boundary. Example 3 obtains a closed form expression for the denominator in (3), which is used in examples 4 and 5 to obtain explicit expressions for the special cases $z = 1$ and $z = 0$ (away from the boundary and at the boundary, respectively).

Example 3. *Let $v = v(z) = 3/(2 + 3z - z^3) = 3/\{(1 + z)^2(2 - z)\}$ and suppose that m_z is a truncated Epanechnikov, i.e. $m_z(t) = v(1 - t^2)\mathbb{1}(-z \le t \le 1)$. If $S = 1$ then the denominator in (3) is $\chi_1(\hat{\boldsymbol{\beta}}_1 h)$, where*

$$\chi_1(t) = \begin{cases} 1, & t = 0, \\ vt^{-3}\left[(2t-2)e^t - \left\{-2 - 2tz + t^2(1 - z^2)\right\}e^{-zt}\right], & t \neq 0. \end{cases}$$

□

Example 4. *Suppose $S = 1$. If $x \ge h$ then $g(t) = -(1 - t^2)\mathbb{1}(|t| \le 1)$, which is proportional to minus the Epanechnikov kernel. So then $\hat{\boldsymbol{\beta}}_1$ is simply the derivative of the logarithm of the RP estimator using the Epanechnikov kernel. The denominator in (3) is then exactly*

$$\begin{cases} \dfrac{3}{2\hat{\boldsymbol{\beta}}_1^3 h^3}\left\{\exp(\hat{\boldsymbol{\beta}}_1 h)(\hat{\boldsymbol{\beta}}_1 h - 1) + \exp(-\hat{\boldsymbol{\beta}}_1 h)(\hat{\boldsymbol{\beta}}_1 h + 1)\right\}, & \hat{\boldsymbol{\beta}}_1 \neq 0, \\ 1, & \hat{\boldsymbol{\beta}}_1 = 0. \end{cases}$$

The denominator can be expanded around $h = 0$ to obtain the approximation $\hat{\mathbf{L}}(x) \approx \log \hat{\boldsymbol{f}}_m(x) - \log\big(1 + \hat{\boldsymbol{\beta}}_1^2 h^2/10\big)$. It is well–known that the bias of $\log \hat{\boldsymbol{f}}_m(x)$ is $h^2 f''(x) \,/\, 5f(x) + o(h^2)$. The bias of $\hat{\mathbf{L}}(x)$ is by the mean value theorem then seen to be $h^2\beta_2/5 + o(h^2)$. Thus, the bias we introduced in the denominator offsets the bias present in the numerator. □

In example 4 we took $x > h$ and hence $z = 1$ to provide intuition. In the following example we consider what happens at the boundary, i.e. if $x = z = 0$.

[10] For $S = 2$ and $\hat{\boldsymbol{\beta}}_1 < 0$ it would however involve a normal distribution function and for $\hat{\boldsymbol{\beta}}_1 > 0$ a similar integral.

[11] For $S = 1$ we get that that the exponential in the denominator in (3) can be written as $1 + \hat{\boldsymbol{\beta}}_1 th + \hat{\boldsymbol{\beta}}_1^2 t^2 h^2 + o_p(h^2)$. For $S = 2$ we get $1 + \hat{\boldsymbol{\beta}}_1 th + (2\hat{\boldsymbol{\beta}}_2 + \hat{\boldsymbol{\beta}}_1^2)t^2h^2 + (3\hat{\boldsymbol{\beta}}_1\hat{\boldsymbol{\beta}}_2 + \hat{\boldsymbol{\beta}}_1^3)t^3h^3 + o_p(h^3)$.

Example 5. *Again suppose that $S = 1$, but now let $x = z = 0$. Then $g(t) = -(1-t)t$, such that $\hat{\boldsymbol{\beta}}_1$ is now the derivative of the kernel density estimator at zero using the kernel $6(1-t)t\mathbb{1}(t \le 1)$.*

If we again use an Epanechnikov in (3) then now the denominator becomes for $\hat{\boldsymbol{\beta}}_1 \neq 0$,

$$\frac{3}{2}\frac{2 + \hat{\boldsymbol{\beta}}_1^2 h^2 - \exp(\hat{\boldsymbol{\beta}}_1 h)(2 - 2\hat{\boldsymbol{\beta}}_1 h)}{\hat{\boldsymbol{\beta}}_1^3 h^3} = 1 + \frac{3\hat{\boldsymbol{\beta}}_1 h}{8} + \frac{\hat{\boldsymbol{\beta}}_1^2 h^2}{10} + o(h^2).$$

Our results show that the bias of $\hat{\boldsymbol{\beta}}_1$ is $\beta_2 h/2 + o(h)$, such that the denominator bias is $3\beta_1 h/8 + 3\beta_2 h^2/16 + \beta_1^2 h^2/10 + o(h^2)$. Further,

$$\mathbb{E}\hat{\boldsymbol{f}}_m(0) = f(0) + \frac{3hf'(0)}{8} + \frac{f''(0)h^2}{10} + o(h^2),$$

such that the bias of $\hat{\boldsymbol{f}}(0)$ is now $-7h^2\beta_2 f(x)/80 + o(h^2)$. The bias of $\hat{\mathbf{L}}(0)$ is by the delta method hence $-7h^2\beta_2/80 + o(h^2)$. Again, the bias we introduced in the denominator offsets the bias in the numerator. □

Example 5 demonstrates that, unlike traditional boundary kernel estimators, the bias of our estimator is $O(h^2)$ at the boundary, also. It may seem odd that the bias in example 5 is less than that in example 4 but note that the variance will be larger at the boundary and one would hence generally choose a greater bandwidth.

We finish this section with a description of a multivariate version of our estimator. For $S = 1$, one could simply solve for $\hat{\boldsymbol{\beta}}_1^{\rightarrow}$ in

$$\frac{1}{nh^{d_x+1}}\sum_{i=1}^{n} \partial_x g^{\rightarrow}\left(\frac{\boldsymbol{x}_i - x}{h}\right) = -\hat{\boldsymbol{\beta}}_1^{\rightarrow}\frac{1}{nh^{d_x}}\sum_{i=1}^{n} g^{\rightarrow}\left(\frac{\boldsymbol{x}_i - x}{h}\right),$$

with d_x the dimension of $\boldsymbol{x}_i$ and $g^{\rightarrow}(x) = \prod_{j=1}^{d_x} g_{(j)}(x_j)$, where the $g_{(j)}$'s satisfy the conditions described above for the univariate case.[12] The corresponding estimator for the density f is then

$$\hat{\boldsymbol{f}}(x) = \frac{\hat{\boldsymbol{f}}_m(x)}{\int_{-z_1}^{1} \cdots \int_{z_{d_x}}^{1} m_z(t)\exp\left(ht^{\intercal}\hat{\boldsymbol{\beta}}_1^{\rightarrow}\right) \mathrm{d}t_{d_x} \cdots \mathrm{d}t_1}.$$

The procedure for values of $S > 1$ is analogous to the univariate case but is substantially messier since the number of parameters to be estimated grows as a power of d_x, just like it would for local polynomial estimation. The theoretical results below are for the univariate case, but extend to the multivariate case.

[12] Away from any boundaries, the $g_{(j)}$'s can all be the same.

3 Limit results

3.1 Derivatives of L

We first derive limit results for the vector $\hat{\boldsymbol{\beta}}$ of estimates of derivatives of L. Since our estimator $\hat{\boldsymbol{\beta}}$ is defined as the inverse of a matrix times a vector, its bias is the inverse of a matrix times a vector, also.

To simplify expressions for the asymptotic bias and variance of our estimator, we introduce the following objects which depend only on the choice of function g (which in turn depends on the proximity to the boundary z), as well as a diagonal matrix that depends on h. Let

$$\Omega_{js} = \Omega_{js}(z) = \frac{1}{(s-1)!}\int_{-z}^{1} -g_j(t)t^{s-1}\,\mathrm{d}t = \sum_{t=0}^{s-1}\frac{(-1)^{s-t+1}(s-t)j!}{(j+s+1-t)!t!}(1+z)^{j+s+1-t}, \qquad j,s = 1,2,\dots, \tag{4}$$

and let Ω, b, Λ be defined as

$$\Omega = \begin{bmatrix} \Omega_{11} & \Omega_{12} & \cdots & \Omega_{1S} \\ \Omega_{21} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \Omega_{S-1,S} \\ \Omega_{S1} & \cdots & \Omega_{S,S-1} & \Omega_{SS} \end{bmatrix}, \qquad b = \Omega^{-1}\begin{bmatrix} \Omega_{1,S+1} \\ \Omega_{2,S+1} \\ \vdots \\ \Omega_{S,S+1} \end{bmatrix}, \qquad \Lambda = \begin{bmatrix} 1 & & \\ & \ddots & \\ & & h^{S-1} \end{bmatrix}.$$

Let further $V \in \mathbb{R}^{S\times S}$ have (j,s) element equal to

$$\int_{-z}^{1} g_j'(t)g_s'(t)\,\mathrm{d}t = \frac{2js(1+z)^{j+s+1}}{(j+s+1)(j+s)(j+s-1)}.$$

We are now in a position to state our first theorem.

Theorem 1. *Assume L is $S+1$ times continuously differentiable (and hence f nonzero) in a neighborhood of x. Let $h \to 0$ and $nh^3 \to \infty$ as $n \to \infty$. For a vector $\tilde{\beta}$ defined in appendix A.1,*

$$\Lambda(\tilde{\beta} - \beta) = h^S\beta_{S+1}b + o_p(h^S), \qquad \sqrt{nh^3}\Lambda(\hat{\boldsymbol{\beta}} - \tilde{\beta}) \xrightarrow{d} N\Big(0,\ (\Omega^{\intercal}V^{-1}\Omega)^{-1} \Big/ f(x)\Big). \qquad \square$$

The "in a neighborhood" condition comes from the fact that we specifically allow $x = zh$. Although x is allowed to vary with h and hence with n, our result is not really uniform since, like with other local nonparametric estimators, tightness obtains only in an h–neighborhood, i.e. a neighborhood that shrinks with the sample size. The rates in theorem 1 are the standard rates for local nonparametric derivative estimation.[13]

[13] With the RP estimator the $\sqrt{nh^3}$ rate comes from the fact that the variance of $h^{-2}k'\{(x-\boldsymbol{x}_i)/h\}$ is $O(h^{-3})$ since squaring produces h^{-4} and then one h is recovered by substitution.

Example 6. *For $S = 1$ the bias and variance expressions of $\hat{\boldsymbol{\beta}}_1$ simplify to $\beta_2 h(1-z)/2$ and $12/\{f(x)(1+z)^3\}$, respectively. The interior case ($z = 1$) is more favorable than the boundary case ($z = 0$), as expected.* □

Example 7. *For $S = 2$ the bias and variance expressions for $\hat{\boldsymbol{\beta}}_1$ are $-\beta_3 h^2(1 - 3z + z^2)/10$ and $48(4 - 7z + 4z^2)/\{f(x)(1 + z)^5\}$, which is again more favorable in the interior than at the boundary.* □

3.2 Density

We now continue with the results for $\hat{\boldsymbol{f}}(x)$.

Let $c_{msz} = \int_{-z}^{1} m_z(t)t^s \, \mathrm{d}t \;/\; s!$ and let c_{mz} be a vector with elements $c_{m1z}, \dots, c_{mSz}$. Let further $\omega_z(t) = m_z(t) - c_{mz}^{\intercal}\Omega^{-1}g'(t)$, and define $\mathscr{V} = f(x)\int_{-z}^{1} \omega_z^2(t) \, \mathrm{d}t$ and $\mathscr{B} = f(x)\beta_{S+1}\Xi_f \int_{-z}^{1} \omega_z(t)t^{S+1} \, \mathrm{d}t \;/\; (S+1)! = f(x)\beta_{S+1}\Xi_f\left(c_{m,S+1,z} - c_{mz}^{\intercal}b\right)$, for Ξ_f a constant defined in the statement of theorem 2. Because m_z integrates to one and $\int_{-z}^{1} g'(t) \, \mathrm{d}t = 0$ and $\int_{-z}^{1} \Omega^{-1}g'(t)t^s \, \mathrm{d}t \;/\; s! = \Omega^{-1}\Omega_{\cdot s}$ is the s–th standard basis vector in $\mathbb{R}^S$, the function ω_z is a kernel of order $S + 1$ or higher. To be clear, ω_z is not used to compute the density estimate; rather, it is a convenient object that arises in the asymptotic theory.

Theorem 2. *Assume L is $S + 1$ times continuously differentiable in a neighborhood of x, that $f(x) > 0$, and that $0 < \Xi_f^2 = \lim_{n\to\infty} nh^{2S+3} < \infty$. Then $\sqrt{nh}\{\hat{\boldsymbol{f}}(x) - f(x)\} \overset{d}{\to} N(\mathscr{B}, \mathscr{V})$.* □

The asymptotic bias of our estimator is zero in some instances. For example, if $S = 2$ and $z = 1$ then the asymptotic bias is zero whenever m is a symmetric kernel function; this is natural since this is effectively equivalent to choosing a higher order kernel ω_z, albeit that unlike higher order kernel density estimates, our estimates cannot be negative.

The following two examples derive the ω_z functions for the case in which both m_z is a uniform and x is at the boundary and the case in which m_z is a truncated Epanechnikov and x is anywhere.

Example 8. *Suppose that m_z is a uniform and $x = z = 0$. If $S = 1$ then $\omega_0(t) = (4 - 6t)\mathbb{1}(0 \le t \le 1)$ and $\omega_1(t) = \mathbb{1}(|t| \le 1)/2$. If instead $S = 2$ then $\omega_0(t) = (9 - 36t + 30t^2)\mathbb{1}(0 \le t \le 1)$ and $\omega_1(t) = 3(3 - 5t^2)\mathbb{1}(|t| \le 1)/8$.* □

Example 9. *If m_z is a truncated Epanechnikov and $S = 1$ then $m_z(t) = \nu(1 - t^2)\mathbb{1}(-z \le t \le 1)$ with $\nu = 3/\{(1 + z)^2(2 - z)\}$, such that $c_{m1z} = \nu(1 - z^2)^2/4 = 3(1 - z)^2/\{4(2 - z)\}$, $c_{m2z} = (2 - 4z + 6z^2 - 3z^3)/\{10(2 - z)\}$, and $b = (1 - z)/2$, which produces*

$$\mathscr{B} = \beta_2\Xi f(x)\frac{3z^3 + 29z - 7 - 21z^2}{40(2 - z)},$$

where the ratio equals $1/10$ *for* $z = 1$ *and* $-7/80$ *for* $z = 0$*. To get* $\mathscr{V}_f$ *note that* $\omega_z(t) = \nu\{2(1+z)(1-t^2) - 6(1-z)^2 t + 3(1-z)^3\} \;/\; \{2(1+z)\}$*, which produces*

$$\mathscr{V} = f(x)\frac{108 - 180(1+z) + 120(1+z)^2 - 36(1+z)^3 + 4.05(1+z)^4}{(1+z)^3(2-z)^2},$$

which equals $0.6f(x)$ *for* $z = 1$ *and* $4.01f(x)$ *for* $z = 0$*.* □

As the above two examples demonstrate, deriving the asymptotic bias and variance for generic z can be a messy but straightforward exercise.

