A Note on the Unconditional Bias of the Nadaraya-Watson Regression Estimator

Kellin Pelrine

Department of Economics
University of Colorado at Boulder
Boulder, CO 80309-0256, USA
Email: kellin.pelrine@colorado.edu

Defended April 9th, 2018

Defense Committee:
Carlos Martins-Filho, Department of Economics, Advisor
Martin Boileau, Department of Economics
Sergei Kuznetsov, Department of Mathematics

Abstract. In this note we investigate the order of the unconditional bias of the Nadarya-Watson (Nadaraya, 1964; Watson, 1964) estimator for a multivariate regression. Surprisingly, previous attempts in establishing this result are either imprecise and technically deficient, or of limited use given the assumptions imposed (see *inter alia,* Glad (1998), Mack and Müller (1988), Pagan and Ullah (1999), and Scott (2015)). The results are also often conflicting (see *inter alia,* Choi et al. (2000), Chu and Marron (1991), Collomb (1981), and Glad (1998)). Unfortunately, our result here is incomplete, but we highlight the issues and suggest further ideas to resolve them.

Keywords and phrases. Nonparametric regression estimation; unconditional bias of the Nadaraya-Watson estimator.

JEL Classifications. C10, C14, C21.

AMS-MS Classifications. 62F12, 62G08, 62G20.

1 Introduction

The Nadaraya-Watson nonparametric regression estimator (Nadaraya, 1964; Watson, 1964) is perhaps the most used and studied smoothing procedure. Despite its popularity, there are few explicit derivations of the structure and order of its bias in the existing literature. Fan (1992) and Scott (2015) give approximate expressions for its bias, but do not explicitly study its asymptotic behavior. Ziegler (2001) gives an exact expression when the regressand Y is bounded or has finite second moment, but does not consider the benefits regression differentiability can bring in terms of faster convergence results of the bias expressions as the sample size $n \to \infty$.

Glad (1998) gives strong and exact results, obtaining faster convergence rates than Ziegler using differentiability assumptions. However, the arguments used in her proof are not explicit. In particular, Glad relies on the estimator being almost surely (a.s.) bounded, but since there is no assumption that the regression error is a.s. bounded, it is not clear that such a bound on the estimator exists. Another part of her argument rests on a random variable with variance decaying at an $O(n^{-1})$ rate. The numerator in the expression that defines the Nadaraya-Watson estimator is then substituted for such arbitrary random variable. But later an expression for the variance of the numerator is given which decays at an $O((nh_n)^{-1})$ rate, where h_n is a nonstochastic bandwidth. Since $h_n \to 0$ this rate is slower than the required $O(n^{-1})$, and thus it is not clear that the substitution is warranted in her proof.

Mack and Müller (1988) gives a result matching Scott (2015) and cites Rosenblatt (1969), but the latter does not even have the result, never mind provide a thorough proof. Sources including Chu and Marron (1991), Collomb (1977), Jennen-Steinmetz and Gasser (1988), and Schimek (2013) cite or have citations leading to Collomb (1976). Seemingly inexplicably, despite drawing on the same source, they give different results. Their orders range over $O(\frac{1}{\sqrt{nh}})$, $O(\frac{1}{nh})$, and $O(\frac{1}{(nh)^2})$. Unfortunately, Collomb's 1976 dissertation is difficult to obtain, having been published only in France, and was unavailable in time to be examined in this paper.

These deficiencies in the extant literature have prompted us to revisit the asymptotic bias of the Nadaraya-

Watson estimator with the aim of providing a thorough proof. We assume regression differentiability, as in Glad (1998), but relax the assumption of finite conditional variance of the regression error by only assuming that its conditional expectation equals zero. Therefore, our result can be applied in some situations where Ziegler's result cannot. We provide a proof for the multivariate regression case, which many of the papers mentioned above do not include. In the single variable case, aside from the relaxed assumption above and some other minor modifications, our result matches that of Glad (1998).

We first present the model in section 2, then give our result in Theorem 2.1. Unfortunately, this proof is not complete. A bound for one of the terms has not yet been derived thoroughly. In subsection 2.1 we show the most extensive attempt so far, which highlights some of the issues, and then we present some avenues to complete the proof.

2 Model, estimator and results

We consider a sequence of independent and identically distributed random vectors $\{(Y_i \; X_i')'\}_{i=1}^n$ where $X_i \in \mathbb{R}^D$, $D \in \mathbb{N}$, is a vector of regressors and $Y_i \in \mathbb{R}$ is a regressand. The Nadaraya-Watson estimator for the regression $E(Y_i|X_i = x) \equiv m(x)$ is given by

