Electronic Journal of Statistics
Vol. 10 (2016) 736–760
ISSN: 1935-7524
DOI: 10.1214/16-EJS1121

Estimating the density of a conditional expectation

Samuel G. Steckley, Shane G. Henderson, David Ruppert

Cornell University

Ran Yang, Daniel W. Apley, and Jeremy Staum

Northwestern University

Abstract: In this paper, we analyze methods for estimating the density of a conditional expectation. We compare an estimator based on a straightforward application of kernel density estimation to a bias-corrected estimator that we propose. We prove convergence results for these estimators and show that the bias-corrected estimator has a superior rate of convergence. In a simulated test case, we show that the bias-corrected estimator performs better in a practical example with a realistic sample size.

Keywords and phrases: Density deconvolution, kernel density estimation, bias-correction, nested simulation, repeated measurements.

Received September 2015.

1. Introduction

This paper proposes and analyzes improved methods for estimating the density of a conditional expectation.The following example illustrates and motivates the problem of estimating the density of a conditional expectation.A pharmaceutical production process produces batches of an ingredient. The true sodium content of the ingredient randomly varies across batches.It is desired to learn the density of the true sodium content.When a sample taken from a batch is subjected to mass spectrometry, it yields an unbiased, noisy measurement of the sodium content in this batch.Due to this noise, the sodium content is measured separately for multiple (homogenized) samples taken from each batch. The goal here is to learn the density of the true sodium content, based on the noisy spectrometry measurements.

We consider a general framework into which this example fits. Let Z be an unobserved random variable. In the sodium measurement example, Z is the true sodium content in a batch. Let X be an observed random variable that has probabilistic dependence with Z. In the sodium measurement example, X is a measurement of sodium content. Let $Y = E(X|Z)$ be the conditional expectation of X given Z. We are seeking to estimate the density of Y based on observations of X. In the sodium measurement example, $Y = Z$, the true sodium content in a batch. Unlike in Berkson's error model that $Z \equiv E[X|Z]$, in general cases of our model, Y and Z can be different. For example, Z could be

a student's standardized test scores and Y could be the conditional expectation of X, the student's income at age 30.

Suppose that m observations of X are available in n samples, each sample being associated with a single value of the random variable Z, as in the following statistical model:

$$X_{ij} = Y_i + U_{ij}, \qquad i = 1, \ldots, n, \quad j = 1, \ldots, m, \tag{1}$$

$$Y_i = \mathrm{E}(X|Z = Z_i), \qquad i = 1, \ldots, n. \tag{2}$$

Here X_{ij} is the jth observation of X in the ith sample, $U_{ij} = X_{ij} - Y_i$ is its observation error relative to Y_i, its conditional expectation given Z_i, and Z_i is the value of the random variable Z associated with the ith sample of observations, $X_{i1}, \ldots, X_{im}$. Our assumptions are:

1. There are *unobserved* i.i.d. random variables $Z_1, \ldots, Z_n$ from the distribution of Z.
2. For each $i = 1, \ldots, n$, $X_{i1}, \ldots, X_{im}$ is the *observed* i.i.d. sample from the conditional distribution of X given $Z = Z_i$.
3. The observation error $U = X - Y$ has mean zero and is uncorrelated with Y, but it need *not* be independent of Y or Z.
4. The normality of U is used to derive the convergence rate. However, for the proposed methods to be effective, the distribution of U need *not* be normal or even known.
5. For any value of Z, the conditional expectation $Y = \mathrm{E}(X|Z)$ exists and is finite. In general, Z need *not* equal to Y, although Z does equal to Y in some interesting examples including the sodium example.
6. The conditional expectation Y has a density with respect to Lebesgue measure.

In the sodium measurement example, the variance of the observation error $U = X - Y$ is larger for larger values of the sodium content Y. Significant heteroscedasticity also appears in stochastic simulation input uncertainty analysis [10], another example that has motivated this work. Substantial bias in density estimation can result from ignoring heteroscedasticity [23].

We estimate the density of conditional expectation using kernel smoothing ([29] and [17]). [29] gave an introduction and review of the subject of kernel smoothing, while [17] proposed to use this method for a nested data structure. The standard setting for kernel smoothing for density estimation is as follows: Suppose $(Y_i : 1 \leq i \leq n)$ is a sequence of independent random variables with density g. The standard kernel smoothing estimator is

$$\hat{g}(x; h) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - Y_i}{h}\right), \tag{3}$$

where the kernel K is typically chosen to be a unimodal probability density function that is symmetric about zero, and h is the bandwidth. The estimator of (3) immediately suggests that we can estimate $f(x)$, the density of $Y = \mathrm{E}(X|Z)$

evaluated at x, from model (1) via

$$\hat{f}(x;m,n,h) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x-\bar{X}_m(Z_i)}{h}\right), \tag{4}$$

with $\bar{X}_m(Z_i) = \sum_{j=1}^{m} X_{ij}/m$. We call this the "standard estimator", and it was analyzed by [17] and [25].

The cost of the experiment from which this estimator is generated is $\delta_1 n + \delta_2 nm$, where δ_1 and δ_2 are the average cost used to generate Z_i and X_{ij} given Z_i, respectively. [17] and [25] gave results on the convergence rate of the mean squared error of the standard kernel estimator as the experiment cost goes to infinity, including an analysis of asymptotically optimal choices of m, n, and h.

Our paper makes three contributions to estimation of the density of a conditional expectation. First, we extend the results of [25], who analyzed only the case in which Z is univariate and the conditional expectation $Y = \mathrm{E}(X|Z)$ is monotone in Z. In this paper, results are presented that apply more broadly to multivariate Z. Second, we propose and analyze a bias-corrected estimator, which has a better rate of convergence than the standard estimator. Third, we create practical methods for selecting the sample sizes n and m in the experiment design and the bandwidth h in the experiment analysis.

Before addressing these contributions in detail, we explain why this paper is based on kernel smoothing instead of kernel deconvolution [29]. One reason is that, when using kernel smoothing, it is easier and more straightforward to derive expressions for asymptotic mean integrated squared error. These expressions provide the foundation for showing that the bias-corrected estimator has a better rate of convergence than the standard estimator, and for the methods to select the sample sizes and the bandwidth. Another reason is that deconvolution is, in a sense, not necessary. In our asymptotic setting in which $m \to \infty$, the standard estimator (4) based on kernel smoothing is consistent. Furthermore, our bias-corrected estimator is an alternative to deconvolution, in the sense that it does something different to reduce the bias caused by observation error. A third reason is that kernel deconvolution can be applied to the problem of estimating the density of a conditional expectation only under more restrictive assumptions than we require when using kernel smoothing.

Most kernel deconvolution methods are based on the assumption that the measurement error U is independent of Y, i.e., the measurement errors U_{ij} have a common distribution for all $i = 1,\ldots,n$ and $j = 1,\ldots,m$ [2]. [17] propose a kernel smoothing estimator for the nested data structure, which is very similar with our standard estimator, yet this paper works with measurement errors U_{ij} that have a common distribution, which we do not assume. [3] have a kernel deconvolution method that allows for errors to have different distributions, but all the error distributions must be known. [4] discuss a kernel deconvolution method for the case of a single unknown error distribution. It seems that it would not be practical to identify the error distributions when they differ and are unknown, unless one could make some very strong assumptions. Other deconvolution methods that allow for heteroscedastic errors assume the errors are normal [23, 16]. The methods we consider do not require the errors to be

normal; we merely perform asymptotic analysis using a normal approximation to the distribution of an average of many errors, which can be justified by the central limit theorem.

We analyze the convergence rate of our proposed estimators in the measure of mean squared error (mse) and also mean integrated squared error (mise). It is well known that if the density function g is continuous, the standard estimator (3) is consistent in quadratic mean. That is to say, $\text{mse}(\hat{g}(x;h))$ converges to zero for all $x \in \mathbb{R}$. It is also well known [18] that if g is twice continuously differentiable such that g'' is bounded and square integrable, mise converges to zero at an optimal rate of $n^{-4/5}$ where n here is the sample size. We will show that our standard estimator (4) is consistent in quadratic mean and mise converges to zero at an optimal rate of $c^{-4/7}$ where c is the experiment budget. This is the same rate that [25] computed for the case in which Z is univariate. We also discuss the convergence of the bias-corrected version of our estimator and show that mse optimally converges to zero at a rate of $c^{-8/11}$. These optimal rates of convergence depend on asymptotically optimal choices of the sample sizes n and m and the bandwidth h.

These questions of optimal rates of convergence and the allocation of an experiment budget c to sample sizes m and n have also been addressed in related but distinct settings. [14], [15], [9], and [1] studied estimation of the distribution function of a conditional expectation. We believe that density estimation is also important because a density is more easily interpreted visually than a distribution function. Estimation of the distribution is rather different from estimation of the density, because techniques such as kernel smoothing are not necessary. [28] studied estimation of the variance of a conditional expectation.

Development of a bias-corrected estimator is another primary contribution of the present paper. [11] review bias-correction in kernel density estimation. The bias they address is caused by the kernel smoothing, while we attempt to address the bias due to both kernel smoothing and noisy observations. We implement a method similar to jackknife bias-correction [6].