3.3 Limit results when $f(x) = 0$

Because the nonnegativity of the proposed density estimator is one of its advantages over some alternative approaches (boundary kernels and local polynomial approximations to the density function), we further investigate the asymptotic behavior of our estimator in cases where one would expect these alternatives to struggle, i.e. when the density is zero. Indeed, nonnegativity would not be an advantage of the density estimator if it performed poorly in other ways whenever the density was close to zero. Fortunately, the bias of the density estimator converges at the usual rate when $f(x) = 0$ for fixed $x > 0$, and the variance converges even faster. Moreover, $\hat{\boldsymbol{f}}(x)\hat{\boldsymbol{\beta}}_1$ converges to $f'(x)$ faster when $f(x) = 0$ than when $f(x) > 0$ if $x > 0$ despite the fact that the log–density and its derivatives are not defined when the density is zero. We further suggest a modification to $\hat{\boldsymbol{\beta}}$ that increases the rate at which $\hat{\boldsymbol{\beta}}_1$ diverges when $f(0) = 0$ and thereby reduces the order of the bias in the density at the boundary. If $f(0) = 0$ then the modified density estimator at the boundary is superconsistent.

Throughout this analysis, we maintain the standard smoothness assumption that f has two continuous derivatives and the assumption that f takes support on an interval $[0, \mathscr{U})$. For fixed $x > 0$, these assumptions imply that $f(x)$ is a relative minimum of the density, i.e. $f(x) = 0$ implies $f'(x) = 0 \leq f''(x)$. At the boundary, however, we admit the possibility that $f'(0) > 0$. Because the dominant terms in the asymptotic expansions depend on whether either or both $f'(x)$ and Ω_{12} are zero, we provide separate results to deal with these cases. The next theorem applies to the interior case where $f(x) = f'(x) = 0$ and $\Omega_{12} = 0$, which will be true if, for example, g is an even function. For simplicity, we specialize the theorems to the local–linear version of the estimator ($S = 1$).

Theorem 3. *Let* $S = 1$*. Assume* f *is twice continuously differentiable at a fixed* $x > 0$*,* $f(x) = 0$*,*

$0 < \Xi_f^2 = \lim_{n\to\infty} nh^5 < \infty$, $\Omega_{12} = 0$, *and* $\Omega_{13} \neq 0$. *Then*

$$n^{2/5}\{\hat{\boldsymbol{f}}(x) - f(x)\} \xrightarrow{p} f''(x)\,\Xi_f c_{m12} \qquad \textit{and} \qquad n^{1/5}\{\hat{\boldsymbol{f}}(x)\hat{\boldsymbol{\beta}}_1 - f'(x)\} \xrightarrow{p} 0. \qquad \square$$

Roughly speaking, the numerator in the definition of $\hat{\boldsymbol{f}}$ converges in probability because the dominant term in the variance expansion is of smaller order when $f(x) = 0$. The denominator of $\hat{\boldsymbol{f}}$ converges in probability to one because $h\hat{\boldsymbol{\beta}}_1$ is $o_p(1)$. Hence, the density estimator is asymptotically equivalent to an RP estimator when $f(x) = f'(x) = 0$ and $\Omega_{12} = 0$.

The fact that $\hat{\boldsymbol{f}}(x)$ in theorem 3 has asymptotic variance equal to zero suggests choosing a smaller bandwidth. Indeed, it can be shown that the optimal bandwidth is then of order $n^{-1/3}$ producing the convergence rate $n^{-2/3}$ and a nonnormal limit distribution.[14] One cannot, however, make the bandwidth dependent on the value of $f(x)$ itself since $f(x)$ is exactly what we wish to estimate. Further, a two step appproach, i.e. replacing $\hat{\boldsymbol{f}}(x)$ with zero if $\hat{\boldsymbol{f}}(x)$ is below some cutoff, would converge even faster if $f(x) = 0$, but would produce discontinuous estimates and erratic behavior near the cutoff.[15]

The fact that the derivative estimator is superconsistent may be surprising at first glance, but it is a property that is shared by the RP derivative estimator. To see this, note that if $f(x) = 0$ for x in the interior then $\mathbb{E}\hat{\boldsymbol{f}}'_{\text{RP}}(x) = o(h^2)$ and $n\mathbb{V}\hat{\boldsymbol{f}}'_{\text{RP}}(x) = o(h^{-3})$, where both results follow by bias–like expansions for nonparametric kernel density estimators, noting that twice continuous differentiability at x in the interior and $f(x) = 0$ imply that $f'(x) = 0$. The convergence rate of both our estimator of f' and the RP derivative estimator is $o_p(n^{-1/5})$, but nothing more can be said.[16]

We now turn to the boundary estimation problem, where we focus on the case $x = 0$ noting that these results easily extend to $x = zh$ for fixed $0 \leq z \leq 1$, as explained in the proof. The following theorem applies to the case where Ω_{12} or $f'(0)$ is nonzero.

Theorem 4. *Let* $S = 1$. *Assume* f *is twice continuously differentiable at zero,* $f(0) = 0$, *and* $h = O(n^{-1/5})$. *Let* $\hat{\boldsymbol{\beta}}_1^* = \hat{\boldsymbol{\beta}}_1 \max\{1, h^q A\hat{\boldsymbol{\beta}}_1\}$ *for some constants* $A > 0$ *and* $0 < q < 1$ *and let* $\hat{\boldsymbol{f}}^*$ *denote the density estimator with* $\hat{\boldsymbol{\beta}}_1$ *replaced by* $\hat{\boldsymbol{\beta}}_1^*$. *If* $f'(0) > 0$ *or* $\Omega_{12} \neq 0$, *then* $n^{2/5}\{\hat{\boldsymbol{f}}^*(0) - f(0)\} \xrightarrow{p} 0$. □

Intuitively, the key idea is that it will become apparent that the density is zero because $\hat{\boldsymbol{\beta}}_1$ diverges at the rate of $1/h$ if $\Omega_{12} \neq 0$ or $1/h^2$ if $f'(0) > 0$ and $\Omega_{12} = 0$. Whenever $\hat{\boldsymbol{\beta}}_1$ is large, $\hat{\boldsymbol{\beta}}_1^*$ gets an extra shove toward infinity so that $h\hat{\boldsymbol{\beta}}_1^*$ diverges under the assumptions of theorem 4 but $\hat{\boldsymbol{\beta}}_1^*$ converges to the same limit as in

[14] Both properties also apply to the RP estimator. Note that $0 \leq \hat{\boldsymbol{f}}(x) = \hat{\boldsymbol{f}}(x) - f(x)$.

[15] And the estimates certainly would not be uniform in f.

[16] For RP, twice continuous differentiability implies that $\mathbb{E}\hat{\boldsymbol{f}}'_{\text{RP}}(x) \simeq f'(x) + hf''(x)\int k(t)t\,\mathrm{d}t = o(h)$ if k is an even kernel.

theorem 1 if $f(0) > 0$. Because $h\hat{\boldsymbol{\beta}}_1^*$ appears in the exponent in the denominator of (3), $\hat{\boldsymbol{f}}^*$ converges to zero very quickly when $f(0) = 0$.

We note that the unmodified estimator $\hat{\boldsymbol{f}}(0)$ works well whenever $\Omega_{12} = 0$ because $h\hat{\boldsymbol{\beta}}_1$ then diverges at the rate of $1/h$. Moreover, practitioners have control over Ω_{12} through the choice of g, so replacing $\hat{\boldsymbol{\beta}}_1$ with $\hat{\boldsymbol{\beta}}^*$ is unnecessary in theorem 4 for suitable choices of g. On the other hand, practitioners may not want to allow the possibility that the density is zero at the boundary to drive this choice. Thus, substituting $\hat{\boldsymbol{\beta}}_1^*$ provides flexibility to consider other criteria in selecting g while maintaining superconsistency when $f(0) = 0$.

For the purposes of estimating f' near the boundary when Ω_{12} is nonzero, we note that the left side of (2) equals $-\Omega_{11} f'(x) - h\Omega_{12} f''(x) + o_p(1/\sqrt{nh^2})$. Dividing the left side by $-\Omega_{11}$ yields a standard RP estimator for $f'(0)$ that converges at the optimal rate whenever $\hat{\boldsymbol{\beta}}_1$ diverges. One can therefore use a similar idea as in theorem 4 to define a piecewise estimator for $f'(zh)$ that equals $\hat{\boldsymbol{f}}\hat{\boldsymbol{\beta}}_1$ if $h^q\hat{\boldsymbol{\beta}}_1 < A$ and equals the RP estimator, otherwise.

4 Asymptotic comparisons

In this section, we explore the optimal (g, m_z) in the local linear case $S = 1$ and compare our optimized estimator with existing methods. We show that the above choice of g and m_z achieves the same asymptotic variance as an optimal RP estimator in the interior ($z = 1$), although their respective biases cannot be compared in general, in the sense that which bias is less depends on the density function and its second derivative at the point of estimation. We then consider the optimal choice of g and m_z at the boundary ($z = 0$), where we show that the truncated Epanechnikov m_z paired with $g(t) = (t + z)(t - 1)^2$ attains the same variance as an optimal boundary kernel (Zhang and Karunamuni, 1998), though the biases are again incomparable because our estimator's bias is proportional to $f(x)L''(x)$ rather than $f''(x)$ as in the case of RP estimators. We are, however, able to derive the relative asymptotic mean squared error (AMSE) of our estimation method compared with a local–likelihood based estimator. We show that our method with the cubic choice $g(t) = (t + z)(1 - t)^2$ attains the same AMSE in the interior and is more efficient at the boundary than the estimator in Loader (1996) with an Epanechnikov kernel.

4.1 Optimal choice of g and m_z in the local linear case

In traditional RP kernel estimation, one typically derives the AMSE–optimal kernel in conjunction with the optimal choice of bandwidth sequence. When the bandwidth sequence is optimally scaled by a multiplicative factor that depends on the roughness and second moment of the kernel, one can show that the Epanechnikov kernel minimizes the AMSE of the estimator for the density among all nonnegative second–order kernels.[17] The nonnegativity restriction ensures that the estimated density is nonnegative and continuous at the point of evaluation. In this section, we similarly explore the optimal choice of inputs to our estimator under an analogous set of restrictions on g and m_z that guarantee that our density estimate $\hat{f}(x)$ is nonnegative.

In addition, we impose the constraint that g does not change sign, i.e. is either nonnegative or nonpositive on $[-z, 1]$. This restriction helps to keep the right side of (2) away from zero whenever the density is nonzero. Although this restriction is not necessary for the asymptotic analysis in section 3 because the probability limit of the right side is (strictly) positive, the estimator may perform poorly when $\int_{-z}^{1} g(t) f(x+th)\,\mathrm{d}t$ is close to zero in finite samples and $f(x) > 0$.

This restriction is necessary in order to derive an interior solution to the input selection problem. Without the sign–change constraint, the optimal choice of g and m_z would yield an asymptotically unbiased estimator, which in turn implies that there is no optimal bandwidth sequence of order $n^{-1/5}$. For example, if $m_z(t) = 3(1-t^2)(1+t)$ and $g(t) = (t+z)(1-t)a_z(t)$, where $a_z(t)$ is a carefully selected quadratic polynomial in t, then $\omega_z(t)$ will be a third–order kernel for $z \in [0, 1)$ and a fourth–order kernel for $z = 1$. One could then choose a bandwidth sequence proportional to $n^{-1/5}$ that is large enough to make the asymptotic variance as small as desired. This approach would be analogous to using a higher–order kernel with a bandwidth of order $n^{-1/5}$ in RP estimation, though the resultant ω_z is in fact less rough than the usual optimal fourth–order kernel because we do not require $\omega_z(1) = \omega_1(-1) = 0$. This asymptotically unbiased estimator can be desirable in some circumstances, but one must be cognizant of the potential risks of using a relatively large bandwidth and a g for which the right side of (2) can be zero in finite samples.

To summarize, we seek to minimize the AMSE subject to the following constraints:

$$\begin{array}{lll} \int_{-z}^{1} g(t)\,\mathrm{d}t = 1, & g(-z) = g(1) = 0, & g \geq 0 \\ \int_{-z}^{1} m_z(t)\,\mathrm{d}t = 1, & m_z(1) = m_1(-1) = 0, & \text{and } m_z \geq 0. \end{array}$$

Note that the constraints that g is nonnegative and integrates to one are without loss of generality because the estimator is invariant under scalar multiplication of g and we have already assumed g does not change sign.

The AMSE of the density estimator depends on g and m_z only through the function ω_z. Provided that the bias is of order $n^{-2/5}$, the optimal bandwidth at a fixed x is $h = \int_{-z}^{1} \omega_z(t)^2 \mathrm{d}t \Big/ \left\{ \int_{-z}^{1} (\omega_z(t) t^2 \mathrm{d}t)^2 f(x) L''(x)^2 n^{1/5} \right\}$ for $z \in \{0, 1\}$ corresponding to $x = 0$ or $x > 0$. The AMSE then factors into

$$\left(\int_{-z}^{1} \omega_z(t) t^2 \,\mathrm{d}t \right)^2 \left(\int_{-z}^{1} \omega_z^2(t)\,\mathrm{d}t \right)^4 \left(\frac{5}{4} f^{6/5}(x) L''(x)^{2/5} n^{-4/5} \right), \tag{5}$$

[17] Specifically, the optimal bandwidth is proportional to $[\int k(t)^2\,\mathrm{d}t]^{1/5}[\int k(t)t^2\,\mathrm{d}t]^{-2/5}$, where k is a nonnegative, symmetric kernel.

which is similar to the corresponding expression for RP estimators, except that the bias depends on $L''(x)$ instead of $f''(x)$. A standard argument using the calculus of variations shows that the Euler equation for optimality implies that ω_z is a quadratic polynomial. Unlike typical derivations of the optimal kernel for RP estimation, however, we do not directly require $\omega_z(1) = 0$, $\omega_1(-1) = 0$, or $\omega_1 \geq 0$. Our problem is slightly more complicated because we must determine the optimal quadratic ω_z that can be obtained from a pair (g, m_z) that satisfy the above constraints.

