Estimating a density under pointwise constraints on the derivatives

Ursula U. Müller
Texas A&M University

Wolfgang Wefelmeyer
Universität zu Köln

Abstract

Suppose we want to estimate a density at a point where we know the values of its first or higher order derivatives. In this case a given kernel estimator of the density can be modified by adding appropriately weighted kernel estimators of these derivatives. We give conditions under which the modified estimators are asymptotically normal. We also determine the optimal weights. When the highest derivative is known to vanish at a point, then the bias is asymptotically negligible at that point and the asymptotic variance of the kernel estimator can be made arbitrarily small by choosing a large bandwidth.

1 Introduction

Consider a point x on the real line and let f be a density that is r times continuously differentiable at x. Suppose we have n independent observations $X_1, \dots, X_n$ with density f. Then $f(x)$ can be estimated with an estimator $\hat{f}(x)$ based on a kernel K_0 of order r. If a bandwidth of order $n^{-1/(2r+1)}$ is used, the estimator $\hat{f}(x)$ will converge at the optimal rate $n^{-r/(2r+1)}$. This goes back to Rosenblatt (1956) and Parzen (1962). Moreover, $n^{r/(2r+1)}(\hat{f}(x) - f(x))$ is asymptotically normal.

Let us extend this statement to (simultaneously) estimating f and its derivatives $f^{(j)}$ at a point x. For this we work with the optimal bandwidth $b = n^{-1/(2r+1)}$ and with j times continuously differentiable kernels K_j of order $r - j$, $j = 0, \dots, r$. Then $f(x)$ and its derivatives $f^{(j)}(x)$ can be estimated with kernel estimators

$$\hat{f}^{(j)}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{b^{j+1}} K_j^{(j)}\Big(\frac{x - X_i}{b}\Big), \qquad j = 0, \dots, r.$$

Each estimator $\hat{f}^{(j)}(x)$ has the optimal rate $n^{-(r-j)/(2r+1)}$ for estimating $f^{(j)}(x)$. We show as a first result (see Proposition 1) that the joint distribution of $n^{(r-j)/(2r+1)}(\hat{f}^{(j)}(x) - f^{(j)}(x))$, $j = 0, \dots, r$, is asymptotically normal, and calculate the asymptotic mean vector and the covariance matrix.

Suppose now that we have auxiliary information in the form of pointwise constraints on the derivatives. This means, for example, that we know the values of some derivatives at x

or, more generally, that certain linear combinations are zero, i.e. $A(f'(x)-a_1,\dots,f^{(r)}(x)-a_r)^\top=0$ for a known matrix A and a known vector $(a_1,\dots,a_r)^\top$. We can then introduce new estimators for $f(x)$ by modifying $\hat{f}(x)$ as follows,

$$\hat{f}_c(x)=\hat{f}(x)-c^\top A\begin{pmatrix} n^{-1/(2r+1)}(\hat{f}'(x)-a_1) \\ \vdots \\ n^{-r/(2r+1)}(\hat{f}^{(r)}(x)-a_r)\end{pmatrix},$$

where c is a vector of constants. We show in Remark 1 that $\hat{f}_c(x)$ can be written as a kernel estimator with bandwidth $b=n^{-1/(2r+1)}$ and kernel $\tilde{K}=K_0-c^\top A(K_1'-a_1,\dots,K_r^{(r)}-a_r)^\top$. This representation makes it easy to use Proposition 1 to establish our main result, the limiting normality of $n^{r/(2r+1)}(\hat{f}_c(x)-f(x))$, which is provided in Theorem 1.

The new estimator $\hat{f}_c(x)$ exploits the auxiliary information and should therefore outperform the ordinary kernel estimator $\hat{f}(x)$, or, at least, be as good as $\hat{f}(x)$. We distinguish two cases. In the first case the constraint implies that the highest derivative of the density vanishes at x, i.e., $f^{(r)}(x)=0$. Then the asymptotic bias of $n^{r/(2r+1)}(\hat{f}(x)-f(x))$ vanishes (see Lemma 1), and the asymptotic MSE equals the asymptotic variance. In this case, we do not need the 'corrected' estimator $\hat{f}_c(x)$. We show that the best kernel with support $[-s,s]$ is the uniform kernel. The asymptotic MSE can then be made arbitrarily small by choosing s large. In the second case, $f^{(r)}(x)$ is not known to be zero. Then we determine the vector c that minimises the MSE of $n^{r/(2r+1)}(\hat{f}_c(x)-f(x))$.