Kernel smoothing methods require the selection of the bandwidth. The performance of kernel smoothing is quite dependent on bandwidth selection, which has received much attention [29]. [21] reviews some modern bandwidth selection methods in the context of local polynomial regression, a type of kernel regression. One such method is the empirical-bias bandwidth selection (EBBS) developed by [20]. In our setting, we must choose the bandwidth, but given an experiment budget c, we must also choose the number n of samples of Z and the number m of samples of X given each Z. Applying the ideas from EBBS, we develop a data-driven method to select each of these parameters.

The rest of the paper is organized as follows. In Section 2 we formulate estimators for the density of the conditional expectation and present convergence results. In Section 3, we develop a data-based selection method for the bandwidth and the sample sizes n and m, based on EBBS. We discuss the reasons for choosing this method and present the algorithm. In Section 4, we then explore the performance of the estimators for a simulated test case and for the sodium measurement example.

2. Estimating the density of the conditional expectation and convergence results

First consider the standard estimator (4) where $\bar{X}_m(Z_i)$ is considered as an observation of $\mathrm{E}(X|Z_i)$ with measurement error. This standard estimator is motivated by the standard estimator for standard kernel density estimation. The measurement error results in additional smoothing beyond that comes from kernel smoothing. A similar double smoothing was noted in [22]. He considered the problem of local polynomial regression in which the covariates are measured with error. The double smoothing increases the bias of our estimator given in (4) as compared with the estimator (3). Specifically, the additional smoothing results in an additional leading term in the bias expansion. This creates an additional leading term in the mse and mise expansions given in Theorems 2 and 3 in Section 2.1, where we present convergence results and proofs for the standard estimator. In Section 2.2 we consider a bias-corrected version. We derive asymptotic expressions of the mse for the estimators and establish an improvement in the optimal rate of convergence.

2.1. Convergence results: Standard estimator

In this section we study the error in the estimator $\hat{f}(x; m(c), n(c), h)$ as the experiment budget c goes to infinity. For any fixed c, the number of internal samples m, and the number of external samples n, must be chosen so that the total cost is c. Note that $m(c)$ and $n(c)$ are thus functions of the experiment budget c. We assume that $m(c) \to \infty$ as $c \to \infty$ so that $\bar{X}_{m(c)}(z_0) \to \mathrm{E}(X|z_0)$ almost surely. Assuming $m(c) \to \infty$, $\delta_1 n(c) + \delta_2 n(c)m(c) \approx \delta_2 n(c)m(c)$. One can assume, by a selection of units, that $\delta_2 = 1$ without loss of generality. Then $m(c)$ and $n(c)$ must be chosen to satisfy the asymptotic relationship $m(c)n(c)/c \to 1$ as $c \to \infty$.

The bandwidth $h = h(c)$ is also a function of c. To keep the notation less cumbersome, the dependence of m, n, and h on c will be suppressed in the calculations.

We will present results concerning the convergence of the estimator as the experiment budget c tends to ∞. We consider the following two error criteria. For all $x \in \mathbb{R}$, define the mean squared error (mse) of the estimator evaluated at x as

$$\mathrm{mse}(\hat{f}(x; m, n, h)) = \mathrm{E}\big(\hat{f}(x; m, n, h) - f(x)\big)^2.$$

Define the mean integrated squared error (mise) of the estimator as

$$\mathrm{mise}(\hat{f}(\cdot; m, n, h)) = \mathrm{E} \int \big(\hat{f}(x; m, n, h) - f(x)\big)^2 \, \mathrm{d}x.$$

These error criteria are not without drawbacks (see [5]) but the mathematical simplicity is appealing.

Before stating our results, we consider the distribution of the observations $(\bar{X}_m(Z_i) : 1 \le i \le n)$ and in doing so, we will collect some of the assumptions

needed for the results. Let $N(\alpha_1, \alpha_2)$ denote a normally distributed random variable with mean α_1 and variance α_2. For two random objects X and Y, define the notation $X =_d Y$ to mean X and Y are equal in distribution. Denote $\mu(\cdot) \equiv \mathrm{E}(X|Z = \cdot)$ and $\sigma^2(\cdot) \equiv \mathrm{var}(X|Z = \cdot)$. Throughout this paper we assume the following:

A1. Conditional on $(Z_i : 1 \leq i \leq n)$, $\bar{X}_m(Z_i) =_d N(\mu(Z_i), m^{-1}\sigma^2(Z_i))$ for $i = 1, \ldots, n$ and $(\bar{X}_m(Z_i) : 1 \leq i \leq n)$ are conditionally independent.

This essentially implies that the internal samples $X(Z)$ conditional on Z are unbiased and normally distributed. Of course, if the assumptions of one of the many versions of the central limit theorem hold, then for large m this assumption is approximately true.

We now turn to the distribution of the observations $(\bar{X}_m(Z_i) : 1 \leq i \leq n)$, which are i.i.d. Under Assumption A1,

$$\bar{X}_m(Z_i) =_d Y_i + S_i \frac{1}{m} \sum_{j=1}^{m} U_{ij} \quad \text{for} \quad i = 1, \ldots, n,$$

where

(i) $((Y_1, S_1), \ldots, (Y_n, S_n))$ are i.i.d. with $(Y_i, S_i) =_d (\mu(Z), \sigma(Z))$;
(ii) $(U_{ij} : 1 \leq i \leq n, \quad 1 \leq j \leq m)$ are i.i.d. with $U_{ij} =_d N(0, 1)$.

Let $U_i^m = m^{-1/2} \sum_{j=1}^{m} U_{ij}$ so that for $i = 1, \ldots, n$,

$$\bar{X}_m(Z_i) \quad =_d \quad Y_i + S_i \frac{1}{m} \sum_{j=1}^{m} U_{ij} = Y_i + \frac{S_i}{\sqrt{m}} U_i^m.$$

Note that $U_i^m =_d N(0, 1)$ for $i = 1, \ldots, n$, and $(U_i^m : 1 \leq i \leq n)$ are i.i.d.

Let F_m denote the distribution function of $\bar{X}_m(Z_i)$. Assuming $\mathrm{P}(S = 0) = 0$,

$$F_m(x) \quad = \quad \mathrm{P}\left(Y_i + \frac{S_i}{\sqrt{m}} U_i^m \leq x\right) = \mathrm{P}\left(U_i^m \leq \frac{(x - Y_i)\sqrt{m}}{S_i}\right).$$

The following is also assumed throughout:

A2. For each $y \in \mathbb{R}$ such that $f(y) > 0$, the conditional density with respect to Lebesgue measure of the conditional distribution $\mathrm{P}(\sigma(Z) \in \cdot \ | \mu(Z) = y)$ exists. Denote this density $g(\cdot|y)$.

Since $\sigma(Z)$ and $\mu(Z)$ are random variables we know that the regular conditional distribution $\mathrm{P}(\sigma(Z) \in \cdot \ | \mu(Z) = y)$ exists for all $y \in \mathbb{R}$. This assumption simply requires that for each $y \in \mathbb{R}$ such that $f(y) > 0$, $\mathrm{P}(\sigma(Z) \in \cdot \ | \mu(Z) = y)$ is absolutely continuous with respect to Lebesgue measure.

We believe that when Z is of dimension 2 or greater, there will be many cases in which A2 is satisfied. By assuming A2 in this paper, we focus on the case in which Z is of dimension 2 or greater. For univariate Z, [25] showed results for

mise that are very similar to the ones presented here but for the sake of space, we omit these results and proofs and refer the reader to [25].

Assuming A2,

$$\begin{aligned} F_m(x) &= \mathrm{P}\left(U_i^m \le \frac{(x-Y_i)\sqrt{m}}{S_i}\right) \\ &= \int\int \mathrm{P}\left(U_i^m \le \frac{(x-y)\sqrt{m}}{s}\right) g(s|y)f(y)\,\mathrm{d}s\,\mathrm{d}y, \end{aligned}$$

where $g(\cdot|y)$ can be defined arbitrarily for $y \in \mathbb{R}$ such that $f(y) = 0$. Let Φ and ϕ denote the standard normal cumulative distribution function and density, respectively. In this notation,

$$\begin{aligned} F_m(x) &= \int\int \Phi\left(\frac{(x-y)\sqrt{m}}{s}\right) g(s|y)f(y)\,\mathrm{d}s\,\mathrm{d}y \\ &= \mathrm{E}\left(\Phi\left(\frac{(x-Y)\sqrt{m}}{S}\right)\right). \end{aligned}$$

Assuming we can differentiate the RHS, and interchange the derivative and expectation, we have that the density f_m of the distribution function F_m exists and is given by

$$f_m(x) = \int\int \frac{\sqrt{m}}{s}\quad \phi\left(\frac{(x-y)\sqrt{m}}{s}\right) g(s|y)f(y)\ \mathrm{d}s\,\mathrm{d}y. \tag{5}$$

A sufficient condition for the interchange is

A3. $\iint (1/s)\, g(s|y)f(y)\ \mathrm{d}s\,\mathrm{d}y < \infty$,

which comes from a result given by [12] and [13]; see also [8], and Lemma 1 of [24] for the application in the present context. Returning to the density of the observations $\bar{X}_m(Z)$ given in (5), the change of variable $z = (x-y)\sqrt{m}$ gives

$$f_m(x) = \int\int \frac{1}{s}\,\phi\left(\frac{z}{s}\right) g(s|x - \frac{z}{\sqrt{m}})f(x - \frac{z}{\sqrt{m}})\,\mathrm{d}s\,\mathrm{d}z.$$

Suppose $f(\cdot)$ is continuous. For y such that $f(y) = 0$, suppose that $g(\cdot|y)$ can be defined so that $g(s|\cdot)$ is continuous for all $s \in \mathbb{R}$. We assume the following:

A4. For almost all $y \in \mathbb{R}$, $g(\cdot|y)$ is nonnegative;
A5. For almost all $y \in \mathbb{R}$, $g(s|y) = 0$ for $s < 0$.