To this end, we note that the above constraints on (g, m_z) imply that $\omega_z(1) \leq 0$ and $\omega_1(-1) \geq 0$. In the interior, the nonnegativity of $\omega_1(-1)$ is binding. Hence, we find that the optimal ω_1 is the Epanechnikov kernel, which can be obtained by selected the Epanechnikov kernel for m_1. Because c_{m11} is zero, the choice of g does not affect the asymptotic distribution given m_1. One can therefore choose g according to another objective. For instance, $g(t) = 1 - t^2$ minimizes the AMSE of the estimator for β_1.

On the other hand, at the boundary ($z = 0$) the constraint $\omega_0(1) \leq 0$ is binding. We find that the optimal ω_0 is given by $\omega_0(t) = 6 - 18t + 12t^2$, which is the same optimal boundary kernel found in Zhang and Karunamuni (1998). One can produce this ω_0 using a truncated Epanechnikov kernel for m_0 and $g(t) = t(t-1)^2$. Because this g differs from the optimal g for estimating f and f' at $z = 1$, we suggest $g(t) = (t + z)(t - 1)(1 + zt - t)$ combined with the Epanechnikov m_z because it is AMSE–optimal for f and f' in the interior and AMSE–optimal for f at the boundary.

If one attempts to derive optimal inputs for estimating the density local to the boundary, i.e. at a drifting sequence $x = zh$ for $z \in (0, 1)$, complications arise quickly. For example at $z = 1/3$, the constraints on (g, m_z) do not bind and an asymptotically unbiased estimator of the density is feasible using m_z proportional to $(1 - t^2)(1 + t)$ and $g(t) = (t + z)(1 - t)(2t^2 - t - 7)$. The preceding analysis immediately goes awry because there is no optimal bandwidth of order $n^{-1/5}$ for $z = 1/3$ given these inputs. At other values of $z \in (0, 1)$ for which the constraints do bind, the issue is that the above bandwidth sequence is generally suboptimal because ω_z depends on h through z, with the result that the first–order condition for optimality of the bandwidth sequence is insufficient for the global minimum. In light of these issues, we conclude that there is no continuous transformation that one can apply to the bandwidth sequences at $z = 0$ and $z = 1$ in order to optimally estimate the density at intermediate values of z. Therefore, one must necessarily choose a somewhat ad hoc bandwidth local to the boundary.

One such rule for specifying the bandwidth near the boundary is to linearly vary the bandwidth between the interior and the boundary: $h = h_0(1 - \min\{x/h_1, 1\}) + h_1 \min\{x/h_1, 1\}$, where h_0 and h_1 are the optimal bandwidths at $z = 0$ and $z = 1$. When g and m_z are also chosen optimally for estimating the density at $z = 1$ and $z = 0$, $h_0 = 2h_1$ and this rule implies that the same window is used to estimate the density and its derivatives for all $x < h_1$.

4.2 Relative AMSE

The AMSE of the local–likelihood estimator for $f(x)$ in Loader (1996) can be written in a form similar to (5). The relative AMSE of our proposed estimator and the local likelihood estimator using an optimal kernel is then given by the ratio of the multiplicative constants that scale $f(x)^{6/5}L''(x)^{2/5}n^{4/5}$. At $z = 1$, if one compares our optimal estimator to the local–likelihood based estimator using an optimal bandwidth sequence and the Epanechnikov kernel, the relative AMSE of our estimators is one. In fact, the limiting distributions are identical. At $z = 0$, the optimal kernel to use with the local likelihood estimator is triangular, i.e. $k(t) = (1 - |t|)\mathbb{1}\{|t| \geq 1\}$, which yields the same asymptotic bias and variance as our estimator using the truncated Epanechnikov and $g(t) = (t + z)(1 - t)^2$.

We should expect our estimator with $g(t) = (t + z)(1 - t)^2$ and the local–likelihood estimator to perform similarly in finite samples. Thus, our density estimator's computational ease is its more salient advantage over the local–likelihood based approach, given these inputs. Of course, one could make an alternative choice of g and increase the bandwidth to widen the gap in AMSE at the possible expense of a larger finite sample bias.

4.3 Optimal inputs with polynomial approximations of higher order

For $S > 1$, the optimal (g, m_z) can again be reformulated as the optimal choice of a higher order kernel ω_z. Unlike in RP estimation, however, the higher order kernel does not necessarily entail the possibility of negative density estimates, since one can achieve a higher order kernel ω_z using a nonnegative m_z and a suitable g. For example, $m_z(t) = 3(1 - t^2)/4$ and $g_j(t) = (t + z)^{j+1}(t - 1)$ for $j = 1, 2, 3$ produces a fourth–order Epanechnikov kernel $\omega_1(t) = \frac{15}{32}(3 - 7t^2)(1 - t^2)$. Like in the linear case, the restrictions $\omega_z(1) = \omega_1(-1) = 0$ are necessary for the fourth–order Epanechnikov kernel to be AMSE–optimal. Without these or similar constraints on g, one can choose the non-bandwidth inputs to achieve zero asymptotic bias, i.e. make the squared bias asymptotically negligible relative to the variance. We must therefore restrict g to ensure an interior solution when we optimize the AMSE over (g, m_z) in the higher order cases, as well.

In general, one also needs to restrict g to derive an AMSE–optimal choice for the estimation of the derivatives of L, as well. The necessary conditions for the calculus of variations problem imply that each optimal g_j is a a polynomial of at most degree $S + 2$ for each j. Because the scale of g does not affect the estimates and $g_j(-z) = g_j(1) = 0$, one only needs to optimize over the remaining roots of g_j for each $j = 1, \ldots, S$. Characterizing the resulting minimization problem is routine, but it does not generally lead to an interior solution for the choice of bandwidth because the bias can be zero. For example, the bias can be made zero when $S = 2$ if $\int g_j t^2 \, \mathrm{d}t = 0$ for all j.

5 Applications

5.1 Treatment effects

It is well–known (see e.g. Hirano et al., 2003) that under an unconfoundedness assumption the average treatment effect can be expressed as

$$\mathbb{E}\Big(\frac{\boldsymbol{y}\boldsymbol{t}}{p(\boldsymbol{x})} - \frac{\boldsymbol{y}(1-\boldsymbol{t})}{1-p(\boldsymbol{x})}\Big),$$

where p is the propensity score, $\boldsymbol{y}$ the outcome variable, $\boldsymbol{x}$ a vector of regressors, and $\boldsymbol{t}$ a binary treatment variable. Let $p_1 = \mathbb{E}\boldsymbol{t}$ be the unconditional treatment probability. Let further f_1 denote the regressor density function conditional on treatment and f_0 the density conditional on nontreatment. Then $f(x) = f_1(x)p_1 + f_0(x)(1-p_1)$, which produces

$$\frac{1}{p(x)} = 1 + \frac{f_0(x)}{f_1(x)}\frac{1-p_1}{p_1}, \qquad \frac{1}{1-p(x)} = 1 + \frac{f_1(x)}{f_0(x)}\frac{p_1}{1-p_1},$$

such that the reciprocals of the propensity scores only depend on the ratios of the densities and the unconditional choice probabilities. In practice, $p(x)$ is often estimated using a logistic functional form and possibly a series expansion in x,[18] which implies the logarithm of the odds ratio is a polynomial in x. Our log–polynomial approximation to f_1 and f_0 similarly imply the odds ratio is log–polynomial. The data $\boldsymbol{x}$ will not typically be scalar–valued, so one could for instance use a linear index of regressors instead of the regressors themselves; see section 5.3 for an example of how one might estimate $p(x) = p^*(x^\intercal \theta_0)$. In any case, our approach is a natural local extension to the logit series estimator for $p(x)$.

5.2 Auctions

In first–price, sealed–bid procurement models with independent private values, it is well–known (see e.g. Guerre et al., 2000) that the inverse bid function is of the form $b - \bar{F}(b)/f(b)$, where $\bar{F}, f$ are the survivor and density functions of the minimum rival bid. Since the support of the bid distribution is assumed to have a lower bound in this literature (costs cannot be less than zero and hence neither are bids), boundary issues are a serious concern.

[18] This casual observation is supported by the fact that the built–in propensity score matching estimator in Stata defaults to the logit model.

So let $\Psi(y) = \bar{F}(y)/f(y)$ be the object of estimation. One way of estimating Ψ is to estimate $\bar{F}, f$ separately where f is estimated using the machinery in the main part of this paper and $\bar{F}$ is estimated by the empirical survivor function. This estimator has all the features of the estimator discussed earlier in the paper, positivity and boundary correction, both of which are desirable in the empirical auction setting. In particular, if the underlying cost distribution is approximately an exponential then so is the bid distribution and our estimator could be expected to work especially well.

The above approach is not specific to auction models. Indeed, consider the hazard function $H(x) = f(x)/\bar{F}(x)$. H can also be estimated using the machinery developed in our paper.

5.3 Semiparametric maximum likelihood

There are many examples of semiparametric maximum likelihood estimators. Here, we only consider a classical one, namely the Klein and Spady (1993) estimator of the coefficients in a semiparametric binary response model, which maximizes

$$\sum_{i=1}^{n} \left[\boldsymbol{y}_i \log \hat{\boldsymbol{p}}(\boldsymbol{x}_i^\intercal \theta) + (1 - \boldsymbol{y}_i) \log\{1 - \hat{\boldsymbol{p}}(\boldsymbol{x}_i^\intercal \theta)\}\right], \tag{6}$$

where $\hat{\boldsymbol{p}}$ is an estimator of the choice probability. Klein and Spady apply techniques to ensure that the estimates $\hat{\boldsymbol{p}}$ are positive and less than one, including trimming and adding a sample–size–dependent constant. Our method can be helpful since $p(t) = \Pr(\boldsymbol{y}_1 = 1 \mid \boldsymbol{x}_1^\intercal \theta = t) = f_1(t) p_1 / f(t)$, where f_j is the density of the linear index for observations with $\boldsymbol{y}_i = j$ and p_1 is the unconditional choice probability. Since $f(t) = p_1 f_1(t) + (1 - p_0) f_0(t)$, the infeasible contribution to the loglikelihood could be written as

$$\boldsymbol{y}_i \log \frac{f_1(\boldsymbol{x}_i^\intercal \theta)}{f_0(\boldsymbol{x}_i^\intercal \theta)} - \log\left(p_1 \frac{f_1(\boldsymbol{x}_i^\intercal \theta)}{f_0(\boldsymbol{x}_i^\intercal \theta)} + (1 - p_1) \right) + \text{constant},$$

such that it is only the ratio of f_1/f_0 that matters.

$$\boldsymbol{y}_i \log f_1(\boldsymbol{x}_i^\intercal \theta) + (1 - \boldsymbol{y}_i) \log f_0(\boldsymbol{x}_i^\intercal \theta) - \log f(\boldsymbol{x}_i^\intercal \theta) + \text{constant},$$

such that only the log of the conditional and unconditional densities matters. Obtaining conditions under which our estimator obtains the semiparametric efficiency bound, like the Klein and Spady estimator does, are well beyond the scope of this paper. In section 6, we provide simulation evidence that suggests our estimates of the log–densities are suitable substitutes for RP estimates in this application. Moreover, we show that an alternative method based on the infeasible score can outperform the Klein and Spady estimator.

5.4 Regression discontinuity design

One context in which the behavior of estimates near or at the boundary is of special importance is that of regression discontinuity design. For instance, Cattaneo et al. (2019) provide a test of continuity of the density function at the boundary using a boundary density estimator that is similar to the estimator in Lejeune and Sarda (1992) in that it is based on a quadratic expansion of the distribution function. Compared to that approach, our method requires the density to be nonzero at the boundary, which is a requirement for the regression discontinuity framework in any case. The bottom line is that our method will work better if the log density is approximately a low order polynomial near the boundary and theirs if the density itself is approximately a low order polynomial. This is borne out by our simulation results.

5.5 Other boundary–correction methods

The boundary correction method of Karunamuni and Zhang (2008) requires a well–behaved estimate of $L'(0)$, which is exactly what our method provides.

6 Simulations

6.1 The density and its derivative

The following simulation exercise compares the performance of our estimator for the density and its derivatives near the boundary with alternatives that also employ local polynomial approximations (Lejeune and Sarda, 1992; Cattaneo et al., 2019; Loader, 1996) and the generalized reflection method, which also estimates the derivative of the log–density near the boundary to remove the boundary effects of the RP estimator (Karunamuni and Alberts, 2005; Karunamuni and Zhang, 2008). For the local polynomial estimators, we use a local linear approximation to the density or log–density, depending on the method.[19]

We simulate 2,000 i.i.d. samples of size $n = 500$ and estimate f at points within one bandwidth of the boundary in order to compare the estimators away from, near, and at the boundary. Since the bandwidths differ across methods, so does the range of values at which the density is estimated in the presented results. The random variables are drawn from each of four distributions whose densities exhibit varying behaviors near their left boundary $x = 0$. The first is a beta distribution with density $f_1(x) = \theta(1 - x)^{\theta-1}$, which is in fact polynomial in x for integer values of θ, which one would expect to favor LS–CJM, although that is not reflected in our simulations if $\theta > 3$. The second design is a normal distribution with mean $\theta/2$ and variance

[19] This corresponds to a local–quadratic polynomial approximation to the distribution function in Lejeune and Sarda (1992) and Cattaneo et al. (2019).

one, left–truncated at zero. This density is log–quadratic, which could favor our method and Loader's. The third and fourth designs are $f_3(x) = (e^{-x} + \theta x e^{-x})/(1 + \theta)$ and $f_4(x) = (e^{-x} + \theta x^2 e^{-x})/(1 + 2\theta)$.

For each simulation design, we use our local–linear estimator to estimate f and its derivative using $g(t) = (t + z)(t - 1)(1 + zt - t)$ (PS),[20] Loader's local likelihood estimator (Loader), a local polynomial regression of the empirical CDF (LS–CJM), and a generalized reflection estimator (KZ).[21] Wherever a kernel is required, we use the Epanechnikov kernel $k(u) = 3(1 - u^2)/4$ or a truncated version thereof. This choice is not optimal at the boundary for Loader's estimator, but the efficiency loss is small.[22] For our method and Loader's, we obtain an estimate of f' by multiplying the estimate of f by the estimate of L'. We do not estimate $f'(x)$ using the generalized reflection method (KZ) for $x > 0$ because Karunamuni and Zhang (2008) do not discuss estimation of the derivatives anywhere except at the boundary.