The main applications are to cases in which we know that certain derivatives are zero at some known point x. For example, the density may have a maximum there, $f'(x)=0$; an inflection point, $f''(x)=0$; or a saddle point, $f'(x)=0$ and $f''(x)=0$. The important special case where $f'(x)$ is zero (or known) is discussed in more detail in Examples 2–4. When we know at which point the density has a maximum or a saddle point, we may also know that it is symmetric (and perhaps bimodal) around this point. This information can be used to improve $\hat{f}_c(x)$ further, by symmetrisation.

The approach described here is not restricted to kernel estimators. Similar improvements can be obtained for other types of density estimator and for combinations of different types of density estimator. The main tool is a result on the pointwise joint asymptotic normality of estimators for the density and some of its derivatives. The approach also extends to multivariate density estimation and to density estimation for dependent data.

We have restricted ourselves to estimating a density under constraints on its derivatives. Similar results can be obtained for estimators of some derivative under constraints on derivatives of higher or lower order.

The idea behind the modification $\hat{f}_c(x)$ of $\hat{f}(x)$ is that the variance may be reduced if we add to $\hat{f}(x)$ an estimator of zero that is correlated to $\hat{f}(x)$. This idea is similar to an additive improvement of *empirical* estimators under *linear* constraints on the underlying distribution. We briefly describe this. Let X have unknown distribution P. Suppose that $f(X)$ is real-valued and square-integrable under P, and that $g(X)$ is r-dimensional with P-square-integrable components. Assume that $Pg=E[g(X)]=0$ constitutes a linear constraint on

P. Let $X_1, \dots, X_n$ be independent copies of X. The best nonparametric estimator of Pf is the empirical estimator $\mathbb{P}f = (1/n)\sum_{i=1}^n f(X_i)$. Its asymptotic variance is $P(f^2) - (Pf)^2$. The constraint $Pg = 0$ gives an r-dimensional 'estimator' $\mathbb{P}g = (1/n)\sum_{i=1}^n g(X_i)$ of zero. It can be combined with $\mathbb{P}f$ to obtain an estimator of the form

$$\mathbb{P}f - c^\top \mathbb{P}g = \frac{1}{n}\sum_{i=1}^n (f(X_i) - c^\top g(X_i)).$$

Such an estimator has asymptotic variance $P(f - Pf - c^\top g)^2$. It is minimised for $c = c^*(P) = (P(gg^\top))^{-1}P(gf)$. Hence the minimal asymptotic variance is

$$P(f^2) - (Pf)^2 - P(fg^\top)(P(gg^\top))^{-1}P(gf).$$

This is strictly smaller than the asymptotic variance $P(f^2) - (Pf)^2$ of $\mathbb{P}f$ unless f and g are uncorrelated. Since $c^*(P)$ depends on P it must be replaced by an estimator, say $c^*(\mathbb{P})$. This does not change the asymptotic variance. Levit (1975) shows that

$$\mathbb{P}f - c^*(\mathbb{P})^\top \mathbb{P}g = \frac{1}{n}\sum_{i=1}^n f(X_i) - \sum_{i=1}^n f(X_i)g^\top(X_i)\Big(\sum_{i=1}^n g(X_i)g^\top(X_i)\Big)^{-1}\frac{1}{n}\sum_{i=1}^n g(X_i)$$

is asymptotically efficient. Müller and Wefelmeyer (2002) consider constraints with $g = g_\vartheta$ depending on an unknown finite-dimensional parameter ϑ.

An asymptotically equivalent improvement of $\mathbb{P}f$ is obtained by using *empirical likelihood*. It replaces the empirical distribution $(1/n)\sum_{i=1}^n \delta_{X_i}$ by a weighted version that obeys the linear constraint; see Owen (1988), (2001).

Our problem of estimating a density f at x under a *pointwise* constraint differs from the problem of estimating $f(x)$ under a *linear* constraint on f, say $E[g(X)] = 0$ for some known function g. Such a linear constraint leads to an improvement of order $n^{-1/2}$. The density estimator $\hat{f}(x)$ converges at a slower rate. Hence the improvement vanishes asymptotically; see e.g. Zhang (1998), who demonstrates (first order) equivalence of a standard kernel estimator and a modified version that uses a linear constraint.