The Assumptions A4 and A5 are certainly true for y such that $f(y) > 0$ since in that case $g(\cdot|y)$ is a density for a nonnegative random variable. Under A4, the order of integration can be changed so that

$$f_m(x) = \int\int \frac{1}{s}\,\phi\left(\frac{z}{s}\right) g(s|x - \frac{z}{\sqrt{m}})f(x - \frac{z}{\sqrt{m}})\,\mathrm{d}z\,\mathrm{d}s. \tag{6}$$

It will be useful to think in terms of the joint density of $\mu(Z)$ and $\sigma(Z)$. Let us denote this density by α. Of course

$$\alpha(x,s) = g(s|x)f(x). \tag{7}$$

Define for nonnegative integer k,

$$\alpha^{(k+1)}(x,s) = \frac{d}{dy}\alpha^{(k)}(y,s)\big|_{y=x}, \tag{8}$$

where $\alpha^{(0)}(x,s) = \alpha(x,s)$. Also define for nonnegative integer k,

$$g^{(k+1)}(s|x) = \frac{d}{dy}g^{(k)}(s|y)\big|_{y=x},$$

where $g^0(s|x) = g(s|x)$.

For ease of notation we define the following set of Assumptions parameterized by nonnegative integer k as A6(k).

1. $f(\cdot)$ is k times continuously differentiable;
2. for all $s \in \mathbb{R}$, $g(s|\cdot)$ is k times continuously differentiable;
3. $\exists B_f > 0$ such that $|f^{(j)}(\cdot)| \le B_f$ for $j = 0, 1, \ldots, k$;
4. $\exists B_g > 0$ such that $|g^{(j)}(\cdot|\cdot)| \le B_g$ for $j = 0, 1, \ldots, k$;
5. $\exists B_S > 0$ such that $\sigma^2(\cdot) \le B_S$ everywhere.

Note that $f^{(0)}$ and $g^{(0)}$ are simply f and g, respectively, and when $k = 0$, Assumptions 1 and 2 imply that $f(\cdot)$ and $g(s|\cdot)$ are continuous.

The following theorem gives sufficient conditions for the consistency in quadratic mean for the estimator formulated in (4).

Theorem 1. Assume A1–A5, and A6(0). Also assume that

1. K is a bounded probability density;
2. $m(c) \to \infty$, $h(c) \to 0$, and $n(c)h(c) \to \infty$, as $c \to \infty$.

Then for all $x \in \mathbb{R}$,

$$\lim_{c\to\infty} \operatorname{mse}(\hat{f}(x; m, n, h)) = 0.$$

A proof is given in the Appendix. (Appendix is presented as a supplementary materials [26])

We now turn to the asymptotic expressions of mse and mise. More restrictive assumptions are needed to compute these asymptotic expansions. For one thing, it is assumed that the function $f(\cdot)$ and the set of functions $\{g(s|\cdot) : s \in \mathbb{R}\}$ are four times continuously differentiable.

For sequences of real numbers a_n and b_n, we say that

$$a_n = o(b_n) \text{ as } n \to \infty \text{ iff } \lim_{n\to\infty} a_n/b_n = 0.$$

For sequences of real numbers a_n and b_n, we say that

$$a_n = O(b_n) \text{ as } n \to \infty \text{ iff } \exists C \text{ s.t. } |a_n| \le C|b_n| \text{ for } n \text{ sufficiently large.}$$

Theorem 2. Assume A1–A5, and A6(4). Also assume

1. K is a bounded probability distribution function symmetric about zero with finite second moment;
2. $m(c) \to \infty$, $n(c) \to \infty$, $h(c) \to 0$, and $n(c)h(c) \to \infty$ as $c \to \infty$.

Then

$$\begin{aligned}\operatorname{mse}(\hat{f}(x; m, n, h)) &= \left(h^2 \frac{1}{2} f''(x) \int u^2 K(u)\,\mathrm{d}u + \frac{1}{m}\frac{1}{2}\int s^2 \alpha^{(2)}(x,s)\,\mathrm{d}s\right)^2 \\ &\quad + \frac{1}{nh} f(x) \int K^2(u)\,\mathrm{d}u + o\left(\left(h^2 + \frac{1}{m}\right)^2 + \frac{1}{nh}\right), \quad (9)\end{aligned}$$

where α is defined in (7) and (8).

Theorem 3. Assume A1–A5 and A6(4). Also assume

1. $f''(\cdot)$ is ultimately monotone, meaning that there exists a $B > 0$ such that f'' is monotone on $[B, \infty)$ and monotone on $(-\infty, -B)$;
2. $f^{(k)}(\cdot)$ is integrable for $k = 1, 2, 3, 4$;
3. K is a bounded probability density function symmetric about zero with finite second moment;
4. $m(c) \to \infty$, $n(c) \to \infty$, $h(c) \to 0$, and $n(c)h(c) \to \infty$ as $c \to \infty$.

Then

$$\begin{aligned}&\operatorname{mise}(\hat{f}(\cdot; m, n, h)) \\ &\quad = \int \left(h^2 \frac{1}{2}\left(\int u^2 K(u)\,\mathrm{d}u\right) f''(x) + \frac{1}{m}\frac{1}{2}\int s^2 \alpha^{(2)}(x,s)\,\mathrm{d}s\right)^2 \mathrm{d}x \\ &\qquad + \frac{1}{nh}\int K^2(u)\,\mathrm{d}u + o\left(\left(h^2 + \frac{1}{m}\right)^2 + \frac{1}{nh}\right), \quad (10)\end{aligned}$$

where α is defined in (7) and (8).

Theorem 3 follows from Theorem 2 provided the o term in (9) is integrable. Proofs of Theorems 2 and 3 are presented in the Appendix ([26]).

Compare (10) to the mise for standard kernel density estimation (e.g., [29]),

$$\begin{aligned}\operatorname{mise}(\hat{g}(\cdot; h)) &= \int \left(h^2 \frac{1}{2}\left(\int u^2 K(u)\,\mathrm{d}u\right) g''(x)\right)^2 \mathrm{d}x + \frac{1}{nh}\int K(u)^2\,\mathrm{d}u \\ &\quad + o\left(h^4 + \frac{1}{nh}\right). \quad (11)\end{aligned}$$

It is known that mise can be decomposed into integrated squared bias and integrated variance. We get similar formulas for the standard kernel density estimator $\hat{g}$. Note that the $O(1/nh)$ terms in the mise expansions in (10) and (11) are the same for both estimators. In the proof of Theorem 3 we show that

this term is the leading term for the integrated variance. The remaining leading terms in (10) and (11) are those of the integrated squared bias.

For our estimator $\hat{f}$, the bias itself can be further decomposed. Suppose that the density of an observation $\bar{X}_m(Z)$ exists and is given by $f_m(\cdot)$. Then

$$\text{bias}(\hat{f}(x;m,n,h)) \quad = \quad (\text{E}(\hat{f}(x;m,n,h)) - f_m(x)) + (f_m(x) - f(x)) \quad (12)$$

The first component, $\text{E}(\hat{f}(x;m,n,h)) - f_m(x)$, is the bias due to kernel smoothing, while the second component is the bias due to measurement error. Both the standard kernel density estimator and our estimator are biased due to the kernel smoothing, and the leading term of this bias for both estimators is $O(h^2)$. However, due to measurement error our estimator has an additional bias whose leading term is $O(1/m)$, and this bias also depends on the distribution of the conditional variance function $\sigma^2(\cdot)$ through α.