Where possible, we use the asymptotically optimal bandwidth sequences for $z = 1$ and $z = 0$. For intermediate values of z, we linearly interpolate the bandwidth.[23] For the generalized reflection method, which requires separate bandwidths and finite–difference approximation to estimate L' in a first step, we do not develop a theory of the asymptotically optimal inputs. Instead, we select a finite–differencing scheme and choose a combination of auxiliary bandwidths so that the pilot estimate of L' at the boundary and the density estimate at $z = 1$ have the same asymptotic variances as our method using $g(t) = (t + z)(1 - t)$. Specifically, we use a main bandwidth equal to the asymptotically optimal bandwidth for our method at $z = 1$, and we estimate $L'(0)$ using $\{\log \hat{f}_{nh_{L'}}(2h_{L'}) - \log \hat{f}_{nh_{L'}}(0)\} \,/\, (2h_{L'})$, where $\hat{f}_{nh_{L'}}(2h_{L'})$ and $\hat{f}_{nh_{L'}}(0)$ are respectively a kernel and a boundary–kernel estimator for the density whose variance is proportional to $1/nh_{L'}$.

Comparing the square root of the mean square error (RMSE) of the density estimates at zero in table 1 and the RMSE for the derivative in table 2, there is no clear ranking of the estimators. Figure 1 depicts the bias and RMSE of the local linear estimators for $\theta = 4$. The linear approximation of f (LS–CJM) performs well for the beta distribution, but has difficulty estimating the truncated normal density and the estimate is often negative.[24] KZ is generally neither the best nor the worst of the estimators we consider here, but we

[20] The choice of g used here is a simple formula that interpolates the optimal choice at the boundary and away from the boundary.

[21] The difference between the estimated coefficients in Cattaneo et al. (2019) and Lejeune and Sarda (1992) is $O_p(1/n)$. We only provide simulation results for CJM's local polynomial regression of the empirical distribution function on x_i, but we note that Lejeune and Sarda's local polynomial approximation to the empirical distribution is very similar.

[22] The relative efficiency of Loader's estimator using the triangular and Epanechnikov kernels at the boundary is about 1.008, meaning the Epanechnikov kernel requires a sample size 1.008 times larger to achieve the same MSE as the triangular kernel. We therefore expect the Epanechnikov kernel with $n = 504$ to have the same MSE as the triangular kernel with $n = 500$.

[23] Specifically, we use $h = h_0\big(1 - \min\{x/h_1, 1\}\big) + h_1 \min\{x/h_1, 1\}$, where h_0 and h_1 are the asymptotically optimal bandwidths at $z = 0$ and $z = 1$ and x is the point of evaluation. This bandwidth selection rule implies that the same window is used to estimate the density and its derivatives at all points within h_1 of the boundary, i.e. $x + h = 2h_1$ for all $x < h_1$.

[24] The LS–CJM estimator for $f(0)$ was negative in 0%, 15%, <0.001%, and 1.2% of the samples from the four respective designs. If we replace negative estimates of $f_2(0)$ with zero, the bias is reduced, but the RMSE of 0.036 is still large relative to the other

acknowledge that we have not optimized the inputs into the KZ estimator as thoroughly as the other estimator's inputs. We also note that KZ provides a familiar benchmark away from the boundary because it is simply the RP density estimator using an Epanechnikov kernel for $z = 1$.

Table 1: RMSE of the estimators for the density at the boundary ($\theta = 4$).

Estimator	f_1	f_2	f_3	f_4
LS–CJM	0.067	0.041	0.082	0.051
PS	0.066	0.023	0.069	0.041
KZ	0.077	0.025	0.076	0.043
Loader	0.065	0.021	0.065	0.039

Table 2: RMSE of the estimators for the derivative of the density at the boundary ($\theta = 4$).

Estimator	f_1	f_2	f_3	f_4
LS–CJM	0.145	0.113	0.397	0.243
PS	0.174	0.028	0.442	0.197
KZ	0.104	0.157	1.287	0.497
Loader	0.174	0.029	0.408	0.180

As expected, PS and Loader perform similarly for all z in each design. All four estimators have the same asymptotic variance away from the boundary, but the differences in the bias due to differences in L'' and f'' cause the relatively minor differences in the AMSE. Local to the boundary, however, we observe more significant differences between the estimators. In three of the four designs, the polynomial approximation to the density function has a greater AMSE than the log–polynomial approximation methods (PS and Loader), and the gap tends to widen at the boundary. The generalized reflection method is relatively well behaved except in the first design in which the bias of the density estimate is very large and negative at the boundary, despite the fact that $f_1'(0)$ is well estimated.

In figure 2 we plot the bias and RMSE for the derivative of f. In three of the four simulation designs, PS has a smaller RMSE than Loader in the interior region, which we expected because $g(t) = (t + z)(t - 1)$ minimizes the AMSE in L' using the given bandwidth sequence.

The following figures show the estimated log–density and its first derivative. Because the density estimated using a polynomial approximation to F may be negative, the logarithm of the LS–CJM estimator is often a large negative number and may be undefined. The sampling distributions of the estimators for the log–density

estimators. Intuitively, the normal asymptotic approximation to the sampling distribution does not accurately capture the bias–variance tradeoff of this censored estimator; hence, it performs suboptimally in finite samples. The MSE could be reduced by re–estimating with a larger bandwidth.

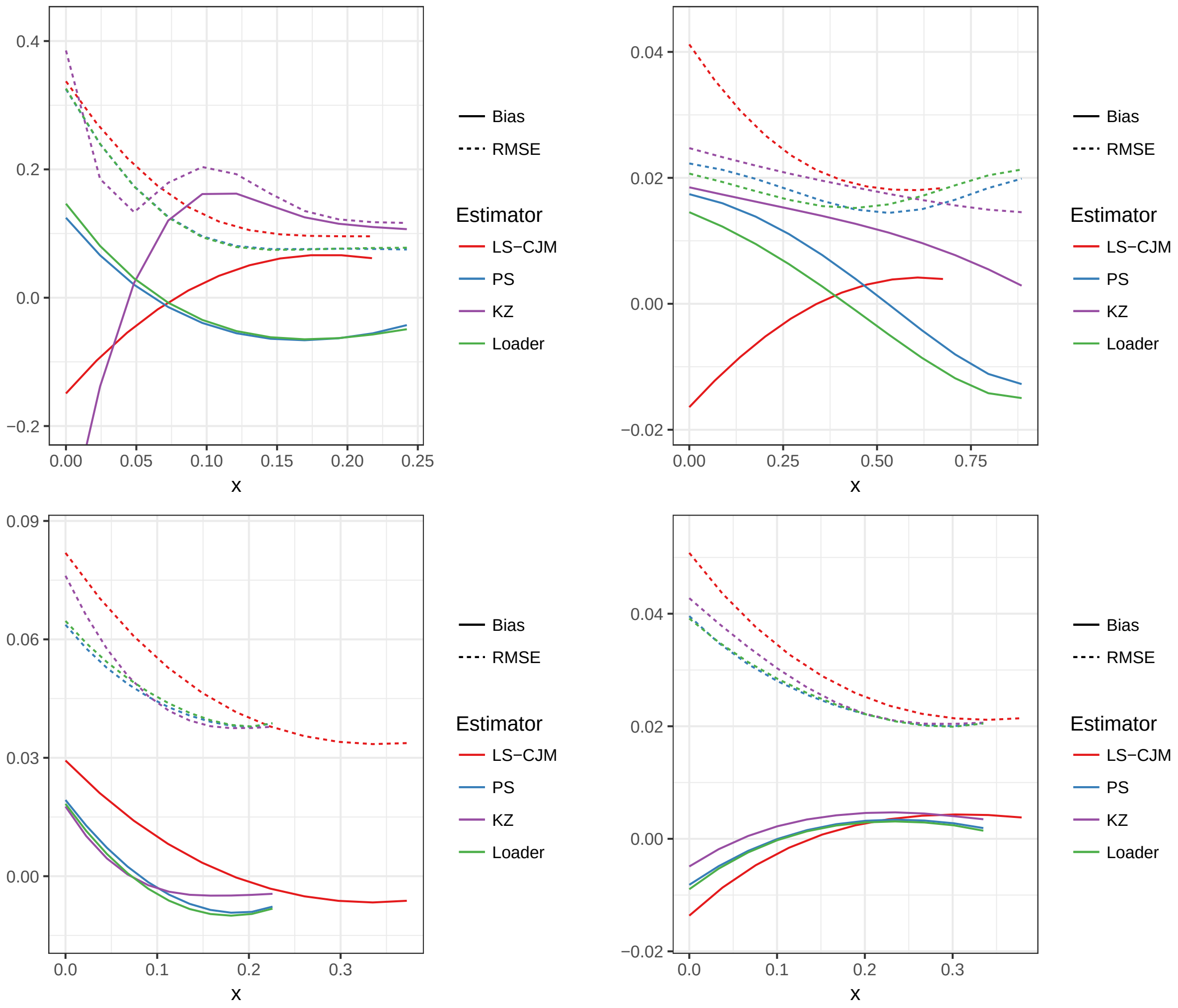

Figure 1: The bias and RMSE for estimators of f_1, f_2, f_3, and f_4 (arranged from left to right) for $\theta = 4$.

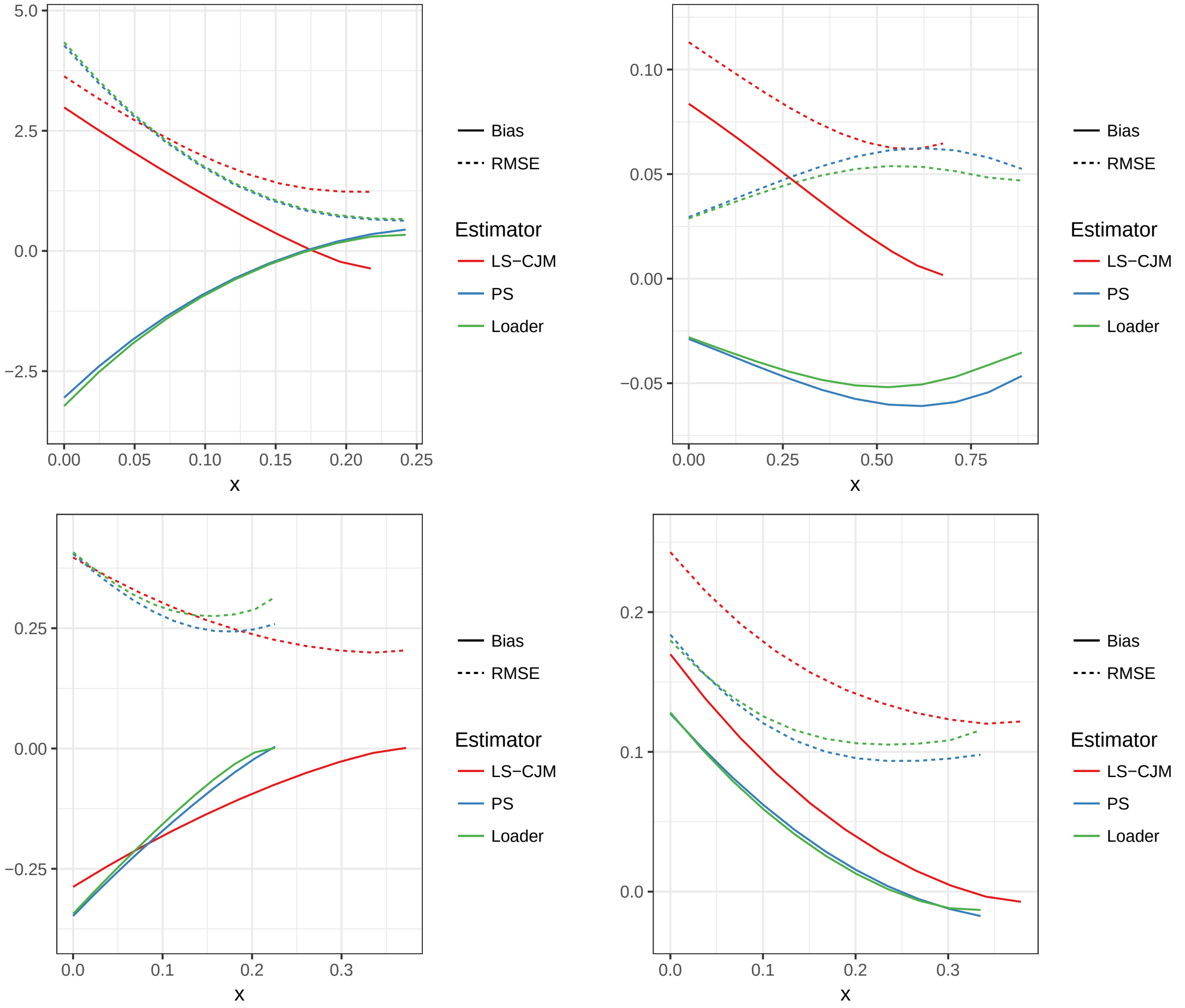

Figure 2: The bias and RMSE for estimators of f_1', f_2', f_3', and f_4' (arranged from left to right) for $\theta = 4$.

and its derivative may be highly skewed as a result, and the bias and RMSE of the density estimate are less informative measures, especially when compared with the other estimators. In order to make reasonable visual comparisons, we plot the median error and median absolute error (MAE) instead of the bias and RMSE.