$$\hat{m}(x) = \frac{A_n(x)}{B_n(x)} \text{ for } B_n(x) \neq 0, \tag{2.1}$$

where $A_n(x) = \frac{1}{n \det(H_n)} \sum_{i=1}^n K\big(H_n^{-1}(X_i - x)\big)Y_i$, $B_n(x) = \frac{1}{n \det(H_n)} \sum_{i=1}^n K\big(H_n^{-1}(X_i - x)\big)$, $H_n = diag\{h_{d,n}\}_{d=1}^D$ with $h_{d,n} > 0$ for all d and n, and K is a multivariate kernel function.[1] In what follows it is convenient to use multi-index notation. Let $v, \alpha \in \mathbb{R}^D$ where the components of α, denoted by $\alpha_i \in \{0, 1, 2, \cdots\}$. We define $v^\alpha = v_1^{\alpha_1} \cdots v_D^{\alpha_D}$, $|\alpha| = \sum_{i=1}^D \alpha_i$, $\alpha! = \alpha_1! \cdots \alpha_D!$ and for a sufficiently differentiable arbitrary function g, $D^\alpha g(x) = \dfrac{\partial^{|\alpha|} g(x)}{\partial x_1^{\alpha_1} ... \partial x_D^{\alpha_D}}$.

Theorem 2.1. *Let f be the marginal density of X_1, and assume that $f(x) > C > 0$. In addition, assume:*

1. *All partial derivatives of m and f up to order 3 exist and are uniformly bounded,*

2. *$E(m(X_1)) < \infty$,*

[1] Specific constraints on K and H_n that are needed in our results will be given in the theorem statement

3. $K(\alpha) = \prod_{d=1}^{D} k(\alpha_d)$, *where* k *compactly supported,* $0 \leq k(\gamma) \leq C$, $\int k(\gamma)d\gamma = 1$, $\int \alpha k(\alpha)d\alpha = 0$ *and* $\mu_2 \equiv \int \gamma^2 k(\gamma)d\gamma < \infty$.

4. $h_{d,n} \to 0 \quad \forall d$, $n \det(H_n) \to \infty$.

Then,

$$E(\hat{m}(x)) - m(x) = \frac{\mu_2}{2} \sum_{i=1}^{D} h_{i,n}^2 \left(\frac{2D_i m(x) D_i f(x)}{f(x)} + D_i^2 m(x) \right) + O\left(\operatorname{tr}(H_n^3) + \frac{\operatorname{tr}(H_n^2)}{n \det(H_n)} \right). \tag{2.2}$$

If we replace third order derivatives by fourth order derivatives, the $\operatorname{tr}(H_n^3)$ *term in the stated order, and in the proof, becomes* $\operatorname{tr}(H_n^4)$.

Proof. Let $a = (a_1 \ a_2)'$ and for $a_2 \neq 0$ let $g(a) = a_1/a_2$. Note that $D^\alpha g(a)$ exists for all α. Hence, for a, $b \in S$ where S is an open and convex subset of $\mathbb{R}^2$, with $a \in S$ implies $a_2 \neq 0$, by Taylor's Theorem we have

$$g(a) = \sum_{|\alpha| \leq k-1} \frac{1}{\alpha!}(a-b)^\alpha D^\alpha g(b) + k \sum_{|\alpha|=k} \frac{1}{\alpha!}(a-b)^\alpha \int_0^1 (1-t)^{k-1} D^\alpha g(b + t(a-b))dt.$$

Under the assumptions on the kernel k, H_n and f, $E(B_n(x)) \to f(x) > C > 0$. Thus, for sufficiently large n, $E(B_n(x)) > C > 0$ and we put $b_n(x) = \begin{pmatrix} E(A_n(x)) \\ E(B_n(x)) \end{pmatrix}$. Since the definition of $\hat{m}(x)$ requires $B_n(x) \neq 0$ we put $a_n(x) = \begin{pmatrix} A_n(x) \\ B_n(x) \end{pmatrix}$. Thus, for sufficiently large n

$$\begin{aligned} \hat{m}(x) = g\left(a_n(x)\right) = &\sum_{|\alpha| \leq k-1} \frac{1}{\alpha!}(a_n(x) - b_n(x))^\alpha D^\alpha g(b_n(x)) \\ &+ k \sum_{|\alpha|=k} \frac{1}{\alpha!}(a_n(x) - b_n(x))^\alpha \int_0^1 (1-t)^{k-1} D^\alpha g(b_n(x) + t(a_n(x) - b_n(x)))dt. \end{aligned}$$

Taking expectations on both sides and expanding the multi-index sum, we have

$$\begin{aligned} E(\hat{m}(x)) = &\frac{E(A_n(x))}{E(B_n(x))} - \frac{1}{(E(B_n(x)))^2} Cov(A_n(x), B_n(x)) + 2\frac{E(A_n(x))}{(E(B_n(x)))^3} V(B_n(x)) + \cdots \\ &+ (-1)^{k-2}(k-2)! \frac{1}{(E(B_n(x)))^{k-1}} E\Big((A_n(x) - E(A_n(x)))(B_n(x) - E(B_n(x)))^{k-2} \Big) \\ &+ (-1)^{k-1}(k-1)! \frac{E(A_n(x))}{(E(B_n(x)))^k} E\Big((B_n(x) - E(B_n(x)))^{k-1} \Big) \\ &+ (-1)^{k-1} k! \, E\left((B_n(x) - E(B_n(x)))^{k-1} \int_0^1 (1-t)^{k-1} \frac{A_n(x) - E(A_n(x))}{(E(B_n(x)) + t(B_n(x) - E(B_n(x))))^k} dt \right) \\ &+ (-1)^k k k! \, E\left((B_n(x) - E(B_n(x)))^k \int_0^1 (1-t)^{k-1} \frac{E(A_n(x)) + t(A_n(x) - E(A_n(x)))}{(E(B_n(x)) + t(B_n(x) - E(B_n(x))))^{k+1}} dt \right). \end{aligned}$$