The next section contains our main results, in particular the limiting normality of the new estimator $\hat{f}_c(t)$. The proofs are in Section 3.

2 Results

Let $X_1, \dots, X_n$ be real random variables with bounded density f. Fix a point x on the real line. Let r be a natural number. Assume that f is r times continuously differentiable at x.

Denote by $\mathcal{K}_{j,s}$ the set of all functions K on the real line that vanish outside a compact set and have bounded and continuous derivatives up to the order j, and that are (signed) kernels *of order* s, i.e. $\int K(t)\,dt = 1$, $\int t^i K(t)\,dt = 0$ for $i = 1, \dots, s-1$, and $\int t^s K(t)\,dt \neq 0$.

For a function K on the real line and a positive bandwidth b, introduce the scaling $K_b(x) = K(x/b)/b$. Note that we can rescale the kernel by multiplying the bandwidth with a positive constant c. This corresponds to replacing K in the definition of K_b by K_c: we have $K_{cb}(x) = K(x/(cb))/(cb) = K_c(x/b)/b$. In the following we will also need the derivatives of K_b which, for appropriately differentiable K, are

$$K_b^{(j)}(x) = \partial_x^j K_b(x) = \frac{1}{b^{j+1}} K^{(j)}\Big(\frac{x}{b}\Big).$$

Let $K_0 \in \mathcal{K}_{0,r}$ be a kernel and b_0 a bandwidth. We estimate $f(x)$ by the kernel estimator

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K_{0b_0}(x - X_i).$$

To estimate the various derivatives, we may use different kernels and bandwidths. For $j = 1, \ldots, r$ let b_j be a bandwidth and $K_j \in \mathcal{K}_{j,r-j}$. Set

$$\hat{f}^{(j)}(x) = \frac{1}{n} \sum_{i=1}^{n} K_{jb_j}^{(j)}(x - X_i).$$

The following lemma describes approximations for the bias and the variance of $\hat{f}^{(j)}(x)$. The result is essentially known. See Bhattacharya (1967), Schuster (1969) and Singh (1977, 1981). We indicate the proof in Section 3.

Lemma 1. *Let f be r times continuously differentiable at x. For $j = 0, \ldots, r$ let $K_j \in \mathcal{K}_{j,r-j}$, and let $b_j \to 0$ and $nb_j \to \infty$. Then*

$$b_j^{-r+j}(E[\hat{f}^{(j)}(x)] - f^{(j)}(x)) = f^{(r)}(x) \frac{(-1)^{r-j}}{(r-j)!} \int t^{r-j} K_j(t)\, dt + o(1),$$

$$nb_j^{2j+1} \operatorname{Var} \hat{f}^{(j)}(x) = f(x) \int K_j^{(j)2}(t)\, dt + o(1).$$

For $j = 0, \ldots, r$ the rate of $\hat{f}^{(j)}(x)$ is optimal if the variance converges at the same rate as the squared bias. This holds if $b_j^{-2(r-j)} \sim nb_j^{2j+1}$, i.e. $b_j \sim n^{-1/(2r+1)}$. In the following we set

$$b_j = b = n^{-1/(2r+1)}$$

and absorb a possible positive factor of the bandwidth as a scale parameter in K_j. The following proposition shows that the joint distribution of $n^{(r-j)/(2r+1)}(\hat{f}^{(j)}(x) - f^{(j)}(x))$, $j = 0, \ldots, r$, is asymptotically normal. Set

$$V_n = \begin{pmatrix} n^{r/(2r+1)}(\hat{f}(x) - f(x)) \\ n^{(r-1)/(2r+1)}(\hat{f}'(x) - f'(x)) \\ \vdots \\ \hat{f}^{(r)}(x) - f^{(r)}(x) \end{pmatrix}$$

and $\mu = (\mu_0, \mu_1, \ldots, \mu_r)^\top$ with

$$\mu = \begin{pmatrix} \frac{(-1)^r}{r!} \int t^r K_0(t)\, dt \\ \frac{(-1)^{r-1}}{(r-1)!} \int t^{r-1} K_1(t)\, dt \\ \vdots \\ \int K_r(t)\, dt \end{pmatrix}.$$