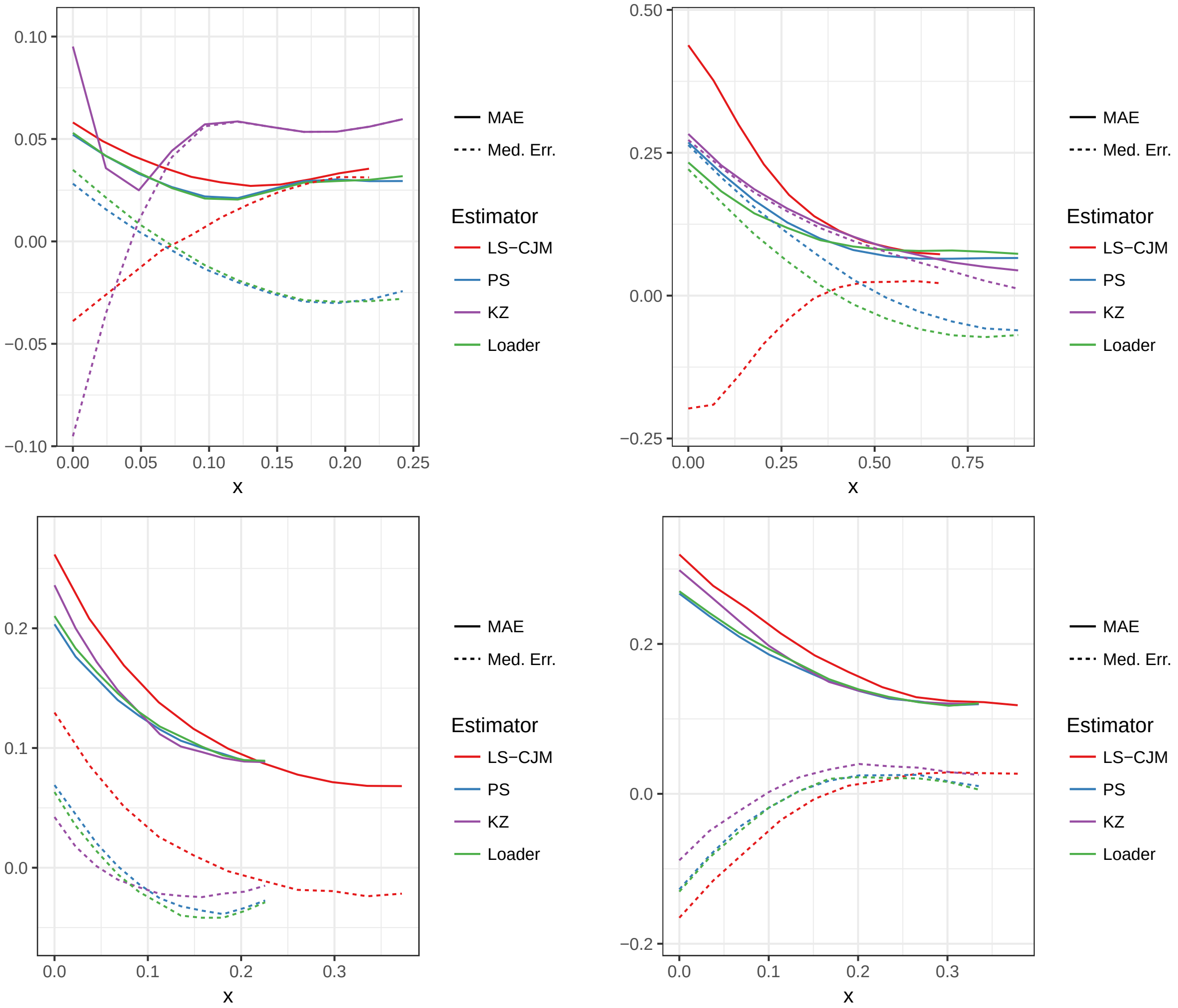

Figure 3: The median error and MAE for estimators of L_1, L_2, L_3, and L_4 (arranged from left to right) for $\theta = 4$.

6.2 Semiparametric maximum likelihood

We now compare our approach with traditional RP estimation when the density is used in the semiparametric estimation of a binary choice model. Let $\boldsymbol{y}_i^* = \boldsymbol{x}_i^\intercal \theta_0 + \boldsymbol{\epsilon}_i$, where $\boldsymbol{\epsilon}_i$ is a random disturbance independent of $\boldsymbol{x}_i$. The data consists of an independent sample of observations $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ for $i = 1, \dots, n$, where $\boldsymbol{y}_i = \mathbb{1}(\boldsymbol{y}_i^* > 0)$. Let $p(x^\intercal \theta) = \mathbb{E}(\boldsymbol{y}_i \mid \boldsymbol{x}_i = x)$. Klein and Spady (1993) show that the quasi-maximum likelihood estimator

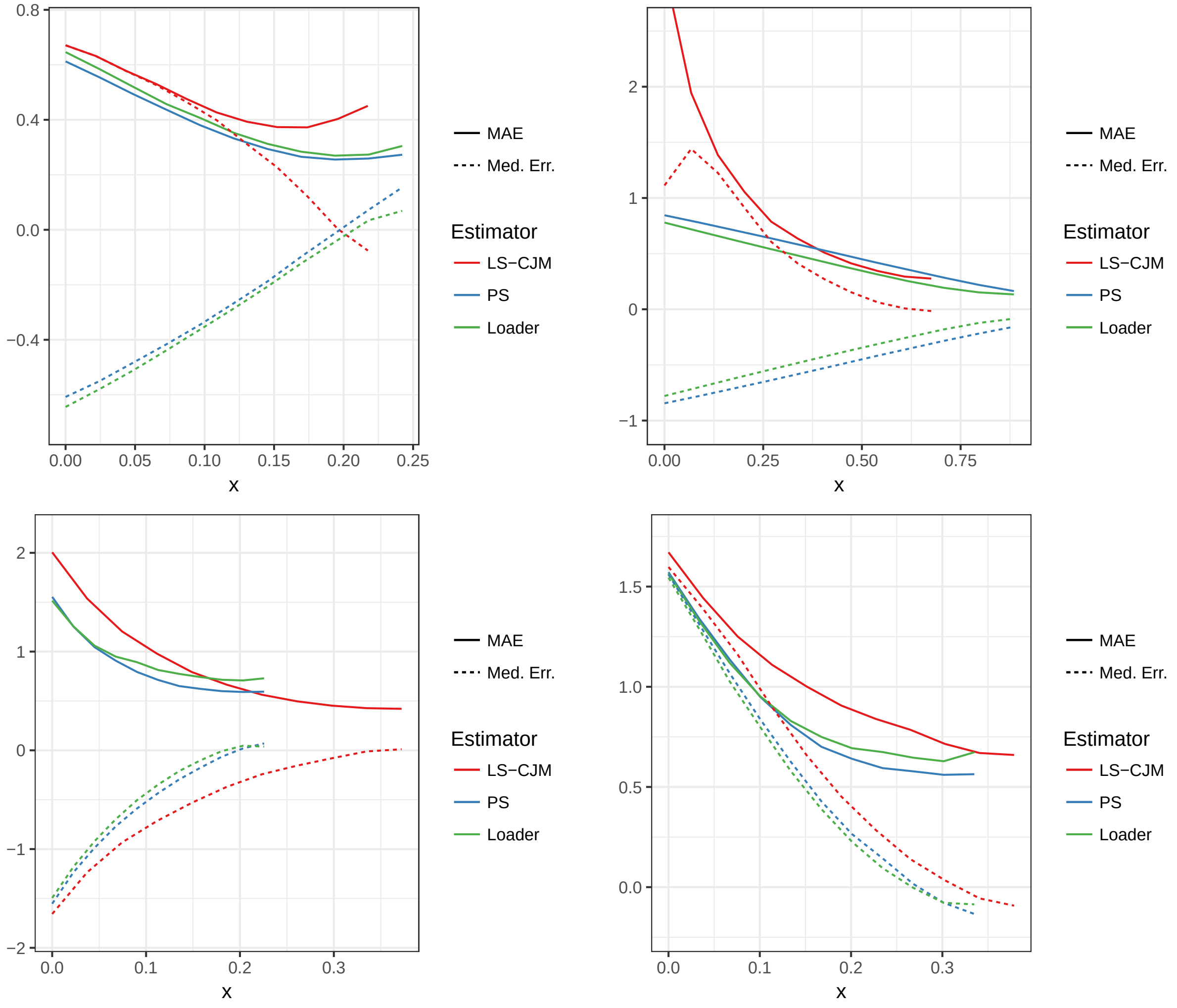

Figure 4: The median error and MAE for estimators of L_1', L_2', L_3', and L_4' (arranged from left to right) for $\theta = 4$.

obtained by substituting a kernel–based estimator for $p(\boldsymbol{x}_i^\intercal\theta)$ in (6) achieves the semiparametric efficiency bound under some smoothness assumptions. In order to adequately control the stochastic order of the bias and variance in the nonparametric estimates of $p(\boldsymbol{x}_i^\intercal\theta)$, Klein and Spady suggest using either a fourth–order kernel with a fixed bandwidth or a second–order kernel with a locally adaptive bandwidth. If one uses a fourth–order kernel, $\hat{\boldsymbol{p}}(\boldsymbol{x}_i^\intercal\theta)$ can be negative. As a result, they recommend replacing $\hat{\boldsymbol{p}}(\boldsymbol{x}_i^\theta)$ and $1-\hat{\boldsymbol{p}}(\boldsymbol{x}_i^\intercal\theta)$ with $\hat{\boldsymbol{p}}^2(\boldsymbol{x}_i^\intercal\theta)$ and $\{1-\hat{\boldsymbol{p}}(\boldsymbol{x}_i^\intercal\theta)\}^2$ in the loglikelihood and then dividing by two. We use the semiparametric estimator based on a fourth–order kernel with a fixed bandwidth to compare with our approach.[25]

Using our approach, one can estimate $\log p(\boldsymbol{x}_i^\intercal\theta)$ with $\log\bar{\boldsymbol{y}}+\log\hat{\boldsymbol{f}}_1(\boldsymbol{x}_i^\intercal\theta)-\log\hat{\boldsymbol{f}}(\boldsymbol{x}_i^\intercal\theta)$, where $\bar{\boldsymbol{y}}=\sum\boldsymbol{y}_i/n$, $\hat{\boldsymbol{f}}_1$ is an estimate of the density of $\boldsymbol{x}_i^\intercal\theta$ conditional on $\boldsymbol{y}_i=1$, and $\hat{\boldsymbol{f}}$ is the estimated unconditional density. For example, we estimate the derivatives of the log–density at $\boldsymbol{x}_i^\intercal\theta$ using

$$\frac{1}{nh^2}\sum_{j\neq i} g'\left(\frac{(\boldsymbol{x}_j-\boldsymbol{x}_i)^\intercal\theta}{h}\right) = -\sum_{s=1}^{S}\hat{\boldsymbol{\beta}}_s\frac{1}{nh}\sum_{j\neq i} g\left(\frac{(\boldsymbol{x}_j-\boldsymbol{x}_i)^\intercal\theta}{h}\right)\frac{[(\boldsymbol{x}_j-\boldsymbol{x}_i)^\intercal\theta]^{s-1}}{(s-1)!}\,.$$

The density at $\boldsymbol{x}_i^\intercal\theta$ is then estimated by

$$\frac{\displaystyle\frac{1}{nh}\sum_{j\neq i} m_z\left(\frac{(\boldsymbol{x}_j-\boldsymbol{x}_i)^\intercal\theta}{h}\right)}{\displaystyle\int_{-z}^{1} m_z(t)\exp\left(\sum_{s=1}^{S}\hat{\boldsymbol{\beta}}_s t^s h^s/s!\right)\,\mathrm{d}t}\,.$$

Notice that we omit the i-th observation when estimating the density at $\boldsymbol{x}_i^\intercal\theta$. Similarly, we can estimate $\log\{1-p(\boldsymbol{x}_i^\intercal\theta)\}$ by $\log(1-\bar{\boldsymbol{y}})+\log\hat{\boldsymbol{f}}_0(\boldsymbol{x}_i^\intercal\theta)-\log\hat{\boldsymbol{f}}(\boldsymbol{x}_i^\intercal\theta)$. As in Klein and Spady (1993), we then maximize the quasi–likelihood to obtain a semiparametric estimate of θ_0, though we do not need to modify the quasi–likelihood to allow for negative density estimates. We will refer to this as the PS-likelihood estimator.

Alternatively, we can write the (population) score as

$$s(\theta)=\mathbb{E}\left[\boldsymbol{y}_i L_0'(\boldsymbol{x}_i^\intercal\theta)+(1-\boldsymbol{y}_i)L_1'(\boldsymbol{x}_i^\intercal\theta)-L'(\boldsymbol{x}_i^\intercal\theta)\right]$$

where L_0', L_1', and L' are the log–derivatives of the density of $\boldsymbol{x}_i^\intercal\theta_0$. We then estimate the score using

$$s_n(\theta,\hat{\boldsymbol{L}}_1',\hat{\boldsymbol{L}}_0',\hat{\boldsymbol{L}}')=\frac{1}{n}\sum_{i=1}^{n}\left\{\boldsymbol{y}_i\hat{\boldsymbol{L}}_1'(\boldsymbol{x}_i^\intercal\theta)+(1-\boldsymbol{y}_i)\hat{\boldsymbol{L}}_0'(\boldsymbol{x}_i^\intercal\theta)-\hat{\boldsymbol{L}}'(\boldsymbol{x}_i^\intercal\theta)\right\},$$

[25] Klein and Spady (1993) introduce two forms of trimming to ensure that the quasi–loglikelihood converges uniformly in θ. In their simulation study, they find that trimming does not appreciably affect the quantitative results. Following their example, we report only the untrimmed versions of the estimators in this section.

where $\hat{\boldsymbol{L}}_1'$, $\hat{\boldsymbol{L}}_0'$, and $\hat{\boldsymbol{L}}'$ are the estimated log–derivatives of the conditional and unconditional densities of $\boldsymbol{x}_i^\intercal \theta$. A semiparametric estimator for θ_0 is then obtained by solving for the value(s) of θ for which the score is zero. In case there are multiple solutions, we take the solution that attains the highest loglikelihood. We will refer to this as the PS-score estimator. Note that we estimate the score of the true model, as opposed to the derivative of the quasi–loglikelihood with respect to θ. Consequently, the PS-score and PS-likelihood estimators can be quite different. In the simulation results reported below, the correlation between the estimates is approximately 0.45, which is in fact weaker than the correlation between the PS-likelihood and Klein-Spady (KS) estimators (0.59).

In the following exercise, we use the same homoskedastic simulation design as in Klein and Spady (1993). The covariates are two–dimensional: $\boldsymbol{x}_{i1}$ is chi-square distributed with three degrees of freedom truncated at six on the right and $\boldsymbol{x}_{i2}$ is an independent standard normal variable truncated at two and negative two. Both covariates are standardized to have zero mean and unit variance. The random disturbance $\boldsymbol{\epsilon}_i$ is standard normal and independent of the $\boldsymbol{x}$'s.