We first consider

$$R_{n,1}(x) = (-1)^{k-1}k!E\left((B_n(x) - E(B_n(x)))^{k-1}\int_0^1 (1-t)^{k-1}\frac{A_n(x) - E(A_n(x))}{(E(B_n(x)) + t(B_n(x) - E(B_n(x))))^k}dt\right).$$

Letting $C_n(x) = \frac{1}{ndet(H_n)}\sum_{i=1}^n K\left(H_n^{-1}(X_i - x)\right)m(X_i)$ we note that by the law of iterated expectations and the triangle inequality

$$\begin{aligned}
R_{n,1}(x) &= (-1)^{k-1}k!E\left((C_n(x) - E(C_n(x))(B_n(x) - E(B_n(x)))^{k-1}\right. \\
&\quad \left.\times \int_0^1 (1-t)^{k-1}\frac{1}{(E(B_n(x)) + t(B_n(x) - E(B_n(x))))^k}dt\right) \\
|R_{n,1}(x)| &\leq (-1)^{k-1}k!E\left(|B_n(x) - E(B_n(x))|^{k-1}\int_0^1 \frac{B_n(x)|C_n(x)/B_n(x) - m(x)|}{|E(B_n(x)) + t(B_n(x) - E(B_n(x)))|^k}dt\right. \\
&\quad \left.+ |B_n(x) - E(B_n(x))|^{k-1}\int_0^1 \frac{|E(C_n(x))| + B_n(x)|m(x)|}{|E(B_n(x)) + t(B_n(x) - E(B_n(x))|^k}dt\right)
\end{aligned}$$

Letting $h_n = (h_{1,n} \cdots h_{D,n})'$ and $\bar{x}_j = \lambda X_j + (1-\lambda)x$ for some $\lambda \in (0,1)$, by Taylor's Theorem and the Triangle Inequality,

$$\begin{aligned}
|C_n/B_n - m(x)| &\leq \frac{1}{B_n(x)}\frac{1}{n\det(H_n)}\sum_{1\leq|\alpha|\leq 2}\frac{1}{\alpha!}|D^\alpha m(x)|\sum_{i=1}^n K\Big(H_n^{-1}(X_i - x)\Big)h_n^\alpha|H_n^{-1}(X_i - x)^\alpha| \\
&\quad + \frac{1}{B_n(x)}\frac{1}{n\det(H_n)}\sum_{|\alpha|=3}\frac{1}{\alpha!}|D^\alpha m(\bar{x})|\sum_{i=1}^n K\Big(H_n^{-1}(X_i - x)\Big)h_n^\alpha|H_n^{-1}(X_i - x)^\alpha| \\
&\leq C\sum_{1\leq|\alpha|\leq 3} h_n^\alpha.
\end{aligned}$$

The last inequality follows from the uniform bound on all partial derivatives of m up to order 3 and the compact support of K. Also, by the Triangle Inequality and the assumption that $E|m(X)| < \infty$ we have

$$E|C_n(x)| \leq \int K(\gamma)|m(x + H_n\gamma)|f(x + H_n\gamma)d\gamma \to |m(x)|f(x) \text{ as } n \to \infty,$$

which implies that $E|C_n(x)|$ is bounded. Thus,

$$\begin{aligned}
|R_{n,1}(x)| &\leq CE\left(\int_0^1 \frac{1}{|(1-t)E(B_n) + tB_n|^k}dt|B_n - E(B_n)|^{k-1} + \int_0^1 \frac{B_n}{|(1-t)E(B_n) + tB_n|^k}dt|B_n - E(B_n)|^{k-1}\right) \\
&\leq C\left(\int\left(\int_0^1 \frac{1}{|(1-t)E(B_n) + tB_n|^k}dt\right)^2 f(\alpha)d\alpha\right)^{1/2}\left(\int(B_n - E(B_n))^{2k-2}f(\alpha)d\alpha\right)^{1/2} \\
&\quad + C\left(\int B_n^4 f(\alpha)d\alpha\right)^{1/4}\left(\int\left(\int_0^1 \frac{1}{|(1-t)E(B_n) + tB_n|^k}dt\right)^4 f(\alpha)d\alpha\right)^{1/4} \\
&\quad \cdot\left(\int(B_n - E(B_n))^{2k-2}f(\alpha)d\alpha\right)^{1/2}
\end{aligned}$$

where the last inequality follows by multiple applications of the Cauchy-Schwarz inequality. The fourth moment of B_n is bounded by an argument similar to that for the first moment. Next,

$$\int_0^1 \frac{1}{|(1-t)E(B_n)+tB_n|^k}dt = \int_0^1 \frac{1}{((1-t)E(B_n)+tB_n)^k}dt$$