The asymptotic mise for our estimator $\hat{f}$ is

$$\int\left(h^2\frac{1}{2}\left(\int u^2K(u)\text{d}u\right)f''(x)+\frac{1}{m}\frac{1}{2}\int s^2\alpha^{(2)}(x,s)\text{d}s\right)^2\text{d}x+\frac{1}{nh}\int K^2(u)\text{d}u. \quad (13)$$

By choosing m, n, and h to minimize this asymptotic mise, we can achieve the optimal asymptotic convergence. Define

$$A=\sqrt{\sqrt{\frac{\int\beta_2(x)^2\,\text{d}x}{2\int\beta_1(x)^2\,\text{d}x}+\frac{(\int\beta_1(x)\beta_2(x)\,\text{d}x)^2}{16(\int\beta_1(x)^2\,\text{d}x)^2}}-\frac{\int\beta_1(x)\beta_2(x)\,\text{d}x}{4\int\beta_1(x)^2\,\text{d}x}},$$

where

$$\beta_1(x)=\frac{f''(x)}{2}\int u^2K(u)\,\text{d}u \quad \text{and} \quad \beta_2(x)=\frac{1}{2}\int s^2\alpha^{(2)}(x,s)\,\text{d}s. \quad (14)$$

Then the optimal m, n, and h, denoted m^*, n^*, and h^*, are

$$m^* \quad = \quad \left(\frac{2A^3\int\beta_1(x)\beta_2(x)\,\text{d}x+2A\int\beta_2(x)\,\text{d}x}{\int K^2(u)\,\text{d}u}\right)^{2/7}c^{2/7}, \quad (15)$$

$$n^* \quad = \quad \left(\frac{\int K^2(u)\,\text{d}u}{2A^3\int\beta_1(x)\beta_2(x)\,\text{d}x+2A\int\beta_2(x)\,\text{d}x}\right)^{2/7}c^{5/7}, \text{ and} \quad (16)$$

$$h^* \quad = \quad A\left(\frac{\int K^2(u)\,\text{d}u}{2A^3\int\beta_1(x)\beta_2(x)\,\text{d}x+2A\int\beta_2(x)\,\text{d}x}\right)^{1/7}c^{-1/7}. \quad (17)$$

Substituting m^*, n^*, and h^* into (13) shows that the optimal rate of convergence is of the order $c^{-4/7}$. In fact, when m, n, and h are chosen such that m is of the order $c^{2/7}$, n is of the order $c^{5/7}$, and h is of the order $c^{-1/7}$ the optimal rate of convergence of mise is achieved. We note that for the case in which Z is assumed to be univariate, the optimal rate of convergence is also $c^{-4/7}$ [25].

The constants in Equations (15–17) are unlikely to be tractable to estimate; the main purpose of the result is to provide the optimal rate of convergence.

In standard kernel density estimation, the optimal rate of convergence is $c^{-4/5}$ ([29]), while the associated constants are often intractable. One of the contributions of this paper is to provide the optimal rate of convergence of our estimator given additional bias due to measurement error. The decrease in the rate of convergence is a consequence of the additional bias. For each of the n observations $\bar{X}_m(Z_i)$, we must use m internal samples to deal with the measurement error bias, and $m \to \infty$ as $c \to \infty$. In the standard kernel density estimation setting, each observation requires only one sample since there is no measurement error. Note that although we phrased the optimal rate of convergence in terms of mise, the same applies to the mse. So the optimal rate of convergence of mse for our estimator $\hat{f}(x; m, n, h)$ is $c^{-4/7}$.

A local kernel estimate can be constructed, based on local kernel density estimation. It allows the bandwidth to be a function of the point at which the density function is being estimated, i.e., the local estimator is constructed by replacing h in Equation (4) with $h(x)$. The mise convergence rate of the local estimator is the same as that of the standard estimator, but the local estimator can have better performance in practice. Results are available in [24].

2.2. A bias-corrected estimator

In this section, we introduce a bias-corrected estimator of the density of the conditional expectation. We motivate the estimator with a discussion of the jackknife bias-corrected estimator; see [6] for an introduction. We present some results on the asymptotic bias and variance of the bias-corrected estimate and show that the optimal rate of mse convergence is faster than for the standard estimator.

The jackknife estimator can be thought of as an extrapolation from one estimate back to another estimate that has nearly zero bias (e.g., [27]). To understand this interpretation of the jackknife estimator, we turn to an example. A similar example was presented in [27]. Suppose we want to estimate $\theta = g(\mu)$ where g is nonlinear and twice continuously differentiable. We are given i.i.d. data $\{X_1, \ldots, X_m\}$ drawn from a $N(\mu, \sigma^2)$ distribution. We take our estimate, denoted $\hat{\theta}_m$, to be $g(\bar{X}_m)$ where $\bar{X}_m$ is the sample mean of the data. Under integrability assumption on the error, we can use Taylor expansion to show that for an estimate based on any sample size m,

$$\mathrm{E}(\hat{\theta}_m) \approx \theta + \frac{1}{m}\beta. \tag{18}$$

We actually know that $\beta = \sigma^2 g''(\mu)/2$, but that is not needed for our discussion. The point is that the bias, $\mathrm{E}(\hat{\theta}_m) - \theta$, is approximately linear in the inverse sample size m. Then if we know β and $\mathrm{E}(\hat{\theta}_m)$ for some m, by extrapolating on the line given in (18) back to $1/m = 0$, we have a nearly unbiased estimate of θ. The remaining bias is from the lower order terms in the Taylor expansion of $\mathrm{E}(\hat{\theta}_m)$.

If we have an estimate of $\mathrm{E}(\hat{\theta}_m)$, all we need is another estimate $\mathrm{E}(\hat{\theta}_{\tilde{m}})$ for $\tilde{m} \neq m$ to estimate β. For the standard jackknife estimator, $\mathrm{E}(\hat{\theta}_m)$ is estimated with $\hat{\theta}_m$ and $\mathrm{E}(\hat{\theta}_{m-1})$ is estimated with $\hat{\theta}_{(\cdot)} = \sum_{k=1}^{m} \hat{\theta}_{(k)}/m$ where for $k = 1, \ldots, m$, $\hat{\theta}_{(k)}$, the leave-one-out estimator, is the estimator based on all the data less X_k. The jackknife bias-corrected estimator $\dot{\theta}$ is then

$$\dot{\theta} \quad = \quad \hat{\theta}_m - (m-1)(\hat{\theta}_{(\cdot)} - \hat{\theta}_m) = m\hat{\theta}_m - (m-1)\hat{\theta}_{(\cdot)}.$$

For our standard estimator (4), we know from Theorem 2 that

$$\mathrm{E}(\hat{f}(x; m, n, h)) \approx f(x) + h^2\beta_1 + \frac{1}{m}\beta_2, \tag{19}$$

where β_1 and β_2 are defined in Equation (14). Here the bias is approximately linear in the square of the bandwidth (h^2) and the inverse of the internal sample size ($1/m$). Given an estimate of $\mathrm{E}(\hat{f}(x; m, n, h))$ for some m and h, we would like to extrapolate back to $1/m = 0$ and $h^2 = 0$ on the plane specified in (19).

Similar to the typical jackknife estimator, we take the standard estimate $\hat{f}(x; m, n, h)$ as an approximation of $\mathrm{E}(\hat{f}(x; m, n, h))$. To determine β_1 and β_2 and thus extrapolate back to $1/m = 0$ and $h^2 = 0$, we need to estimate $\mathrm{E}(\hat{f}(x; m, n, h))$ at two other pairs of (m, h). Alternatively, we can save ourselves a bit of work by choosing only one other pair $(\tilde{m}, \tilde{h})$ such that $(1/\tilde{m}, \tilde{h^2})$ lies on the line determined by $(0, 0)$ and $(1/m, h^2)$.

We could estimate $\mathrm{E}(\hat{f}(x; \tilde{m}, n, \tilde{h}))$ as the average of the leave-one-out estimators as is done for the typical jackknife estimator. This will require m computations of the density estimator. As a computationally attractive alternative, consider instead taking $\tilde{m} = m/2$ and $\tilde{h} = \sqrt{2}h$ and take the estimate $\hat{f}(x; \tilde{m}, n, \tilde{h})$ as an approximation of $\mathrm{E}(\hat{f}(x; \tilde{m}, n, \tilde{h}))$. Note that $(1/\tilde{m}, \tilde{h^2})$ lies on the line determined by $(0, 0)$ and $(1/m, h^2)$.

Using the data points $\hat{f}(x; m, n, h)$ and $\hat{f}(x; m/2, n, \sqrt{2}h)$ and extrapolating back to $1/m = 0$ and $h^2 = 0$ gives the bias-corrected estimator

$$\dot{f}(x; m, n, h) = 2\hat{f}(x; m, n, h) - \hat{f}(x; m/2, n, \sqrt{2}h). \tag{20}$$

We emphasize that just like the leave-one-out jackknife estimator, the data can be reused to estimate $\hat{f}(x; m/2, n, \sqrt{2}h)$. That is to say, the estimator $\hat{f}(x; m/2, n, \sqrt{2}h)$ can be computed with the same data set with which $\hat{f}(x; m, n, h)$ is computed less half of the internal samples. However in some cases, it would be possible to generate a new data set to estimate $\hat{f}(x; m/2, n, \sqrt{2}h)$. For the remainder of this section, we consider the asymptotic bias and variance of the bias-corrected estimator given in (20). The results cover both the case where the data is reused in computing $\hat{f}(x; m/2, n, \sqrt{2}h)$ and the case where a new data set is generated.