We take $\theta_{01} = \theta_{02} = 1$. Because θ_0 is only identified up to scale, we normalize $\|\theta\| = \sqrt{2}$ and $\theta_1 > 0$.[26] For the KS estimator, we use a fourth–order Epanechnikov kernel to estimate the density of $\boldsymbol{x}_i^\intercal \theta$. For our estimators, we choose $g_j(t) = (t + z_\ell)^{S-j+2}(t - z_r)^{j+1}$, where z_ℓ and z_r measure how close the point of evaluation is to the boundary of the support of the $\boldsymbol{x}_i^\intercal \theta$, and we use a truncated second–order Epanechnikov kernel for m_z. In the interior, the resulting ω_1 is a fourth–order Epanechnikov kernel when $S = 3$. Hence, the pointwise variance of the dominant terms in the density estimates are the same for both KS and PS-likelihood. The observed differences between them can be largely attributed to differences in the bias and the boundary correction. For the PS-score estimator, we use $S = 2$ and recommend a bandwidth sequence proportional to $n^{-2/7}$ so that $|\boldsymbol{s}_n(\theta, \hat{\boldsymbol{L}}_1', \hat{\boldsymbol{L}}_0', \hat{L}') - s_n(\theta, L_1', L_0', L')|$ is asymptotically negligible compared to $|\boldsymbol{s}_n(\theta, L_1', L_0', L') - s(\theta)|$.[27] This bandwidth sequence converges to zero more quickly than the bandwidths used by Klein and Spady (1993) because the bias in the estimates of the log–derivatives of the density are proportional to h^2 when $S = 2$.

For the sake of comparison, however, we use the same interior bandwidth $h \approx 1.4$ with all three methods.[28] Near the boundaries of the support of the data, we multiply h by $2 - \min(z_l, z_r)$ so that the same window is

[26] Assuming $\theta_1 > 0$ is a normalization because the propensity $p(x^\intercal \theta)$ is not restricted to be increasing in $x^\intercal \theta$.

[27] One should trim the observations as discussed in Klein and Spady (1993) to ensure convergence in the quasi–score is uniform in θ, but we do not use any trimming in the reported simulation results.

[28] Following Klein and Spady (1993) we select h proportional to $n^{-1/6.02}$. Unlike in their simulation exercise, however, we must also select the constant of proportionality in the bandwidth sequence because we do not use a locally adaptive bandwidth. To do so, we make a somewhat arbitrary appeal to Silverman's rule–of–thumb for the fourth–order Epanechnikov kernel $h = 2.1276\sigma n^{-1/9}$, where σ is the standard deviation of the observations. After normalizing $\|\theta\|_2 = \sqrt{2}$, the standard deviation of $\boldsymbol{x}_i^\intercal \theta$ is $\sqrt{2}$. Hence, we use $h = 2.1276\sqrt{2}n^{-1/6.02}$.

used to estimate the log–density and its derivatives at all points within distance h of one of the boundaries.

The simulation was repeated 1000 times. Table 3 reports the simulated bias, median error, variance, and median absolute deviation in the estimates of θ_2 from a sample of $n = 100$ observations. By design, the PS-likelihood and KS estimators are as similar as possible given the two different approaches to density estimation. It is therefore unsurprising that neither dominates the other in terms of bias and variance. The variance of the PS-likelihood estimator is noticeably larger than that of the KS estimator, which one would expect in light of the boundary correction which reduces bias near the boundaries at the cost of an increased variance.

The PS-score estimator has the smallest variance, which one would also expect because it uses the same bandwidth with a lower order polynomial approximation to the log–density. Interestingly, the PS-score estimator has the smallest bias, as well, despite using the same bandwidth. The smaller bias is partially due to the fact that the distribution of the PS-score estimator is less skewed than the other two sampling distributions, as evidenced by the median errors. The median error of the PS-likelihood estimator is much closer to zero and is less than that of PS-score, as we would expect from the higher order local polynomial approximation. On the other hand, the PS-score estimator improves upon the KS estimator according to all four performance measures. The relatively large bias and median error in KS may be symptomatic of the frequently observed phenomenon in which higher order kernels and polynomial approximations do not demonstrate their bias–reducing properties in small samples. In addition, the squaring of $\hat{\boldsymbol{p}}$ and $1 - \hat{\boldsymbol{p}}$ in the modified quasi–likelihood adds another source of bias to the KS estimator. Asymptotically, this modification has no effect, but it essentially replaces the estimated $\hat{\boldsymbol{p}}$ with its absolute value since $\log(x^2)/2 = \log|x|$, which naturally tends to decrease the variance and increase the bias of the KS estimator in finite samples.

Table 3: Simulation results for semiparametric estimates of the linear parameter in a binary choice model.

Estimator	Bias	Median Error	Variance	Median Absolute Deviation
KS	−0.029	−0.008	0.040	0.122
PS-likelihood	−0.026	−0.004	0.045	0.138
PS-score	−0.013	−0.007	0.022	0.088

This example illustrates that estimating the log–derivative of the density can provide a more direct route to the researcher's ultimate goal and may improve on the finite–sample performance of existing methods, even semiparametrically efficient ones. Though it would not be prudent to draw general conclusions based on this example, we hope it is uncontroversial to note that correcting for boundary issues and nonpositive densities estimates has predictable effects on the bias and variance of semiparametric estimators. The approach developed in this paper provides a single framework for dealing with boundary and nonpositivity issues so that researchers can choose the inputs (h, g) and possibly m_z to meet their needs.

7 Conclusion

We develop asymptotically normal nonparametric estimators based on a log–polynomial approximation of the unknown density function. By approximating the log–density with a polynomial, we can guarantee that our estimated density is nonnegative; and by using a polynomial approximation instead of a local constant approximation, we achieve the optimal rates of convergence at the boundary of the support as well as in the interior.

Because our approach allows for a relatively larger degree of customization—the researcher must specify a bandwidth, a kernel, and a vector–valued function g that is zero at the extremes of its support—we explore the optimal set of inputs. Unlike the standard analysis of optimal kernel and bandwidth inputs, our estimator is nonnegative at the point of evaluation under a relaxed set of constraints. Because these constraints were needed in order to derive the optimal kernel for use with alternative methods, there is no interior solution to the optimal choice of inputs using our approach. If one imposes the additional restriction that g cannot cross zero, the AMSE–minimizing inputs achieve the same asymptotic variance as the standard kernel density estimator with the Epanechnikov kernel in the interior and the optimal boundary kernel derived in Zhang and Karunamuni (1998) at the boundary. Otherwise, if the researcher uses a larger bandwidth, marginal reductions in the asymptotic mean square error are possible. In the extreme, an asymptotically unbiased estimator can be obtained using a suitable choice of g. This approach would be analogous to the use of a higher–order kernel in standard kernel density estimation, except that our methods would still guarantee a nonnegative estimate.

More generally, our approach is based on a sample analog to partial integration and can be applied to other settings, as well, such as the estimation of hazard functions or propensity scores.

In simulation exercises, we explore the finite–sample behavior of our approach to density estimations as well as the estimation of a semiparametric binary choice model. Using the constrained–optimal choice of inputs designed to mimic the asymptotic behavior of alternative density estimators as closely as possible, we unsurprisingly find comparable finite–sample performance, especially when compared with the local likelihood estimator of Loader (1996). The advantage of our approach compared with Loader (1996) is that our estimator is computationally simpler, which can be an important feature when the density is used as an input in a estimation algorithm. As an illustration, we apply our approach to the estimation of the semiparametric binary choice model studied by Klein and Spady (1993) and find that our method reduces the finite–sample bias and variance in their Monte-Carlo experiment.

A Proofs

A.1 Definitions

To simplify the argument of g_j, define $g_{hxj}(y) = g_j\left(\frac{y-x}{h}\right)$, noting that $g'_{hxj}(y) = \frac{1}{h}g'_j\left(\frac{y-x}{h}\right)$. Rewrite (1) as $\ell_{hx} = \boldsymbol{R}_{hx}\beta + r_{hx}$, where ℓ_{hx}, $\boldsymbol{R}_{hx}$ are a vector and a matrix, whose elements are

$$\begin{cases} \ell_{hxj} = \mathbb{E}\left(g'_{hxj}(\boldsymbol{x}_1)\,\middle|\, x - zh \le \boldsymbol{x}_1 \le x + h\right), \\ R_{hxjs} = -\mathbb{E}\left(g_{hxj}(\boldsymbol{x}_1)\dfrac{(\boldsymbol{x}_1 - x)^{s-1}}{(s-1)!}\,\middle|\, x - zh \le \boldsymbol{x}_1 \le x + h\right), \end{cases} \qquad j, s = 1, 2, \dots, \tag{7}$$

and where $r_{hx} = \ell_{hx} - \boldsymbol{R}_{hx}\beta$ is the vector with elements

$$r_{hjx} = \ell_{hjx} - \sum_{s=1}^{S} R_{hjsx}\beta_s = R_{hj,S+1,x}\beta_{S+1} + o\left(h^S\right). \tag{8}$$

Let $\hat{\boldsymbol{\ell}}_{hx}$, $\hat{\boldsymbol{R}}_{hx}$ be sample analogs of ℓ_{hx}, $\boldsymbol{R}_{hx}$, e.g.

$$\hat{\boldsymbol{\ell}}_{hxj} = \frac{\displaystyle\sum_{i=1}^{n} g'_{hxj}(\boldsymbol{x}_i)}{n\{\hat{\boldsymbol{F}}(x+h) - \hat{\boldsymbol{F}}(x - zh)\}},$$

where $\hat{\boldsymbol{F}}$ is the empirical distribution function.

Define $\xi_{js}(t) = -\left\{g_j(t)t^{s-1}/(s-1)! = (z+t)^j(1-t)t^{s-1}/(s-1)!\right\}\mathbb{1}(-z \le t \le 1)$ and $\tilde{\beta} = R_{hx}^{-1}\ell_{hx}$.

A.2 Estimation of derivatives

A.2.1 Bias

Lemma 1. $\Omega_{js} = \int_{-z}^{1} \xi_{jsz}(t)\,\mathrm{d}t = \frac{1}{(s-1)!}\int_{-z}^{1}(z+t)^j(1-t)t^{s-1}\,\mathrm{d}t$.

Proof. The right hand side equals

$$\frac{1}{(s-1)!}\int_{-z}^{1}(z+t)^j\{1 + z - (z+t)\}(z+t-z)^{s-1}\,\mathrm{d}t =$$
$$\frac{1}{(s-1)!}\left((1+z)\int_{-z}^{1}(z+t)^j(z+t-z)^{s-1}\,\mathrm{d}t - \int_{-z}^{1}(z+t)^{j+1}(z+t-z)^{s-1}\,\mathrm{d}t\right) =$$

$$\frac{1}{(s-1)!}\sum_{i=0}^{s-1}\binom{s-1}{i}(-z)^i\left((1+z)\int_{-z}^{1}(z+t)^{j+s-1-i}\,\mathrm{d}t-\int_{-z}^{1}(z+t)^{j+s-i}\,\mathrm{d}t\right)=$$

$$\sum_{i=0}^{s-1}\frac{(-z)^i(1+z)^{j+s+1-i}}{i!(s-1-i)!(j+s-i)(j+s-i+1)}=\sum_{i=0}^{s-1}\sum_{t=0}^{i}\binom{i}{t}\frac{(-1)^{i+t}(1+z)^{j+s+1-t}}{i!(s-1-i)!(j+s-i)(j+s-i+1)}$$

$$=\sum_{t=0}^{s-1}(1+z)^{j+s+1-t}\sum_{i=t}^{s-1}\frac{(-1)^{i+t}}{t!(i-t)!(s-1-i)!(j+s-i)(j+s-i+1)}=\sum_{t=0}^{s-1}(1+z)^{j+s+1-t}\frac{(-1)^{s+1+t}(s-t)j!}{(j+s+1-t)!t!}.\quad\square$$

Lemma 2. $r_{hjx}=\ell_{hjx}-\sum_{s=1}^{S}\beta_s R_{hjsx}=\beta_{S+1}h^S\Omega_{j,S+1}/(1+z)+o(h^S)$.

Proof. From (7) and (8) it follows that

$$r_{hxj}=\ell_{hxj}-\sum_{s=1}^{S}\beta_s R_{hxjs}=-\frac{1}{F(x+h)-F(x-zh)}\int_{x-zh}^{x+h}g_{hxj}(y)L'(y)\,\mathrm{d}y-\sum_{s=1}^{S}\beta_s R_{hjsx}=$$

$$-\frac{\beta_{S+1}}{F(x+h)-F(x-zh)}\int_{x-zh}^{x+h}g_{hxj}(y)\frac{(y-x)^S}{S!}f(y)\,\mathrm{d}y+o(h^S)$$

$$=-\frac{h^{S+1}\beta_{S+1}}{F(x+h)-F(x-zh)}\int_{-z}^{1}\left\{(t+z)^{j+1}-(1+z)t^j\right\}\frac{t^S}{S!}f(x+th)\,\mathrm{d}t+o(h^S)$$

$$=-\frac{h^S\beta_{S+1}}{f(x)(1+z)}\int_{-z}^{1}\xi_{j,S+1,z}(t)f(x)\,\mathrm{d}t+o(h^S)$$

$$=\frac{h^S\beta_{S+1}\Omega_{j,S+1}(z)}{1+z}+o(h^S),$$

by the mean value theorem and lemma 1. □

Lemma 3. $R_{hxjs}=h^{s-1}\Omega_{js}/(1+z)+o(h^{s-1})$.

Proof. Repeat the steps of the proof of lemma 2. □

A.2.2 Distribution

Let $\hat{\boldsymbol{B}}_{hxjs}=\hat{\boldsymbol{R}}_{hxjs}\{\hat{\boldsymbol{F}}(x+h)-\hat{\boldsymbol{F}}(x-zh)\}/h^s$, $B_{hxjs}=R_{hxjs}\{F(x+h)-F(x-zh)\}/h^s$, $\hat{\boldsymbol{a}}_{hx}=\hat{\boldsymbol{\ell}}_{hx}\{\hat{\boldsymbol{F}}(x+h)-\hat{\boldsymbol{F}}(x-zh)\}/h$, and $a_{hx}=\ell_{hx}\{F(x+h)-F(x-zh)\}/h$.

Lemma 4. $B_{hxjs}=f(x)\Omega_{js}+o(1)$.

Proof. Follows from lemma 3. □

Lemma 5. $\hat{\boldsymbol{B}}_{hx}-B_{hx}=O_p(1/\sqrt{nh})$ *and* $\sqrt{nh^3}(\hat{\boldsymbol{a}}_{hx}-a_{hx})\xrightarrow{d}N\{0,f(x)(1+z)V\}$.