since B_n is positive, and then,

$$\int_0^1 \frac{1}{((1-t)E(B_n)+tB_n)^k}dt = \frac{\sum_{i=0}^{k-2} E(B_n)^i B_n^{k-2-i}}{(k-1)E(B_n)^{k-1}B_n^{k-1}}$$

by direct integration. Note that the powers in the numerator are smaller than those in the denominator, so one can expand the sum and cancel, leaving $\sum_{i=1}^{k-1} \frac{1}{(k-1)E(B_n)^i B_n^{k-i}}$. For sufficiently large n, since $E(B_n)$ is bounded away from 0 by $C > 0$ and non-stochastic, we can replace it with $min(C, C^{k-1}) = C_1$, making the entire sum larger. Finally,

$$\sum_{i=1}^{k-1} \frac{1}{(k-1)C_1 B_n^{k-i}} \leq \frac{1}{C_1} max(1/B_n, 1/B_n^{k-1}).$$

If $B_n \geq 1$, in which case the maximum is $1/B_n$, this is bounded by $1/C_1$. Otherwise the maximum is $1/B_n^{k-1}$. We examine bounding the expectation of this term in subsection 2.1. In the rest of this section we assume it is bounded.

Similar arguments apply to the second remainder term, noting that we can separate $E(A_n)+t(A_n-E(A_n))$ into two pieces, $E(A_n)$ which does not depend on t and $t(A_n - E(A_n))$. Since t is bounded over its domain of integration $[0, 1]$, this reduces to an expression like the first remainder term.

Thus, we are left with obtaining the decay rate of $\int (B_n - E(B_n))^{2k} f(\alpha) d\alpha$. We will show that by choosing k, it can be made to decay faster than $(n \det(H_n))^l$ for any l. Note that the case where the exponent is $2k-2$ will follow likewise. For notational simplicity, in the following we replace $2k$ with k alone, noting that it is even.

$$B_n - E(B_n) = \frac{1}{n \det(H_n)} \sum_{i=1}^{n} \left(K\left(H_n^{-1}(X_i - x) \right) - E\left(K\left(H_n^{-1}(X_i - x) \right) \right) \right) = \frac{1}{n \det(H_n)} \sum_{i=1}^{n} G_i.$$

By the Multinomial Theorem, $(B_n - E(B_n))^k = \frac{1}{(n \det(H_n))^k} \sum_{k_1+\ldots+k_n=k} \frac{k!}{k_1!\ldots k_n!} G_1^{k_1} \ldots G_n^{k_n}$, where the k_i are nonnegative integers. Note that $\frac{k!}{k_1!\ldots k_n!} \leq k!$. We separate terms and then group them by combinations

of values of k_i. First, suppose $k_i = 1$. Then $E(G_i)$ can be factored out of the rest of the expectation by independence. But $E(G_i) = 0$, so all terms in the multinomial sum where any $k_i = 1$ are 0.

Consider the terms where $k_i \geq 2$ or $k_i = 0$ $\forall i$. Fix the values but not the permutations of k_i. Then let $S = \{k_1, ..., k_n\}$ be a particular permutation of k_i with those fixed values, and let $r = |S|$. Noting we have at most $\frac{n!}{(n-r)!} < n^r$ terms,

$$\begin{aligned} C \int \frac{1}{(n \det(H_n))^k} \sum_{k_1+\ldots+k_n=k} \prod_{i=1}^{n} G_i^{k_i} f(\alpha) d\alpha &\leq C \frac{1}{(ndet(H_n))^k} \sum_{k_1+\ldots+k_n=k} \prod_{i=1}^{n} \int |G_i^{k_i}| f(\alpha) d\alpha \\ &\leq C \frac{1}{(n \det(H_n))^{k-r}} \prod_{k_i \in S} \int |G_i^{k_i}| f(x + H_n \gamma) d\gamma \\ &= O((n \det(H_n))^{-(k-r)}) \end{aligned}$$

Note that we use independence for the first inequality and the identical distribution assumption for the second. The order uses the fact that the G_i are bounded. r is at most $k/2$ since every nonzero $k_i \geq 2$. The number of different sets of values of k_i is bounded independent of n by $\binom{2k-1}{k-1}$. Thus, by choosing k, the whole remainder can be made to decay faster than any fixed decay rate.

Next, consider the terms of order 2 or greater. Call the order d, so they have the form

$$C_1 \frac{1}{E(B_n)^d} \int (B_n - E(B_n))^{d-1} (A_n - E(A_n)) dP + C_2 \frac{E(A_n)}{E(B_n)^{d+1}} \int (B_n - E(B_n))^d dP$$

The first term corresponds to taking the derivative of the denominator d times. The second term corresponds to taking the derivative of the numerator once and the derivative of the denominator $d-1$ times. The numerator derivative can be permuted d ways. Therefore, $C_1 = -C_2$. We multiply the first term by $\frac{E(B_n)}{E(B_n)}$ for a common denominator and factor out the constant and denominator, looking only at the rest (as before, the denominator is bounded away from 0).