Based on Equation (12), the bias of the estimate $\dot{f}(x; m, n, h)$ can be expressed as

$$\begin{aligned}\text{bias}(\dot{f}(x;m,n,h)) &= \mathrm{E}(\dot{f}(x;m,n,h)) - f(x)\\ &= 2\left[\mathrm{E}(\hat{f}(x;m,n,h)) - f(x)\right] - \left[\mathrm{E}\hat{f}(x;m/2,n_0,\sqrt{2}h) - f(x)\right]\\ &= 2\left[(\mathrm{E}(\hat{f}(x;m,n,h)) - f_m(x)) + (f_m(x) - f(x))\right]\\ &\quad - \left[(\mathrm{E}(\hat{f}(x;m/2,n_0,\sqrt{2}h)) - f_{m/2}(x)) + (f_{m/2}(x) - f(x))\right]. \quad (21)\end{aligned}$$

From Lemma 6 in the Appendix ([26]),

$$\begin{aligned}\mathrm{E}(\hat{f}(x;m,n,h)) - f_m(x)) &= h^2\frac{1}{2}f^{(2)}(x)\int u^2K(u)\,\mathrm{d}u\\ &+ \frac{h^2}{m}\frac{1}{4}\int s^2\alpha^{(4)}(x,s)\,\mathrm{d}s\int u^2K(u)\,\mathrm{d}u + h^4\frac{1}{24}f^{(4)}(x)\int u^4K(u)\,\mathrm{d}u\\ &+ o\left(\frac{h^2}{m}+h^4\right)\end{aligned}$$

and

$$\begin{aligned}\mathrm{E}(\hat{f}(x;m/2,n_0,\sqrt{2}h)) &= 2h^2\frac{1}{2}f^{(2)}(x)\int u^2K(u)\,\mathrm{d}u\\ &+ 4\frac{h^2}{m}\frac{1}{4}\int s^2\alpha^{(4)}(x,s)\,\mathrm{d}s\int u^2K(u)\,\mathrm{d}u + 4h^4\frac{1}{24}f^{(4)}(x)\int u^4K(u)\,\mathrm{d}u\\ &+ o\left(\frac{h^2}{m}+h^4\right).\end{aligned}$$

From Lemma 5 in the Appendix ([26]),

$$f_m(x) - f(x) = \frac{1}{m}\frac{1}{2}\int s^2\alpha^{(2)}(x,s)\,\mathrm{d}s + \frac{1}{m^2}\frac{1}{8}\int s^4\alpha^{(4)}(x,s)\,\mathrm{d}s + o\left(\frac{1}{m^2}\right)$$

and

$$f_{m/2}(x) - f(x) = 2\frac{1}{m}\frac{1}{2}\int s^2\alpha^{(2)}(x,s)\,\mathrm{d}s + 4\frac{1}{m^2}\frac{1}{8}\int s^4\alpha^{(4)}(x,s)\,\mathrm{d}s + o\left(\frac{1}{m^2}\right).$$

Substituting into (21) proves the following theorem.

Theorem 4. Assume A1–A5 and A6(6). Also assume

1. K is a bounded probability distribution function symmetric about zero with finite fourth moment;
2. $m \to \infty$ and $h \to 0$ as $c \to \infty$.

Then

$$\begin{aligned}\text{bias}(\dot{f}(x;m,n,h)) \quad = \quad & -h^4\frac{1}{12}f^{(4)}(x)\int u^4K(u)\,\mathrm{d}u\\ & -\frac{h^2}{m}\frac{1}{2}\int s^2\alpha^{(4)}(x,s)\,\mathrm{d}s\int u^2K(u)\,\mathrm{d}u\\ & -\frac{1}{m^2}\frac{1}{4}\int s^4\alpha^{(4)}(x,s)\,\mathrm{d}s + o\left(\left(h^2+\frac{1}{m}\right)^2\right).\end{aligned}$$

As for the variance of $\dot{f}(x;m,n,h)$, note that from the proof of Theorem 2,

$$\mathrm{var}(\hat{f}(x;m,n,h)) \quad = \quad \frac{1}{nh} f(x) \int K^2(u)\,\mathrm{d}u + o\left(\frac{1}{nh}\right)$$

and

$$\mathrm{var}(\hat{f}(x;m/2,n,\sqrt{2}h)) \quad = \quad \frac{1}{\sqrt{2}nh} f(x) \int K^2(u)\,\mathrm{d}u + o\left(\frac{1}{nh}\right).$$

Also,

$$\begin{aligned}
&|\mathrm{cov}(\hat{f}(x;m,n,h), \hat{f}(x;m/2,n,\sqrt{2}h))| \\
&\quad \le \sqrt{\mathrm{var}(\hat{f}(x;m,n,h))\mathrm{var}(\hat{f}(x;m/2,n,\sqrt{2}h))} \\
&\quad \le \frac{1}{2^{1/4}} \frac{1}{nh} f(x) \int K^2(u)\,\mathrm{d}u + o\left(\frac{1}{nh}\right).
\end{aligned}$$

Then

$$\begin{aligned}
\mathrm{var}(\dot{f}(x;m,n,h)) &= \mathrm{var}(2\hat{f}(x;m,n,h) - \hat{f}(x;m/2,n,\sqrt{2}h)) \\
&= 4\mathrm{var}(\hat{f}(x;m,n,h)) + \mathrm{var}(\hat{f}(x;m/2,n,\sqrt{2}h)) \\
&\quad - 4\mathrm{cov}(\hat{f}(x;m,n,h), \hat{f}(x;m/2,n,\sqrt{2}h)) \\
&\le 4\frac{1}{nh} f(x) \int K^2(u)\,\mathrm{d}u + \frac{1}{\sqrt{2}nh} f(x) \int K^2(u)\,\mathrm{d}u \\
&\quad + 4\frac{1}{2^{1/4}} \frac{1}{nh} f(x) \int K^2(u)\,\mathrm{d}u + o\left(\frac{1}{nh}\right) \\
&= \left(4 + \frac{1}{2^{1/2}} + \frac{4}{2^{1/4}}\right) \frac{1}{nh} f(x) \int K^2(u)\,\mathrm{d}u + o\left(\frac{1}{nh}\right). \quad (22)
\end{aligned}$$

This shows that $\mathrm{var}(\dot{f}(x;m,n,h))$ is $O(\frac{1}{nh})$. Similarly,

$$\mathrm{var}(\dot{f}(x;m,n,h)) \ge \left(4 + \frac{1}{2^{1/2}} - \frac{4}{2^{1/4}}\right) \frac{1}{nh} f(x) \int K^2(u)\,\mathrm{d}u + o\left(\frac{1}{nh}\right). \quad (23)$$

Since

$$4 + \frac{1}{2^{1/2}} - \frac{4}{2^{1/4}} \approx 1.34,$$

we conclude that the asymptotic variance of $\dot{f}(x;m,n,h)$ is greater than the asymptotic variance of the standard estimator $\hat{f}(x;m,n,h)$. Therefore, it is likely the actual variance of the bias-corrected estimate is greater than the variance for the standard estimate. This is a common theme for bias-corrected estimates ([6]).

The above asymptotic bias and variance results for $\dot{f}(x;m,n,h)$ imply that if m, n, and h are chosen such that m is of the order $c^{2/11}$, n is of the order $c^{9/11}$, and h is of the order $c^{-1/11}$ the optimal rate of convergence of mse is

obtained and that optimal rate is $c^{-8/11}$. Recall the optimal rate of mse for the standard estimator $\hat{f}(x; m, n, h)$ was $c^{-4/7}$. Thus, the bias-correction leads to improved convergence. But as we noted above, the variance is greater for the bias-corrected estimate and this can adversely affect performance, especially for modest sample sizes.

3. Estimation implementation and bandwidth selection

In this section, we address the implementation of our estimators for the density of the conditional expectation discussed in Section 2 and study their performance. Implementation requires the specification of a number of inputs. For the standard kernel density estimator presented in (3), one must choose the kernel K and the bandwidth h. For the estimators of the density of the conditional expectation including the standard kernel density estimator (4), and the bias-corrected estimator (20), one must choose K, h, as well as the number of external samples n and the number of internal samples m.

We choose K to be the Epanechnikov kernel which is $K(x) = 0.75(1 - x^2)$ $I(|x| < 1)$. [7] showed this kernel was optimal in terms of minimizing the mise for the standard kernel density estimator (3); see [29].

The rest of this section deals with the choice of the parameters m, n, and h. In Section 3.1 we consider the selection of these parameters for the standard kernel density estimator (4). We present a data-based method to select these parameters based on EBBS developed by [20]. We present the algorithm and briefly discuss why we chose this method. In Section 3.2, the data-based parameter selection method is applied to the bias-corrected estimator (20).

3.1. Standard estimator

In Section 2 we saw how to choose the bandwidth h, the number of internal samples m, and the number of external samples n for the standard estimator $\hat{f}(x; m, n, h)$ to obtain optimal convergence: see (15). However the expressions for m, n, and h given in (15) involve unknowns such as $f''(x)$, the second derivative of the target density, and $\int s^2 \alpha^{(2)}(x, s)\, \mathrm{d}s$ where $\alpha^{(2)}$ is defined in (7) and (8) as the second derivative with respect to the first argument of the function $\alpha(y, s) = g(s|y)f(y)$.