Proof. We first show the result for $\hat{\boldsymbol{a}}_{hx}$.

$$\sqrt{nh^3}(\hat{\boldsymbol{a}}_{hx} - a_{hx}) = \frac{1}{\sqrt{nh}}\sum_{i=1}^{n}\big\{g'_{hxj}(x_i) - \mathbb{E}g'_{hxj}(\boldsymbol{x}_1)\big\} \xrightarrow{d} N\{0, f(x)V\}\,. \tag{9}$$

The normality of the limit follows from a standard central limit theorem, e.g. Eicker (1966). Because the estimator is linear and $\{x_i\}$ is i.i.d., the asymptotic mean and variance are easily obtained, e.g. the (j,s)–element of the asymptotic variance matrix is

$$nh^3\,\mathrm{Cov}\big(\hat{\boldsymbol{a}}_{hxj}, \hat{\boldsymbol{a}}_{hxs}\big) = h\,\mathrm{Cov}\big(g'_{hxj}(\boldsymbol{x}_1), g'_{hxs}(\boldsymbol{x}_1)\big) = f(x)\int_{-z}^{1} g'_j(t)g'_s(t)\,\mathrm{d}t + o(1) = f(x)V + o(1).$$

Now, for any $j, s = 1, \ldots, S$,

$$\sqrt{nh}(\hat{\boldsymbol{B}}_{hxjs} - B_{hxjs}) = \frac{1}{\sqrt{nh}}\sum_{i=1}^{n}\left\{\xi_{js}\Big(\frac{\boldsymbol{x}_i - x}{h}\Big) - \mathbb{E}\xi_{js}\Big(\frac{\boldsymbol{x}_1 - x}{h}\Big)\right\},$$

which by standard kernel estimation theory has a limiting mean zero normal distribution and is hence $O_p(1)$. □

Proof of theorem 1. Recall that $\tilde{\beta} = R_{hx}^{-1}\ell_{hx}$. The bias result then follows immediately from lemmas 2 and 3. Now the asymptotic distribution. Note that $\tilde{\beta} = \Lambda_h^{-1}B_{hx}^{-1}a_{hx}$. Likewise $\hat{\boldsymbol{\beta}} = \Lambda_h^{-1}\hat{\boldsymbol{B}}_{hx}^{-1}\hat{\boldsymbol{a}}_{hx}$. Thus,

$$\begin{aligned}\sqrt{nh^3}\Lambda_h(\hat{\boldsymbol{\beta}} - \tilde{\beta}) = \sqrt{nh^3}(\hat{\boldsymbol{B}}_{hx}^{-1} - B_{hx}^{-1})(\hat{\boldsymbol{a}}_{hx} - a_{hx}) + \sqrt{nh^3}(\hat{\boldsymbol{B}}_{hx}^{-1} - B_{hx}^{-1})a_{hx} + \sqrt{nh^3}B_{hx}^{-1}(\hat{\boldsymbol{a}}_{hx} - a_{hx}) =\\ O_p\big\{(nh)^{-1/2}\big\} + O_p(h) + B_{hx}^{-1}\sqrt{nh^3}(\hat{\boldsymbol{a}}_{hx} - a_{hx}) = B_{hx}^{-1}\sqrt{nh^3}(\hat{\boldsymbol{a}}_{hx} - a_{hx}) + o_p(1).\end{aligned} \tag{10}$$

Since $\lim_{n\to\infty} B_{hx} = f(x)\Omega_{js}$ by lemma 4, the stated result follows from lemma 5. □

A.3 Estimation of L

Let ψ_n denote the convergence rate of $\max_{s=1,\ldots,S}\big|h^s(\hat{\boldsymbol{\beta}}_s - \beta_s)\big|$, i.e. $n^{-(S+1)/(2S+3)}$, which is $n^{-2/5}$ if $S = 1$. If $S = 2$ then the rate is $n^{-3/7}$.

Call the numerator and denominator in (3) $\hat{\boldsymbol{N}}$ and $\hat{\boldsymbol{D}}$, respectively, and let $N = f(x)$, $D = 1$. Then, our proofs will be based on an expansion of the form

$$\frac{\hat{\boldsymbol{N}}}{\hat{\boldsymbol{D}}} - \frac{N}{D} \simeq \frac{(\hat{\boldsymbol{N}} - N) - (N/D)(\hat{\boldsymbol{D}} - D)}{D} \simeq \{\hat{\boldsymbol{N}} - f(x)\} - f(x)(\hat{\boldsymbol{D}} - 1), \tag{11}$$

where $\simeq$ means that the omitted terms are asymptotically negligible.

The first lemma is concerned with addressing the influence of the $\hat{\boldsymbol{\beta}}$'s from $\hat{\boldsymbol{D}}$.

Lemma 6.

$$\int_{-z}^{1} m_z(t)\Big\{\exp\Big(\sum_{s=1}^{S}\frac{\hat{\boldsymbol{\beta}}_s t^s h^s}{s!}\Big)-\exp\Big(\sum_{s=1}^{S}\frac{\beta_s t^s h^s}{s!}\Big)\Big\}\,\mathrm{d}t=\sum_{s=1}^{S}h^s(\hat{\boldsymbol{\beta}}_s-\beta_s)c_{msz}+o_p(\psi_n), \tag{12}$$

Proof. Expanding $\exp\big(\sum_{s=1}^{S} b_s h^s t^s/s!\big)$ about $[\beta_1 h \quad \cdots \quad \beta_S h^S]$, the left side in (12) equals

$$\int_{-z}^{1} m_z(t)\sum_{s=1}^{S}(\hat{\boldsymbol{\beta}}_s-\beta_s)\frac{t^s h^s}{s!}\exp\Bigg(\frac{\sum_{s=1}^{S}\beta_s t^s h^s}{s!}\Bigg)+o_p(\psi_n)=\sum_{s=1}^{S}(\hat{\boldsymbol{\beta}}_s-\beta_s)h^s\int_{-z}^{1}\frac{m_z(t)t^s}{s!}\frac{f(x+th)}{\exp(\beta_0)}\mathrm{d}t+o_p(\psi_n)$$
$$=\sum_{s=1}^{S}(\hat{\boldsymbol{\beta}}_s-\beta_s)h^s\int_{-z}^{1}\frac{m_z(t)t^s}{s!}\frac{f(x)}{\exp(\beta_0)}\,\mathrm{d}t+o_p(\psi_n)=\sum_{s=1}^{S}(\hat{\boldsymbol{\beta}}_s-\beta_s)h^s c_{msz}+o_p(\psi_n),$$

as asserted. □

A.3.1 Bias

The next few lemmas are concerned with the asymptotic bias.

Lemma 7. *The bias in* $\hat{\boldsymbol{N}}-f(x)$ *is* $f(x)\sum_{s=1}^{S+1}\mathscr{P}_s(\beta_1,\dots,\beta_s)c_{msz}h^s+o(h^{S+1})$, *where* $\mathscr{P}_s$ *is a complete exponential Bell polynomial.*

Proof. We have

$$\mathbb{E}\hat{\boldsymbol{N}}=\frac{1}{h}\int_{x-zh}^{x+h}m_z\Big(\frac{y-x}{h}\Big)f(y)\,\mathrm{d}y=\int_{-z}^{1}m_z(t)f(x+th)\,\mathrm{d}t=f(x)+\sum_{s=1}^{S+1}f^{(s)}(x)c_{msz}h^s+o(h^{S+1})$$
$$=f(x)+f(x)\sum_{s=1}^{S+1}h^s\mathscr{P}_s(\beta_1,\dots,\beta_s)c_{msz}+o(h^{S+1}),$$

by Faà di Bruno's theorem. □

Lemma 8. *Letting* $\beta_s^*=\mathbb{1}(s\le S)\beta_s$, *the bias in* $\hat{\boldsymbol{D}}-1$ *is* $\sum_{s=1}^{S+1}\mathscr{P}_s(\beta_1^*,\dots,\beta_s^*)c_{msz}h^s+\beta_{S+1}h^{S+1}\sum_{s=1}^{S}c_{msz}b_s+$ $o(\psi_n)$.

Proof. The bias has two components: one is due to the estimation of β and the other to the finite polynomial approximation. Consider the latter first. We have

$$\frac{1}{h}\int_{x-zh}^{x+h}m_z\Big(\frac{y-x}{h}\Big)\exp\Big(\sum_{s=1}^{S}\frac{\beta_s(y-x)^s}{s!}\Big)\,\mathrm{d}y-1=\int_{-z}^{1}m_z(t)\exp\Big(\sum_{s=1}^{\infty}\frac{\beta_s^* t^s h^s}{s!}\Big)\,\mathrm{d}t-1=$$

$$\sum_{s=1}^{S+1}\mathscr{P}_s(\beta_1^*,\dots,\beta_s^*)c_{msz}h^s + o(h^{S+1}).$$

For the bias due to the estimation of β note that by lemma 6 and theorem 1 this bias is equal to $\sum_{s=1}^{S} h^s c_{msz}(\beta_{S+1}h^{S+1-s}b_s) + o(\psi_n) = \beta_{S+1}h^{S+1}\sum_{s=1}^{S} c_{msz}b_s + o(\psi_n)$. □

Lemma 9. *The bias in* $\hat{\boldsymbol{N}} - f(x)\hat{\boldsymbol{D}}$ *is* $f(x)\beta_{S+1}h^{S+1}\bigl(c_{m,S+1,z} - \sum_{s=1}^{S} c_{msz}b_s\bigr) + o(\psi_n)$.

Proof. By lemmas 7 and 8, the desired bias is

$$\begin{aligned} f(x)\sum_{s=1}^{S+1}\mathscr{P}_s(\beta_1,\dots,\beta_s)c_{msz}h^s - f(x)\Biggl(\sum_{s=1}^{S+1}\mathscr{P}_s(\beta_1^*,\dots,\beta_s^*)c_{msz}h^s + \beta_{S+1}h^{S+1}\sum_{s=1}^{S}c_{msz}b_s\Biggr) + o(\psi_n) \\ = f(x)\beta_{S+1}h^{S+1}\Biggl(c_{m,S+1,z} - \sum_{s=1}^{S}c_{msz}b_s\Biggr) + o(\psi_n), \end{aligned}$$

as claimed. □

Lemma 10. $\hat{\boldsymbol{D}} - 1 = o_p(1)$.

Proof. Follows from lemmas 6 and 8 and theorem 1. □

A.3.2 Distribution

Lemma 11. $\sqrt{nh}\bigl\{\hat{\boldsymbol{N}} - f(x)\hat{\boldsymbol{D}}\bigr\} \xrightarrow{d} N(\mathscr{B},\mathscr{V})$.

Proof. The bias result follows from lemma 9 and the definition of Ξ. For asymptotic normality and the variance formula, note that

$$\hat{\boldsymbol{N}} - \mathbb{E}\hat{\boldsymbol{N}} = \frac{1}{nh}\sum_{i=1}^{n}\Biggl\{m_z\Bigl(\frac{\boldsymbol{x}_i - x}{h}\Bigr) - \mathbb{E}m_z\Bigl(\frac{\boldsymbol{x}_1 - x}{h}\Bigr).\Biggr\}$$

Note further that by lemma 6, the asymptotic distribution of $\hat{\boldsymbol{D}}$ net of bias is governed by $\sum_{s=1}^{S} h^s(\hat{\boldsymbol{\beta}}_s - \tilde{\beta}_s)c_{msz}$, which by (9) and (10) and lemma 4 is

$$hc_{mz}^{\intercal}\Lambda_h(\hat{\boldsymbol{\beta}} - \tilde{\beta}) = \frac{c_{mz}^{\intercal}\Omega^{-1}}{f(x)}\frac{1}{nh}\sum_{i=1}^{n}\Biggl\{g'\Bigl(\frac{\boldsymbol{x}_i - x}{h}\Bigr) - \mathbb{E}g'\Bigl(\frac{\boldsymbol{x}_1 - x}{h}\Bigr)\Biggr\} + o_p(\psi_n).$$

Thus,

$$\sqrt{nh}\{\hat{\boldsymbol{N}} - f(x)\hat{\boldsymbol{D}} - \mathbb{E}(\cdot)\} = \frac{1}{\sqrt{nh}}\sum_{i=1}^{n}\Biggl\{m_z\Bigl(\frac{\boldsymbol{x}_i - x}{h}\Bigr) - c_{mz}^{\intercal}\Omega^{-1}g'\Bigl(\frac{\boldsymbol{x}_i - x}{h}\Bigr) - \mathbb{E}(\cdot)\Biggr\} + o_p(1)$$

$$= \frac{1}{\sqrt{nh}} \sum_{i=1}^{n} \left\{ \omega_z\left(\frac{\boldsymbol{x}_i - x}{h}\right) - \mathbb{E}\omega_z\left(\frac{\boldsymbol{x}_1 - x}{h}\right) \right\} + o_p(1),$$

which has the stated limit distribution by e.g. Eicker (1966). □

Lemma 12. *The omitted terms in* (11) *are* $o_p(\psi_n)$.