By Taylor's Theorem, conditioning on the X_i,

$$E(A_n) = (mf)(x) + \frac{1}{2} \sum_{i=1}^{D} h_{i,n}^2 (D_i^2 (mf)(x)) \mu_2 + O(\mathrm{tr}(H_n^3))$$

and

$$E(B_n) = f(x) + \frac{1}{2} \sum_{i=1}^{D} h_{i,n}^2 (D_i^2 f(x)) \mu_2 + O(\mathrm{tr}(H_n^3)).$$

Note that these expansions are the same except for the m and its derivatives in the former. Then,

$$\begin{aligned}
&E(B_n)\int (B_n - E(B_n))^{d-1}(A_n - E(A_n))dP - E(A_n)\int (B_n - E(B_n))^d dP \\
&= E(B_n)\int (B_n - E(B_n))^{d-1}(A_n)dP - E(A_n)\int (B_n - E(B_n))^{d-1}(B_n)dP \\
&= E(B_n)\int A_n B_n^{d-1} - (d-1)A_n B_n^{d-2}E(B_n) + ... \pm A_n E(B_n)^{d-1}dP \\
&\quad - E(A_n)\int B_n B_n^{d-1} - (d-1)B_n B_n^{d-2}E(B_n) + ... \pm B_n E(B_n)^{d-1}dP.
\end{aligned}$$

For the first integral, this gives (after conditioning) terms of form

$$E(B_n)^t \int (n\det(H_n))^{-r} K^{d-t}(\gamma) m(x + H_n\gamma) f(x + H_n\gamma) d\gamma,$$

where t corresponds to the term in question and r depends, similar to the remainder, on indices and independence. The order is at worst $O((n\det(H_n))^{-1})$ (from $r = 1$, corresponding to a second moment). The second integral has the same form without the m.

We take a Taylor expansion of $(mf)(x+H_n\gamma)$ to order 3 remainder of each term. The first order part is 0 by symmetry of K; the second and third order parts are $O(\frac{1}{n\det(H_n)})O(\mathrm{tr}(H_n^2))$ and $O(\frac{1}{n\det(H_n)})O(\mathrm{tr}(H_n^3))$ respectively. Therefore, everything is $O(\frac{\mathrm{tr}(H_n^2)}{n\det(H_n)})$ if the 0th order parts are. The 0th order parts are already $O(\frac{1}{n\det(H_n)})$. Therefore when multiplied by $E(B_n)$ or $E(A_n)$ respectively, the second order and beyond terms of those expectations give $O(\frac{\mathrm{tr}(H_n^2)}{n\det(H_n)})$ or better. The 0th order terms of those expectations differ by a factor of m. Since the 0th order parts of the integral also differ by a factor of m, in the opposite order and with opposite sign, they cancel. Thus in the original Taylor expansion, everything of order 2 or higher, including the remainder, is $O(\frac{\mathrm{tr}(H_n^2)}{n\det(H_n)})$.

Finally, consider $\frac{E(A_n)}{E(B_n)} = \frac{(mf)(x)+\frac{1}{2}\sum_{i=1}^D h_{i,n}^2(D_i^2(mf)(x))\mu_2+O(\mathrm{tr}(H_n^3))}{f(x)+\frac{1}{2}\sum_{i=1}^D h_{i,n}^2(D_i^2 f(x))\mu_2+O(\mathrm{tr}(H_n^3))}$. First, because $\frac{1}{1+c} = (1-c) - \frac{c^2}{1+c}$, we can apply this repeatedly to the denominator as a convergent power series with order better than its first term and eliminate the $O(\mathrm{tr}(H_n^3))$ term there (leaving it in the numerator). Note $D_i^2(mf) = mD_i^2 f +$

$2D_imD_if + fD_i^2m$. Then,

$$\begin{aligned}\frac{E(A_n)}{E(B_n)} &= \frac{(mf)(x) + \frac{1}{2}\sum_{i=1}^{D} h_{i,n}^2 (D_i^2(mf)(x))\mu_2 + O(\mathrm{tr}(H_n^3))}{f(x) + \frac{1}{2}\sum_{i=1}^{D} h_{i,n}^2 (D_i^2 f(x))\mu_2} \\ &= \frac{(mf)(x) + \frac{1}{2}\sum_{i=1}^{D} h_{i,n}^2 (mD_i^2 f + 2D_imD_if + fD_i^2 m)(x)\mu_2}{f(x) + \frac{1}{2}\sum_{i=1}^{D} h_{i,n}^2 (D_i^2 f(x))\mu_2} + O(\mathrm{tr}(H_n^3)) \\ &= m(x) + \frac{1}{2}\sum_{i=1}^{D} h_{i,n}^2 \mu_2 \left(\frac{2(D_imD_if)(x)}{f(x)} + D_i^2 m(x) \right) + O(\mathrm{tr}(H_n^3))\end{aligned}$$

where the last equality uses another application of the $\frac{1}{1+c}$ identity. Thus, the bias is

$$\frac{\mu_2}{2}\sum_{i=1}^{D} h_{i,n}^2 \left(\frac{2(D_imD_if)(x)}{f(x)} + D_i^2 m(x) \right) + O\left(\mathrm{tr}(H_n^3) + \frac{\mathrm{tr}(H_n^2)}{n \det(H_n)} \right).$$