To implement the estimator $\hat{f}(x; m, n, h)$ in an optimal way, one could attempt to estimate these unknown quantities and plug these estimates into the expressions given in (15). This type of estimator is known as a plug-in estimator ([29]). In fact it is quite doable to estimate the unknowns f and f'' needed for the plug-in estimator. Other needed estimates, including an estimate of the second derivative of α, appear very difficult to obtain.

To choose the parameters m, n, and h needed to implement the estimator $\hat{f}(x; m, n, h)$ we turn from optimizing the asymptotic mise to optimizing an

approximation of mise. Note that mise can be decomposed as

$$\operatorname{mise}(\hat{f}(\cdot; m, n, h)) = \int \operatorname{bias}^2(\hat{f}(x; m, n, h))\,\mathrm{d}x + \int \operatorname{var}(\hat{f}(x; m, n, h))\,\mathrm{d}x.$$

It was shown in the proof of Theorem 3 that

$$\int \operatorname{var}(\hat{f}(x; m, n, h))\,\mathrm{d}x = \frac{1}{nh}\int K^2(u)\,\mathrm{d}u + o\left(\frac{1}{nh}\right).$$

An approximation for the variance component in mise is the asymptotic approximation,

$$\frac{1}{nh}\int K^2(u)\,\mathrm{d}u,$$

which is readily available. Also in the proof of Theorem 3, it was shown that

$$\begin{aligned}\int \operatorname{bias}^2(\hat{f}(x; m, n, h))\,\mathrm{d}x &= \int \left(h^2\frac{1}{2}\left(\int u^2 K(u)\,\mathrm{d}u\right) f''(x) + \frac{1}{m}\frac{1}{2}\int s^2\alpha^{(2)}(x, s)\,\mathrm{d}s\right)^2 \mathrm{d}x \\ &\quad + o\left(\left(h^2 + \frac{1}{m}\right)^2\right).\end{aligned}$$

As explained above, the asymptotic approximation

$$\int \left(h^2\frac{1}{2}\left(\int u^2 K(u)\,\mathrm{d}u\right) f''(x) + \frac{1}{m}\frac{1}{2}\int s^2\alpha^{(2)}(x, s)\,\mathrm{d}s\right)^2 \mathrm{d}x$$

is not immediately useful given the unknowns in the approximation. To approximate the bias component in mise we will instead build and estimate a model of bias for each x. Squaring the bias and numerically integrating will then provide an empirical model of integrated squared bias. Adding the integrated variance approximation to this gives an empirical model of mise which can then be optimized with respect to m, n, and h.

The idea of building and empirically estimating a model of bias to be used in the selection of an estimator's parameters was introduced in [20]. It is called the empirical-bias bandwidth selection (EBBS) method, which is developed in [20] for local polynomial regression. EBBS uses a model of bias suggested by the asymptotic expression of the expected value of the estimator.

In our case, by Lemmas 5 and 6 in the Appendix ([26]),

$$\begin{aligned}\mathrm{E}\big(\hat{f}(x; m, n, h)\big) &= f(x) + h^2\frac{1}{2}f''(x)\int u^2 K(u)\,\mathrm{d}u + \frac{1}{m}\frac{1}{2}\int s^2\alpha^{(2)}(x, s)\,\mathrm{d}s \\ &\quad + o(h^2 + \frac{1}{m}).\end{aligned}$$

The asymptotic expression

$$\mathrm{E}\big(\hat{f}(x; m, n, h)\big) = f(x) + h^2\frac{1}{2}f''(x)\int u^2 K(u)\,\mathrm{d}u + \frac{1}{m}\frac{1}{2}\int s^2\alpha^{(2)}(x, s)\,\mathrm{d}s, \quad (24)$$

suggests the following model:

$$\mathrm{E}\big(\hat{f}(x;m,n,h)\big) = \beta_0(x) + \beta_1(x)h^2 + \beta_2(x)\frac{1}{m}. \tag{25}$$

Here $\beta_0(x)$ approximately corresponds to $f(x)$, the target density evaluated at x. The bias of $\hat{f}(x;m,n,h)$ is then approximately given by

$$\beta_1(x)h^2 + \beta_2(x)\frac{1}{m}. \tag{26}$$

The EBBS model of bias used in local polynomial regression is a polynomial in h ([20, 22]). In our case the model of bias is polynomial in h as well as $1/m$. Lemmas 5 and 6 in the Appendix ([26]) allow for more terms used in the asymptotic expression of $\mathrm{E}\big(\hat{f}(x;m,n,h)\big)$ given in (24) which would give more terms in model (25). Such a model would be a better approximation of $\mathrm{E}\big(\hat{f}(x;m,n,h)\big)$ but would require the estimation of additional parameters. In this paper, we use the model (25).

Though approximate, notice that the model of bias does capture the fact that as $h \to 0$ and $1/m \to 0$, bias tends to zero. Suppose that we can estimate the model (25). This not only gives us an empirical model of bias that can be used in selecting the needed parameters m, n, and h but also gives another estimator which will be of some use. Extrapolating the estimated model to $h = 1/m = 0$ gives an approximately unbiased estimate of $f(x)$. This approximately unbiased estimate of $f(x)$ is of course $\hat{\beta}_0$, the estimate of β_0. Based on the discussion of jackknife bias-correction, one can argue $\hat{\beta}_0$ is essentially a jackknife estimate. For more on this see [22].

The estimation procedure of the model (25) at x_0 for a given experiment budget c is outlined in Appendix ([26]) 3.

3.2. Bias-corrected estimator

Now we turn to the implementation of the bias-corrected estimator presented in Section 2.2. We use the same data to compute the estimators on the RHS. We again would like to use an expression for asymptotic mise to guide the modeling of mise. Recalling the decomposition of mise, we thus need asymptotic expressions for integrated, squared bias and integrated variance. Theorem 4 gives an asymptotic expression for bias. Let us assume that we can integrate squared bias so that we have the asymptotic expression of integrated, squared bias

$$\int \left(-h^4 \frac{1}{12} f^{(4)}(x) \int u^4 K(u)\,\mathrm{d}u - \frac{h^2}{m}\frac{1}{2}\int s^2 \alpha^{(4)}(x,s)\,\mathrm{d}s \int u^2 K(u)\,\mathrm{d}u \right.$$
$$\left. -\frac{1}{m^2}\frac{1}{4}\int s^4 \alpha^{(4)}(x,s)\,\mathrm{d}s \right)^2 \mathrm{d}x.$$

This suggests that we model the expectation of $\dot{f}_L(x;m,n,h)$ as

$$\mathrm{E}\big(\dot{f}_L(x;m,n,h)\big) = \beta_0(x) + \beta_1(x)h^4 + \beta_2(x)\frac{h^2}{m} + \beta_3(x)\frac{1}{m^2}.$$

The bias of $\dot{f}_L(x; m, n, h)$ is then approximately

$$\beta_1(x)h^4 + \beta_2(x)\frac{h^2}{m} + \beta_3(x)\frac{1}{m^2}. \tag{27}$$

Let us also assume that the upper and lower bounds on variance given in (22) and (23) integrate. Moreover, since we are reusing the data, assume that the covariance of $\hat{f}_L(x; m, n, h)$ and $\hat{f}_L(x; m/2, n, \sqrt{2}h)$ is equal to the approximate upper bound

$$\frac{1}{2^{1/4}}\frac{1}{nh}f(x)\int K^2(u)\,\mathrm{d}u,$$

so that we can approximate the variance component in mise with the integrated asymptotic expression from the lower bound of $\mathrm{var}(\dot{f}_L(x; m, n, h)$. This approximation is

$$\left(4 + \frac{1}{2^{1/2}} - \frac{4}{2^{1/4}}\right)\frac{1}{nh}\int K^2(u)\,\mathrm{d}u. \tag{28}$$

We thus have an approximation for the variance component of mise (28) and a model for the bias (27). The tuning parameter values for standard estimators mentioned in Appendix ([26]) 3 work well here.

4. Numerical experiments

In this section we examine the performance of the implementations discussed in the previous section on the sodium measurement example, along with another test case and a financial risk management example. To assess performance we consider representative plots and the behavior of estimated mise.

4.1. Test case

In this two-dimensional test case, $Z = (Z_1, Z_2)$ has a standard bivariate normal distribution. Conditional on Z,

$$X(Z) =_d N\left(Z_1 + Z_2, \left(1 - \frac{1}{1 + 2^{-1/2}|Z_1 - Z_2|}\right)^2\right).$$

Then the random variable $\mathrm{E}(X|Z) = Z_1 + Z_2$ is normally distributed with mean 0 and variance 2. This is a straightforward example in which all the assumptions for Theorem 3 are satisfied. We consider this example mainly to numerically verify that the rate of mise convergence for the standard estimator is $c^{-4/7}$ as suggested by Theorem 3.

In Figure 1, the standard density estimator is plotted for two different experiment budgets along with the target density for the first test case. The figure shows that, as expected, the performance of the estimator improves as the experiment budget increases.