Proof. Write $\hat{\boldsymbol N}/\hat{\boldsymbol D} - f(x)/1 = \{\hat{\boldsymbol N} - f(x)\hat{\boldsymbol D}\} + \{\hat{\boldsymbol N} - f(x)\hat{\boldsymbol D}\}(1/\hat{\boldsymbol D} - 1) \simeq \hat{\boldsymbol N} - f(x)\hat{\boldsymbol D}$. by lemma 10. Apply lemmas 9 and 11. □

Proof of theorem 2. We show the result for $\hat{\boldsymbol f}$ where the result for $\hat{\boldsymbol L}$ follows from the delta method. By lemma 12 we only need to consider $\hat{\boldsymbol N} - f(x)\hat{\boldsymbol D}$. Apply lemmas 9 and 11. □

A.4 Estimation of $f(x)$ and $f'(x)$ when $f(x) = 0$

Lemma 13. *If* $S = 1$, *the rates in the following table apply to the variance of* $\hat{\boldsymbol f}_m$, $\hat{\boldsymbol \ell}_{hx}$, *and* $\hat{\boldsymbol R}_{hx}$.

condition	$\mathbb{V}\hat{\boldsymbol f}_m$, $\mathbb{V}\hat{\boldsymbol R}_{hx}$	$\mathbb{V}\hat{\ell}_{hx}$
$0 < f(x)$	$\frac{1}{\sqrt{nh}}$	$\frac{1}{\sqrt{nh^3}}$
$f(x) = 0 < f'(x)$	$\frac{1}{\sqrt{n}}$	$\frac{1}{\sqrt{nh^2}}$
$f(x) = f'(x) = 0 < f''(x)$	$\sqrt{\frac{h}{n}}$	$\frac{1}{\sqrt{nh}}$

Proof. We show the rates for $\hat{\ell}_{hx}$ where the other rates obtain similarly. By a standard kernel variance expansion, we get

$$\mathbb{V}\hat{\boldsymbol\ell}_{hx}(x) \simeq \frac{1}{nh^4} \int_{x-zh}^{x+h} g'\left(\frac{t-x}{h}\right)^2 f(t)\,\mathrm{d}t \simeq \frac{1}{nh^3} \int_{-z}^{1} g'(t)^2 f(x+th)\,\mathrm{d}t$$
$$= f(x)O\left(\frac{1}{nh^3}\right) + f'(x)O\left(\frac{1}{nh^2}\right)\mathbb{1}(z<1) + f''(x)O\left(\frac{1}{nh}\right).$$

The stated results then follow by taking square roots of the relevant rate on the right hand side in the last displayed equation. □

Lemma 14. *If* $S = 1$, $-\Omega_{11}(z) = c_{m0z} = 1$ *and* $\Omega_{12}(1) = c_{m11} = 0$, *the dominant terms in* $\mathbb{E}\hat{R}_{hx} - f(x)$ *and* $\mathbb{E}\hat{\ell}_{hx} - f'(x)$ *are contained in the following table.*

condition	$\mathbb{E}\hat{f}_m - f(x)$	$\mathbb{E}\hat{R}_{hx} - f(x)$	$\mathbb{E}\hat{\ell}_{hx} - f'(x)$
$0 < f(x),\ x = zh$	$hc_{mz1}f'(0)$	$-h\Omega_{12}f'(0)$	$-h\Omega_{12}f''(0)$
$0 < f(x),\ 0 < x$	$h^2c_{m12}f''(x)$	$-h^2\Omega_{13}f''(x)$	$o(h)$
$f(x) = 0 < f'(x)$	$hc_{mz1}f'(0)$	$-h\Omega_{12}f'(x)$	$-h\Omega_{12}f''(x)$
$f(x) = f'(x) = 0 < f''(x)$	$h^2c_{mz2}f''(x)$	$-h^2\Omega_{13}f''(x)$	$o(h)$

Proof. These follow from standard kernel bias expansions. □

Proof of theorem 3. Because $\Omega_{11} \neq 0$ and g can be multiplied by any constant without affecting the estimator, we assume without loss of generality that $\Omega_{11} = -1$. Next, we observe that

$$\hat{\boldsymbol{\beta}}_1 \simeq \frac{-\frac{1}{nh^2}\sum_i g'\left(\frac{\boldsymbol{x}_i - x}{h}\right) + \mathbb{E}\frac{1}{nh^2}\sum_i g'\left(\frac{\boldsymbol{x}_i - x}{h}\right) + f'(x) - h\Omega_{12}f''(x)}{\frac{1}{nh}\sum_i g\left(\frac{\boldsymbol{x}_i - x}{h}\right) - \mathbb{E}\frac{1}{nh}\sum_i g\left(\frac{\boldsymbol{x}_i - x}{h}\right) + f(x) - h\Omega_{12}f'(x) - h^2\Omega_{13}f''(x)}. \tag{13}$$

Hence, $h\hat{\boldsymbol{\beta}}_1 = o_p(1)$ by lemmas 13 and 14 when $f(x) = f'(x) = 0$ and g is symmetric so that $\Omega_{12} = 0$. Thus,

$$\hat{\boldsymbol{f}}(x) - f(x) = \hat{\boldsymbol{f}}(x) = \frac{\frac{1}{nh}\sum_i m\left(\frac{\boldsymbol{x}_i - x}{h}\right) - \mathbb{E}\frac{1}{nh}\sum_i m\left(\frac{\boldsymbol{x}_i - x}{h}\right) + h^2 c_{m12}f''(x) + o(h^2)}{1 + o_p(1)},$$

which implies the stated result for the density estimator since $\frac{1}{nh}\sum_i m\left(\frac{\boldsymbol{x}_i - x}{h}\right) - \mathbb{E}\frac{1}{nh}\sum_i m\left(\frac{\boldsymbol{x}_i - x}{h}\right) = O_p(n^{-3/5})$ by lemma 13.

The result involving $\hat{\boldsymbol{f}}\hat{\boldsymbol{\beta}}_1$ follows from the fact that $\hat{\boldsymbol{f}} = O_p(h^2)$ and $h\hat{\boldsymbol{\beta}}_1 = o_p(1)$. □

Proof of theorem 4. Because $\Omega_{11} \neq 0$ and g can be multiplied by any constant without affecting the estimator, we assume without loss of generality that $\Omega_{11} = -1$.

If $f'(0) > 0$, $\Omega_{12} = 0$, and $x = 0$, then $h\hat{\boldsymbol{\beta}}_1 \simeq -f'(0)/\{h\Omega_{13}f''(0)\}$ diverges by (13).

If $\Omega_{12} \neq 0$, by (13) we have $h\hat{\boldsymbol{\beta}}_1 \simeq \frac{\Omega_{11}}{\Omega_{12}} = \frac{-1}{\Omega_{12}}$ if $f'(0) > 0$ or $h\hat{\boldsymbol{\beta}}_1 \simeq \frac{\Omega_{12}}{\Omega_{13}}$ if $f'(0) = 0$. Hence, if $\Omega_{12} \neq 0$ then

$$h\hat{\boldsymbol{\beta}}_1^* = h\hat{\boldsymbol{\beta}}_1 \max\left(1, h^q A\hat{\boldsymbol{\beta}}_1\right) \simeq Ah^{q-1}(h\hat{\boldsymbol{\beta}}_1)^2 \simeq Ah^{q-1}\frac{\mathbb{1}\{f'(0) > 0\} + \mathbb{1}\{f'(0) = 0\}\Omega_{12}^2}{\mathbb{1}\{f'(0) > 0\}\Omega_{12}^2 + \mathbb{1}(f'(0) = 0)\Omega_{13}^2}$$

diverges to ∞.

In either case, the denominator in the definition of $\hat{\boldsymbol{f}}^*$ explodes. Thus, although the numerator is $O_p(n^{1/5})$ because c_{mz1} is generically nonzero, $\hat{\boldsymbol{f}}^*$ converges exponentially toward zero.

If one considers the case $x = zh$ then much the same derivations arise except that one has to take additional expansions. For instance, in the $\Omega_{12} \neq 0$ and $f'(0) = 0$ case, we would have

$$h\hat{\beta}_1 \simeq h\frac{f'(zh) - h\Omega_{12}f''(zh)}{-h\Omega_{12}f'(zh) - h^2\Omega_{13}f''(zh)} \simeq \frac{\Omega_{12} - z}{\Omega_{12}z + \Omega_{13}}. \qquad \square$$

B Existing methods

This appendix presents the key equations in a standardized notation for related estimators.

B.1 Local polynomial regression

The estimator analyzed in Cattaneo et al. (2019) is a local polynomial regression of $F_n(x_i)$ on x_i. That is, they estimate a smooth distribution and its derivatives by solving

$$\min_{b \in \mathbb{R}^{S+1}} \sum_i \left[F_n(x_i) - p_{S+1}(x_i - x)^\intercal b\right]^2 k\left(\frac{x_i - x}{h}\right)$$

where $p_{S+1}(u) = [1 \; u \; \ldots \; u^{S+1}]^\intercal$. The first–order conditions yield the linear system of equations

$$\sum_i k\left(\frac{x_i - x}{h}\right)(x_i - x)^j F_n(x_i) = \sum_i k\left(\frac{x_i - x}{h}\right)(x_i - x)^j p_{S+1}(x_i - x)^\intercal b \qquad j = 0,\, 1, \ldots,\, S+1\,. \tag{14}$$

Meanwhile, Lejeune and Sarda (1992) use a closely related local polynomial approximation to the function F_n. They solve

$$\min_{b \in \mathbb{R}^{S+1}} \int_0^{\mathcal{U}} \left[F_n(y) - p_{S+1}(y - x)^\intercal b\right]^2 k\left(\frac{y - x}{h}\right) \mathrm{d}y\,,$$

which yields the linear system

$$\int_0^{\mathcal{U}} k\left(\frac{y - x}{h}\right)(y - x)^j F_n(y)\,\mathrm{d}y = \int_0^{\mathcal{U}} k\left(\frac{y - x}{h}\right)(y - x)^j p_{S+1}(y - x)^\intercal b\,\mathrm{d}y \qquad j = 0,\, 1, \ldots,\, S+1\,. \tag{15}$$

The difference between the estimated polynomial coefficients using (14) and (15) is $O_p(1/n)$, which derives from the difference between $n\int_{x_{(i-1)}}^{x_{(i)}} k\left(\frac{y-x}{h}\right)(y - x)^j\,\mathrm{d}y$ and $k\left(\frac{x_{(i)}-x}{h}\right)(x_{(i)} - x)^j$ for $j = 0, \ldots, 2(S+1)$, where the subscripts denote the i–th order statistic.

B.2 Local likelihood density estimation

The estimator in Loader (1996) solves

$$\max_{b \in \mathbb{R}^{S+1}} \sum_i k\left(\frac{x_i - x}{h}\right) p_{S+1}(x_i - x)^\intercal b - n \int_0^{\mathcal{U}} k\left(\frac{y - x}{h}\right) \exp\{p_{S+1}(y - x)^\intercal b\} \, \mathrm{d}y \, .$$

The first term in the objective represents the locally weighted log–likelihood of the data drawn from a log–polynomial density, while the second term comes from the constraint that the density must integrate to one. The first–order conditions for this problem yield the nonlinear system

$$\frac{1}{n} \sum_i k\left(\frac{x_i - x}{h}\right) (x_i - x)^j = \int_0^{\mathcal{U}} k\left(\frac{x_i - x}{h}\right) (y - x)^j \exp\{p_{S+1}(y - x)^\intercal b\} \, \mathrm{d}y \qquad j = 0, 1, \ldots, S + 1 \, .$$

For $j = 0$, the above equation resembles (3), while for $j > 0$, the left side is a local sample moment, and the right side is the local population moment implied by the polynomial approximation to the log–density.

B.3 Generalized reflection method

The generalized reflection method uses a transformation of the data such that the transformed data takes support on the entire real line, has a density that coincides with the density of the original data on $[0, \mathcal{U})$, and has a square–integrable second derivative. A transformation that achieves these aims is the mapping $\{x_i\} \mapsto \{x_i, -\rho(x_i)\}$ where the function ρ is given by $\rho(x) = x + \beta_1 x^2 + A\beta_1^2 x^3$ for some $A > 1/3$. The constant A ensures the transformation is strictly increasing; Karunamuni and Alberts (2005) and Karunamuni and Zhang (2008) use $A = 0.55$ in their simulations. The density estimator is given by

$$\frac{1}{nh} \sum_i k\left(\frac{x - x_i}{h}\right) + k\left(\frac{x + \rho_n(x_i)}{h}\right)$$

where ρ_n is an estimator of the function ρ. The authors suggest estimating ρ by substituting β_1 with

$$\frac{\log f_n(h) - \log f_n(0)}{h} \, ,$$

where

$$f_n(h) = \frac{1}{n^2} + \frac{1}{nh} \sum_i k\left(\frac{x - x_i}{h}\right) , \qquad f_n(0) = \max\left\{ \frac{1}{n^2}, \frac{1}{nh_0} \sum_i k_0\left(\frac{-x_i}{h}\right) \right\} ,$$

h_0 is a bandwidth and k_0 is a boundary kernel that satisfies $\int_{-1}^0 k_0(t) \, \mathrm{d}t = 1$, $\int_{-1}^0 k_0(t) t \, \mathrm{d}t = 0$, and $\int_{-1}^0 k_0(t) t^2 \, \mathrm{d}t \neq 0$. We note, however, that the details of the finite–difference approximation to the derivative of $\log f$ can be modified as long as the estimate of β_1 converges at the rate of $n^{-1/5}$, which is the optimal rate of convergence when f is assumed to have two continuous derivatives.

References

Bell, E. T. (1927). Partition polynomials. *Annals of Mathematics*, pages 38–46.

Cattaneo, M. D., Jansson, M., and Ma, X. (2019). Simple local polynomial density estimators. *Journal of the American Statistical Association*, pages 1–7.

Cheng, M.-Y., Fan, J., and Marron, J. S. (1997). On automatic boundary corrections. *Annals of Statistics*, 25(4):1691–1708.

Eicker, F. (1966). A multivariate central limit theorem for random linear vector forms. *Annals of Mathematical Statistics*, pages 1825–1828.

Gasser, T. and Müller, H.-G. (1979). Kernel estimation of regression functions. In *Smoothing techniques for curve estimation*, pages 23–68. Springer.

Guerre, E., Perrigne, I., and Vuong, Q. (2000). Optimal nonparametric estimation of first–price auctions. *Econometrica*, 68(3):525–574.

Hickman, B. R. and Hubbard, T. P. (2015). Replacing sample trimming with boundary correction in nonparametric estimation of first-price auctions. *Journal of Applied Econometrics*, 30(5):739–762.

Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189.

Hjort, N. L. and Jones, M. C. (1996). Locally nonparametric density estimation. *Annals of Statistics*, 24(4):1619–1647.

Jones, M. and Foster, P. (1996). A simple nonnegative boundary correction method for kernel density estimation. *Statistica Sinica*, pages 1005–1013.

Karunamuni, R. J. and Alberts, T. (2005). On boundary correction in kernel density estimation. *Statistical Methodology*, 2(3):191–212.

Karunamuni, R. J. and Zhang, S. (2008). Some improvements on a boundary corrected kernel density estimator. *Statistics and Probability Letters*, 78(5):499–507.

Klein, R. W. and Spady, R. H. (1993). An efficient semiparametric estimator for binary response models. *Econometrica*, pages 387–421.

Lejeune, M. and Sarda, P. (1992). Smooth Estimators of Distribution and Density Functions. *Computational Statistics & Data Analysis*, 14:457–471.

Lewbel, A. and Schennach, S. M. (2007). A simple ordered data estimator for inverse density weighted expectations. *Journal of Econometrics*, 136(1):189–211.

Loader, C. R. (1996). Local likelihood density estimation. *Annals of Statistics*, 24(4):1602–1618.

Pinkse, J. and Schurter, K. (2019). Estimation of auction models with shape restrictions. Pennsylvania State University working paper.

Zhang, S. and Karunamuni, R. J. (1998). On kernel density estimation near endpoints. *Journal of Statistical Planning and Inference*, 70:301–316.