□

2.1 Bounding $E(1/B_n^k)$

Unfortunately, we have not yet bounded this term successfully. Here, we outline the main attempt so far, followed by some other ideas. We start by stating two related results that are used:

Lemma 1. *Let $a_{0,n} = 0$ and let a_{mn}, $m \geq 1$ and $n \geq 1$, be a positive double sequence that is monotone increasing in m. Suppose also the difference $c_{mn} = a_{mn} - a_{m-1,n}$ is monotone increasing in n. Then the limit can be taken in either order, that is,*

$$\lim_{n\to\infty} \lim_{m\to\infty} a_{mn} = \lim_{m\to\infty} \lim_{n\to\infty} a_{mn}$$

Proof. Because a_{mn} is monotone increasing in m, $c_{mn} \geq 0$. Let μ be the counting measure. Since c_{mn} is monotone increasing in n, by the monotone convergence theorem, $\lim_{n\to\infty} \int c_{mn} d\mu = \int \lim_{n\to\infty} c_{mn} d\mu$. This integral is an infinite sum, so

$$\lim_{n\to\infty} \lim_{m\to\infty} \sum_{q=0}^{m} c_{qn} = \lim_{m\to\infty} \lim_{n\to\infty} \sum_{q=0}^{m} c_{qn}$$

Every term in the sum cancels except the first and last. The first is 0 as defined, and the last is a_{mn}. So the limits can be interchanged. □

Corollary 2.1.1. *Suppose a_{mn} is a double sequence as above, but now the difference c_{mn} is monotone decreasing in n. If $\lim_{m\to\infty} a_{m,1}$ exists, then we can again take the limits in either order. In addition, if c_{mn} is always negative, we can still swap the limit (by factoring out -1).*

Proof. This follows like Lemma 1 by a standard corollary for decreasing sequences to the monotone convergence theorem. To apply this corollary, the first term in the monotone sequence must be integrable. This requires the additional condition on the limit of $a_{m,1}$. □

We now attempt to deal with $E(1/B_n^k)$ using an identity from the integral of an exponential function,

$$\lim_{n\to\infty} E\left(\frac{1}{B_n^k}\right) = \lim_{n\to\infty} E\left(\lim_{m\to\infty}\int_0^{b_m} e^{-\lambda B_n^k} d\lambda\right)$$

where $b_m \to \infty$ monotonically as $m \to \infty$. The inner integral is monotonically increasing in m and positive, so by the monotone convergence theorem,

$$\begin{aligned}
\lim_{n\to\infty} E\left(\lim_{m\to\infty}\int_0^{b_m} e^{-\lambda B_n^k} d\lambda\right) &= \lim_{n\to\infty}\lim_{m\to\infty} E\left(\int_0^{b_m} e^{-\lambda B_n^k} d\lambda\right) \\
&= \lim_{n\to\infty}\lim_{m\to\infty}\int_0^{b_m} E\left(e^{-\lambda B_n^k}\right) d\lambda \\
&= \lim_{n\to\infty}\lim_{m\to\infty} a_{mn}
\end{aligned}$$

where the second equality follows by Tonelli's theorem. Note a_{mn} is positive and increasing in m. Let $c_{mn} = a_{mn} - a_{m-1,n}$. Without loss of generality, let the indexing start for c_{mn} start at $n = m = 1$ and set $a_{0,n} = 0$. By Taylor's Theorem $e^{-\lambda B_n^k} = \sum_{i=0}^{\infty} \frac{(-\lambda B_n^k)^i}{i!}$. Note that taking the absolute value of each term gives the series for $e^{\lambda B_n^k}$, so this sum is absolutely convergent. Therefore, Fubini's Theorem,

$$c_{mn} = \int_{b_{m-1}}^{b_m} E\left(e^{-\lambda B_n^k}\right) d\lambda = \int_{b_{m-1}}^{b_m} \sum_{i=0}^{\infty} (-1)^i \frac{E(B_n^{ki})}{i!}$$

We now examine $E(B_n^{ki}) = E(B_n^r)$, with r = 2, 3, ... Suppose $n \geq r$. Then,

$$\begin{aligned}
E(B_n^r) &= \frac{1}{(n\det(H_n))^r} E\left(\sum_{|\alpha|} \frac{r!}{\alpha!} K^{\alpha_1}\big(H^{-1}(X_1 - x)\big) \cdots K^{\alpha_n}\big(H^{-1}(X_n - x)\big) \right) \\
&= \frac{n}{(n\det(H_n))^r} \int K^r(H_n^{-1}(\alpha - x)) f(\alpha) d\alpha \\
&\quad + \frac{n(n-1)r!}{(r-1)!(n\det(H_n))^r} \int K^{r-1}(H_n^{-1}(\alpha - x)) f(\alpha) d\alpha \int K(H_n^{-1}(\alpha - x)) f(\alpha) d\alpha \\
&\quad + \cdots + \frac{n(n-1)\cdots(n-r+1)}{(n\det(H_n))^r} \left(\int K(H_n^{-1}(\alpha - x)) f(\alpha) d\alpha \right)^r \\
&= \frac{1}{(n\det(H_n))^{k-1}} \int K(\gamma) f(x + H_n\gamma) d\gamma + \cdots + \left(\int K(\gamma) f(x + H_n\gamma) d\gamma \right)^r + O(n^{-1})
\end{aligned}$$