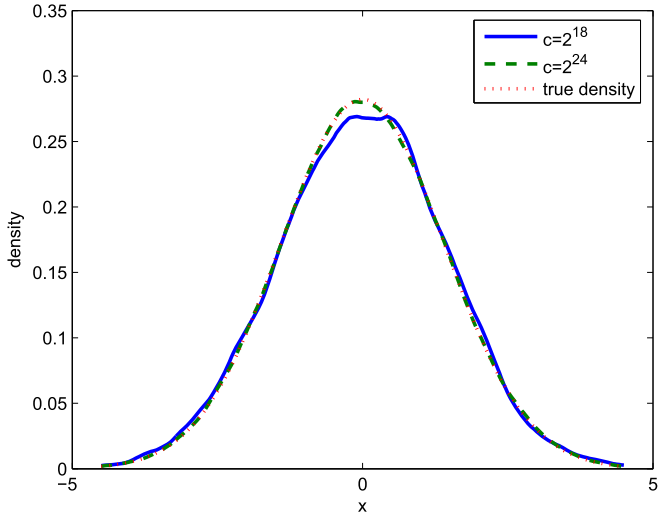

FIG 1. *The standard kernel density estimator for two different experiment budgets along with the target density.*

We now turn to mise convergence. For clarity, we no longer suppress the dependence of the various estimators and parameters on the experiment budget c. To estimate $\text{mise}(c)$, mise at a given experiment budget c, we first replicate the density estimator 50 times:

$$\{\hat{f}(\cdot; m(c), n(c), h(c))_k : k = 1, \ldots, 50\}.$$

We define integrated squared error (ise) as follows:

$$\text{ise}(c) = \int [\hat{f}(x; m(c), n(c), h(c)) - f(x)]^2 \, \mathrm{d}x.$$

For each $k = 1, \ldots, 50$, we use numerical integration to compute

$$\text{ise}_k(c) = \int [\hat{f}(x; m(c), n(c), h(c))_k - f(x)]^2 \, \mathrm{d}x.$$

Our estimate for $\text{mise}(c)$ is then

$$\hat{\text{mise}}(c) = \frac{1}{50} \sum_{k=1}^{50} \text{ise}_k(c).$$

In Figure 2, we plot $\log(\text{mise}(c))$ vs. $\log(c)$ at $c = 2^{18}$, 2^{20}, 2^{22}, 2^{24} and the least squares regression line for the standard estimator. The linearity of the plot suggests that over the particular range of experiment budgets c, the estimator's $\text{mise}(c)$ has the form $\text{mise}(c) = Vc^{\gamma}$ for some constants V and γ. Suppose that $\hat{\delta}_0$ and $\hat{\delta}_1$ are the estimated intercept and slope of the regression line plotted

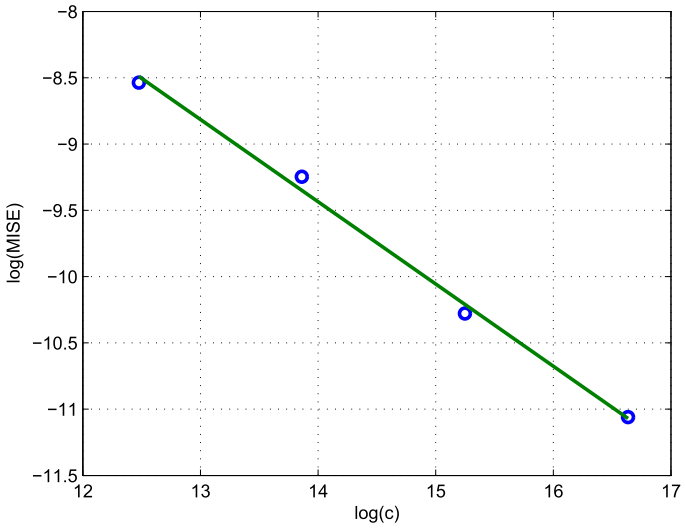

FIG 2. *Plot of log(mise(c)) vs. log(c) at* $c = 2^{18}, 2^{20}, 2^{22}$ *for the standard kernel density estimator.*

in the figures. Then $\hat{\delta}_1$ estimates γ and $\exp(\hat{\delta_0}/\hat{\delta_1})$ estimates V. Given that the optimal mise convergence rate is $c^{-4/7}$ we expect that, asymptotically, $\gamma = -4/7 \approx -0.57$. The estimated intercept and slope in Figure 2 are -7.51 and -0.62, respectively. So it appears that the estimator performs as expected.

4.2. Sodium measurement example

In this section, we return to the sodium measurement example described in Section 1 and show that the bias-corrected estimator we proposed in Section 2.2 outperforms the standard estimator even when the experiment budget is small.

Consider an experiment with $m = 5$ repeated sodium measurements for each of $n = 300$ batches of ingredient having true sodium content $Z_i, i = 1, \ldots, n$. For the purpose of assessing the performance of the estimators, we use simulation to generate data for this example. In the simulation, The distribution of Z is a three parameter lognormal distribution: lognormal$(\mu = 1.544, \sigma = 0.5, t = 2)$, which has mean 7.307 and standard deviation 2.828. Note that parameters μ and σ correspond to the mean and standard deviation of the variable's natural logarithm, while t is the location parameter. For any i and j, we take the measurement error $U_{ij} = X_{ij} - Z_i$ to be normally distributed with standard deviation $0.6 + 0.5Z_i$. The bias-corrected estimator significantly outperforms the standard estimator in terms of mean integrated squared error: when the experiment was repeated 5000 times, the estimators had mise of 0.0061 and 0.0103, respectively.

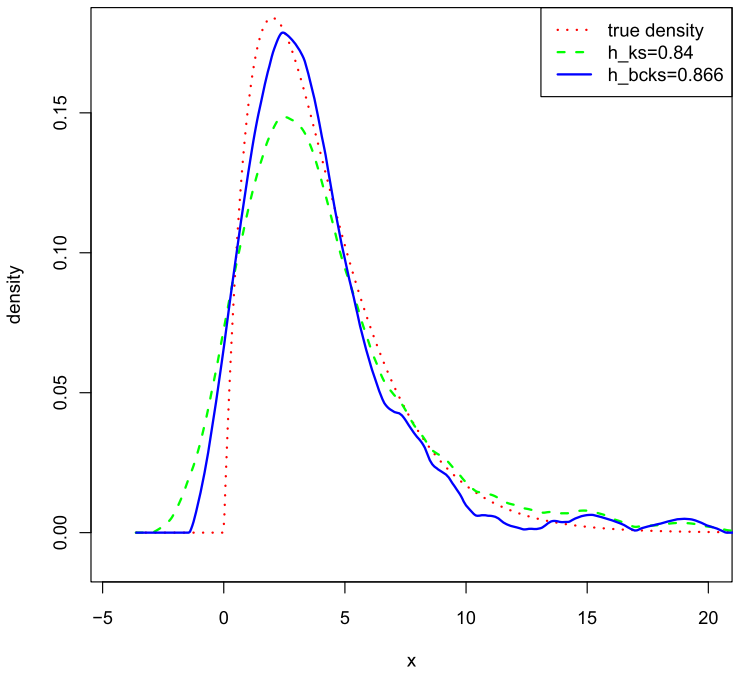

FIG 3. *The true density, standard kernel-smoothing estimator, and bias-corrected kernel-smoothing estimator, in the sodium measurement example.* h_{ks} *and* h_{bcks} *are bandwidths used for standard kernel-smoothing estimator, and bias-corrected kernel-smoothing estimator respectively.*

Figure 3 plots the bias-corrected estimator (solid curve) compared with the standard estimator (dashed curve) for one simulated data set. The bias-corrected estimator is much closer to the true density in the center of the distribution and in the left tail. Although the bias-corrected estimator is less smooth than the standard estimator in the right tail, this is mainly due to the presence of few observations near extreme quantiles. In this situation, poor density estimation in the tail is a typical problem for kernel density estimators [29].

4.3. Financial risk management example

In this section, we apply our bias-corrected estimator to an example in financial risk management. This financial simulation example was used and described in detail by [28]. The scenario $Z = (S_1, \cdots, S_s)$ is an s-dimensional random vector, representing a simulated time series of future stock prices. We want to know the density of Y, the profit and loss (P&L) that would be earned by a particular financial strategy in this scenario. The P&L in this example is

$$Y = (p_0 + \Delta_0 S_0)e^{rT} + \sum_{k=1}^{s} (\Delta_k - \Delta_{k-1}) S_k e^{r(T-t_k)} - |S_s - Q|, \qquad (29)$$

where p_0, Δ_0, S_0, r, T, Q and $t_1, \ldots, t_s$ are known parameters, $S_1, \cdots, S_s$ are the simulated future stock prices that make up the scenario Z, and $\Delta_1, \ldots, \Delta_s$ are amounts of stock that will be owned at the future times $t_1, \ldots, t_s$. However, $\Delta_1, \ldots, \Delta_s$ are unknown functions of Z. Therefore Y cannot be directly observed based on simulating the scenario Z. However, Y can be observed with noise. To get such an observation, X, we simulate $\psi_1, \ldots, \psi_s$ conditional on Z, where each ψ_k is an unbiased, noisy observation of Δ_k. This is possible because it is known how to simulate a random variable ψ_k whose conditional expectation given Z is Δ_k. Then

$$X = (p_0 + \Delta_0 S_0)e^{rT} + \sum_{k=1}^{s} (\psi_k - \psi_{k-1}) S_k e^{r(T-t_k)} - |S_s - Q| \tag{30}$$

and the P&L $Y = E[X|Z]$. The resulting simulation is a nested simulation: the outer level of simulation samples the scenario $Z = (S_1, \cdots, S_s)$, and the inner level of simulation samples $\psi_1, \ldots, \psi_s$ given Z and then computes X. The steps in this simulation are:

- For $i = 1, \ldots, n$, simulate scenario $Z_i = (S_{i1}, \cdots, S_{is})$.
  - For $j = 1, \cdots, m$, simulate $\psi_{ij1}, \ldots, \psi_{ijs}$ conditional on Z_i and calculate

$$X_{ij} = (p_0 + \Delta_0 S_0)e^{rT} + \sum_{k=1}^{s} (\psi_{ijk} - \psi_{i,j,k-1}) S_{ik} e^{r(T-t_k)} - |S_{is} - Q|.$$

Figure 4 plots the bias-corrected estimator (solid curve) compared with the standard estimator (dashed curve) for one simulated data set. The bias-corrected estimator outperforms the standard estimator on this data set where $m = 40$ and $n = 5000$. An inner-level sample size of 40 was found to be nearly optimal for the purpose of [28], which was to estimate the variance of the P&L Y.

5. Conclusions

We proposed a bias-corrected estimator for the density of a conditional expectation, based on kernel smoothing. We derived results about the convergence rates of this estimator and a standard kernel smoothing estimator; the bias-corrected estimator has a superior convergence rate. Using the asymptotic analysis and EBBS, we created algorithms for choosing the bandwidth and the sample sizes given an experiment budget. When applied to a practical example with moderate sample sizes, the bias-corrected estimator performed better than the standard estimator.

Supplementary Material

Appendix to "Estimating the density of a conditional expectation" (doi: 10.1214/16-EJS1121SUPP; .pdf).

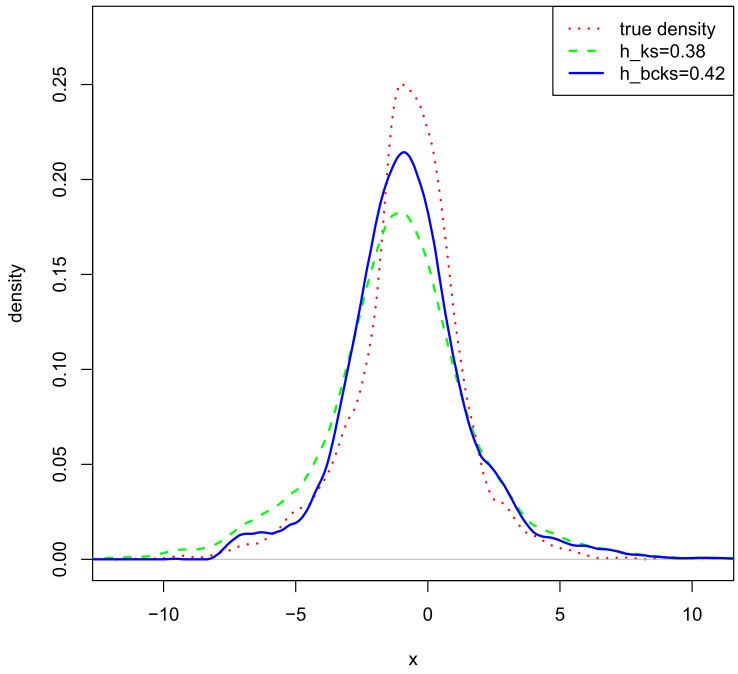

FIG 4. *The true density, standard kernel-smoothing estimator, and bias-corrected kernel-smoothing estimator, in the P&L example.* h_{ks} *and* h_{bcks} *are bandwidths used for standard kernel-smoothing estimator, and bias-corrected kernel-smoothing estimator respectively.*

References

[1] M. Broadie, Y. Du, and C. C. Moallemi. Efficient risk estimation via nested sequential simulation. *Management Science*, 57:1172–1194, 2011.

[2] R. J. Carroll and P. Hall. Optimal rates of convergence for deconvolving a density. *Journal of the American Statistical Association*, 83(404):1184–1186, Dec 1988. MR0997599

[3] A. Delaigle and A. Meister. Density estimation with heteroscedastic error. *Bernoulli*, 14(2):562–579, 2008. MR2544102

[4] A. Delaigle, P. Hall, and A. Meister. On deconvolution with repeated measurements. *Annals of Statistics*, 36(2):665–685, 2008. MR2396811

[5] L. Devroye and G. Lúgosi. *Combinatorial Methods in Density Estimation.* Springer, New York, 2001. MR1843146

[6] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap.* Chapman & Hall, New York, 1993. MR1270903

[7] V. A. Epanechnikov. Non-parametric estimation of a multivariate probability density. *Theory of Probability and its Applications*, 14(1):153–158, 1967.

[8] P. Glasserman. Performance continuity and differentiability in Monte Carlo optimization. In M. Abrams, P. Haigh, and J. Comfort, editors, *Proceedings*

of the 1988 Winter Simulation Conference, pages 518–524, Piscataway, NJ, 1988. IEEE.

[9] M. B. Gordy and S. Juneja. Nested simulation in portfolio risk measurement. *Management Science*, 56:1833–1848, 2010.

[10] S. G. Henderson. Input model uncertainty: why do we care and what should we do about it? In S. E. Chick, P. J. Sánchez, D. J. Morrice, and D. Ferrin, editors, *Proceedings of the 2003 Winter Simulation Conference*, page To appear, Piscataway, NJ, 2003. IEEE.

[11] M. C. Jones and D. F. Signorini. A comparison of higher-order bias kernel density estimators. *Journal of the American Statistical Association*, 92 (439):1063–1073, September 1997. MR1482137

[12] P. L'Ecuyer. A unified view of the IPA, SF and LR gradient estimation techniques. *Management Science*, 36:1364–1383, 1990.

[13] P. L'Ecuyer. On the interchange of derivative and expectation for likelihood ratio derivative estimators. *Management Science*, 41:738–748, 1995.

[14] S. H. Lee. *Monte Carlo Computation of Conditional Expectation Quantiles*. PhD thesis, Stanford University, Stanford, CA, 1998.

[15] S. H. Lee and P. W. Glynn. Computing the distribution function of a conditional expectation via monte carlo: discrete conditioning spaces. *ACM Transactions on Modeling and Computer Simulation*, 13(3):238–258, July 2003.

[16] J. McIntyre and L. A. Stefanski. Density estimation with replicate heteroscedastic measurements. *Annals of the Institute of Statistical Mathematics*, 63:81–99, 2011. MR2748935

[17] P. Patil. A note on deconvolution density estimation. *Statistics & Probability Letters*, 29(1):79–84, 1996. MR1411171

[18] B. L. S. Prakasa Rao. *Nonparametric Functional Estimation*. Academic Press, New York, 1983. MR0740865

[19] W. Rudin. *Real and Complex Analysis*. McGraw-Hill, New York, 1987. MR0924157

[20] D. Ruppert. Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation. *Journal of the American Statistical Association*, 92(439):1049–1062, September 1997. MR1482136

[21] A. E. Schulman. *A Comparison of Local Bandwidth Selectors for Local Polynomial Regression*. PhD thesis, Cornell University, Ithaca, NY, 1998.

[22] J. Staudenmayer and D. Ruppert. Local polynomial regression and simulation-extrapolation. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 66(1):pp. 17–30, 2004. ISSN 13697412. URL http://www.jstor.org/stable/3647624. MR2035756

[23] J. Staudenmayer, D. Ruppert, and J. P. Buonaccorsi. Density estimation in the presence of heteroscedastic measurement error. *Journal of the American Statistical Association*, 103(482):726–736, 2008. MR2524005

[24] S. G. Steckley. *Estimating the density of a conditional expectation*. PhD thesis, Cornell University, Ithaca, NY, 2005.

[25] S. G. Steckley and S. G. Henderson. A kernel approach to estimating the density of a conditional expectation. In S. E. Chick, P. J. Sánchez, D. J.

Morrice, and D. Ferrin, editors, *Proceedings of the 2003 Winter Simulation Conference*, pages 383–391, Piscataway, NJ, 2003. IEEE.

[26] S. G. Steckley, S. G. Henderson, D. Ruppert, R. Yang, D. W. Apley, and J. Staum. Appendix (supplemental materials) for "Estimating the density of a conditional expectation". *Electronic Journal of Statistics*, 2016, DOI: 10.1214/16-EJS1121SUPP.

[27] L. A. Stefanski and J. R. Cook. Simulation-extrapolation: the measurement error jacknife. *Journal of the American Statistical Association*, 90(432): 1247–1256, Dec 1995. MR1379467

[28] Y. Sun, D. W. Apley, and J. Staum. Efficient nested simulation for estimating the variance of a conditional expectation. *Operations Research*, 59:998–1007, 2011. MR2844419

[29] M. P. Wand and M. C. Jones. *Kernel Smoothing*. Chapman & Hall, London, 1995. MR1319818