where the second equality uses the fact the X_i are i.i.d., and the third is by substitution. The $O(n^{-1})$ accounts for the mismatch between the factors of form $n(n-1)(n-2)...$ in the numerators and n^r in the denominator. Once more, by Taylor's Theorem,

$$f(x + H_n\gamma) = f(x) + \sum_{|\alpha|=1,2} \frac{1}{\alpha!} (H_n\gamma)^\alpha D^\alpha f(x) + \sum_{|\alpha|=3} \frac{1}{\alpha!} (H_n\gamma)^\alpha D^\alpha f(x + \zeta H_n\gamma)$$

for some $\zeta \in [0, 1]$. Suppose we make an additional assumption that the order of decay rates is fixed. For example, suppose each $h_{i,n}$ is a function of form $h_{i,n} = C_i n^{-v}$, $v > 0$. Let h_n^* be the slowest decaying $h_{i,n}$ $(i = 1, ..., d)$, let $h_n = (h_{1,n}\ ...\ h_{D,n})$, and let $\nabla^2 f(x) = diag\{\frac{\partial^2 f(x)}{\partial x_d^2}\}_{d=1}^D$. Substituting in $E(B_n^r)$ above,

$$E(B_n^r) = f^r(x) + \frac{r}{2} h_n^T f^{r-1}(x) \nabla^2 f(x) h_n + O((h_n^*)^3) + O(n^{-1})$$

Note that the big-O terms depend on r, but the coefficients involving r in those terms do not exceed r!, and the powers of $f(x)$ do not exceed $f^r(x)$. Add the additional assumption that $nh_{d,n}^2 \to \infty \quad \forall d$. Then we can rewrite the $O(n^{-1})$ term as $O((h_n^*)^2 \psi_n)$ with a sequence $\psi_n \to 0$ monotonically. Thus, there exists $N_1 > r!$ such that for $n > N_1$, the change in the second term when incrementing from n to $n + 1$ dominates the change in the big-O terms. Because the ordering of decay rates is fixed by assumption 4, there exists $N_2 > N_1$ such that for all $n > N_2$, the change in the second term is monotonic.

Now consider $\sum_{i=0}^{\infty} (-1)^i \frac{E(B_n^{ki})}{i!}$. We truncate the sum as $\sum_{i=0}^{g_q} (-1)^i \frac{E(B_n^{ki})}{i!}$, where g_q is the sequence of odd integers 1, 3, 5, ... Now, $\sum_{i=0}^{g_q} \frac{w^i}{i!}$ has derivative $S_q(w) = \sum_{i=0}^{g_q - 1} \frac{w^i}{i!}$, and since g_q is odd, $g_q - 1$ is even. For $w < 0$, consider the terms present in the exponential function that are missing in S_q. The first

term has odd power. If we factor that term out, we get a product of that term and the normal exponential function $\sum_{i=0}^{\infty} \frac{x^i}{i!}$. The exponential function is positive, and the other factor is negative for all $w < 0$. So their product is negative. Now, if S_q took on a negative value for some w and some q, then S_q plus the missing terms would also be negative. But S_q plus missing terms is again the exponential function, so it cannot be negative. Therefore $S_q(w)$ must be nonnegative for all q and all $w < 0$. Since the derivative is nonnegative, $\sum_{i=0}^{g_q} \frac{w^i}{i!}$ is monotonic in w. Since the composition of two monotone functions is also monotone, $\sum_{i=0}^{g_q} \frac{(-\lambda B_n^k)^i}{i!}$ is monotonic in n.

Suppose we add two more terms to the truncated sum, that is, we go from g_q to $g_q + 2$. For some Q, and $q > Q$, the sum of these two terms is positive: the first term is positive, and it is larger in magnitude than the second because the factorial in the denominator dominates the polynomial in the numerator. So the sum $\sum_{i=0}^{g_q}(-1)^i \frac{E(B_n^{ki})}{i!}$ is eventually monotone increasing in q.

We now try to put these results together to show that

$$\lim_{n\to\infty}\lim_{m\to\infty}\int_0^{b_m}\lim_{q\to\infty}\sum_{i=0}^{g_q}\frac{(-\lambda)^i E(B_n^{ki}))}{i!}d\lambda = \lim_{m\to\infty}\int_0^{b_m}\lim_{q\to\infty}\sum_{i=0}^{g_q}\frac{(-\lambda)^i \lim_{n\to\infty} E(B_n^{ki}))}{i!}d\lambda$$

First, we take the limit in q out. It goes outside the $d\lambda$ integral by dominated convergence, since it is dominated by its limit, which is integrable on the compact integration domain (m here is fixed). Next, the whole function is increasing in q (by above). The difference in q is the last two terms, whose sum as above is positive, and therefore increasing in m because of the larger integration domain. So by Lemma 1, we can interchange the limit in q and the limit in m. However, after taking m inside, integrating and taking the limit in m gives infinity. So this does not seem to work.

2.1.1 Other Possibilities

It is possible that there is nonetheless a way to swap the limits in n and m. Aside from mistakes in the above, a tighter bound might replace one of the inequalities and permit the swap. This would complete the proof. However, given the number of limits, integrals, and Taylor expansions involved, and the past attempts, if there is a solution here at all it is difficult to obtain.

The most promising route at the moment seems to be to explore the identity used by Ziegler (2001):

$$E\Big(\frac{A}{B}\Big) = \frac{E(A)}{E(B)} - \frac{1}{(E(B))^2}E(A(B - E(B))) + \frac{1}{(E(B))^2}E\Big(\frac{A}{B}(B - E(B))^2\Big)$$

There are two possibilities here. Since Ziegler doesn't use differentiability assumptions, it is possible that by using that identity directly but modifying other parts of Ziegler's proof, likely by using Taylor expansions to directly incorporate the differentiability, one could get a better result than Ziegler does. The second possibility is related to the order k of the expansion in Theorem 2.1 above. The arbitrary order ensures that the remainder decays fast enough not to conflict with decay in other terms. If one could increase the order of this identity, or perhaps derive a general version, it might be possible to improve on Ziegler's order or relax his assumptions, perhaps even without assuming differentiability.

Another avenue to investigate is a simulation to try to pin down the actual order. Such a simulation could not be used as any sort of proof. But since there is disagreement in the literature on the actual order, a simulation could show what result to aim for, and head off fruitless attempts to prove a nonexistent faster convergence.

3 Conclusion

In this note we attempt to provide an expression for the unconditional bias of the traditional Nadaraya-Watson estimator for a multivariate regression. Our proof clarifies some issues in the literature. Besides their importance from a purely technical perspective, these issues are important to address in developing alternatives to the Nadaraya-Watson estimator that try to reduce the leading terms of the bias. Unfortunately, we have so far been unable to complete the proof. However, some ideas that may lead to resolving the remaining problems are presented above, and we are hopeful that developing them further will lead to a complete proof.

References

Choi, E., Hall, P., Rousson, V., 2000. Data Sharpening Methods for Bias Reduction in Nonparametric Regression. The Annals of Statistics 28 (5), 1339–1355.
URL http://www.jstor.org/stable/2674096

Chu, C.-K., Marron, J. S., Nov. 1991. Choosing a Kernel Regression Estimator. Statistical Science 6 (4), 404–419.
URL https://projecteuclid.org/euclid.ss/1177011586

Collomb, G., 1976. Estimation non paramtrique de la rgression par la mthode du noyau. Universit Paul Sanatier, Toulouse, France.

Collomb, G., 1977. Quelques proprits de la mthode du noyau pour l'estimation non paramtrique de rgression en un point fix 285(368), 289–292.

Collomb, G., 1981. Estimation Non-paramétrique de la Régression: Revue Bibliographique. International Statistical Review / Revue Internationale de Statistique 49 (1), 75–93.
URL http://www.jstor.org/stable/1403039

Fan, J., 1992. Design-adaptive Nonparametric Regression. Journal of the American Statistical Association 87 (420), 998–1004.
URL http://www.jstor.org/stable/2290637

Glad, I. K., 1998. A note on unconditional properties of a parametrically guided nadaraya-watson estimator. Statistics & Probability Letters 37, 101–108.

Jennen-Steinmetz, C., Gasser, T., 1988. A Unifying Approach to Nonparametric Regression Estimation. Journal of the American Statistical Association 83 (404), 1084–1089.
URL http://www.jstor.org/stable/2290140

Mack, Y. P., Müller, H.-G., Dec. 1988. Convolution type estimators for nonparametric regression. Statistics & Probability Letters 7 (3), 229–239.
URL http://www.sciencedirect.com/science/article/pii/0167715288900569

Nadaraya, E. A., 1964. Some new estimates for distribution functions. Theory of Probability and Applications 15 (1), 497–500.

Pagan, A., Ullah, A., 1999. Nonparametric econometrics. Cambridge University Press, Cambridge, UK.

Rosenblatt, M., 1969. Multivariate Analysis II. Wiley, New York, NY.

Schimek, M. G., May 2013. Smoothing and Regression: Approaches, Computation, and Application. John Wiley & Sons, google-Books-ID: ffa0evMuDNoC.

Scott, D. W., 2015. Multivariate Density Estimation: Theory, Practice, and Visualization. John Wiley & Sons, New York, NY.

Watson, G. S., 1964. Smooth regression analysis. Sankhya A 26, 359–372.

Ziegler, K., 2001. On approximations to the bias of the nadaraya-watson regression estimator. Journal of Nonparametric Statistics 13, 583–589.