Define $\Sigma = (\sigma_{jk})_{j,k}$ with

$$\sigma_{jk} = \int K_j^{(j)}(t) K_k^{(k)}(t)\, dt, \qquad j, k = 0, \dots, r.$$

Proposition 1. *Let f be r times continuously differentiable at x. For $j = 0, \dots, r$ let $K_j \in \mathcal{K}_{j,r-j}$. Then V_n is asymptotically normal with mean vector $f^{(r)}(x)\mu$ and covariance matrix $f(x)\Sigma$.*

A similar result for polynomial estimators of regression functions is in Masry and Fan (1997); see also Fan and Yao (2003), Theorem 5.2. Analogous results hold for time series. Univariate asymptotic normality for density estimators in time series is proved in Bradley (1983) and Lu (2001). In the following we write briefly $\int t^k K_j^m$ instead of $\int t^k K_j^m(t)\, dt$.

Example 1. (Vanishing highest derivative.) Suppose we have a constraint $A(f^{(\cdot)}(x) - a) = 0$, where $f^{(\cdot)}$ denotes the vector of derivatives, $f^{(\cdot)} = (f', \dots, f^{(r)})^\top$, A is some known matrix and a a known vector. We first address the case in which the constraint implies that the highest derivative of the density vanishes at x, i.e. $f^{(r)}(x) = 0$. This is a special case where the ordinary kernel density estimator $\hat{f}(x)$ for estimating $f(x)$ cannot be improved: the asymptotic bias of $n^{r/(2r+1)}(\hat{f}(x) - f(x))$ is $f^{(r)}(x)\mu_0$ (see Lemma 1), i.e. it vanishes. The asymptotic MSE of $\hat{f}(x)$ therefore equals the asymptotic variance $f(x)\sigma_{00} = f(x)\int K_0^2$. Let us suppose that K_0 is supported by the bounded interval $[-s, s]$. Then $\int K_0^2$ is minimised by the box kernel $K_0 = B_s = (2s)^{-1}\mathbf{1}[-s, s]$. This follows from the Cauchy inequality

$$1 = \int K_0 = \int \mathbf{1}[-s,s] K_0 \le \Big(\int \mathbf{1}[-s,s]^2 \int K_0^2 \Big)^{1/2} = (2s)^{1/2} \Big(\int K_0^2 \Big)^{1/2},$$

which implies $\int B_s^2 = (2s)^{-1} \le \int K_0^2$. (This is plausible because the best nonparametric estimator of an expectation is the *unweighted* sample mean, and here we estimate the expectation of $(2bs)^{-1}\mathbf{1}[x - bs, x + bs](X)$.) This means that we can make the asymptotic MSE of $\hat{f}(x)$ arbitrarily small by taking $K_0 = B_s$ with s large. $\square$

In the following we address the general case in which $f^{(r)}(x)$ is not known to vanish. For an arbitrary vector c we can then introduce a 'corrected' estimator $\hat{f}_c(x)$ for $f(x)$,

$$\hat{f}_c(x) = \hat{f}(x) - c^\top A \begin{pmatrix} n^{-1/(2r+1)}(\hat{f}'(x) - a_1) \\ \vdots \\ n^{-r/(2r+1)}(\hat{f}^{(r)}(x) - a_r) \end{pmatrix}.$$

Remark 1. (Alternative presentation of the estimator.) Set $K^{(\cdot)} = (K_1', \dots, K_r^{(r)})^\top$. We can write $\hat{f}_c(x)$ as an ordinary kernel estimator

$$\hat{f}_c(x) = \frac{1}{n} \sum_{i=1}^n \tilde{K}_b(x - X_i)$$

with bandwidth $b = n^{-1/(2r+1)}$ and kernel $\tilde{K} = K_0 - c^\top A(K^{(\cdot)} - a)$. $\square$

Write μ and Σ from Lemma 1 as

$$\mu = \begin{pmatrix} \mu_0 \\ \nu \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} \sigma_{00} & \lambda^\top \\ \lambda & \Lambda \end{pmatrix},$$

where $\nu = (\mu_1, \dots, \mu_r)^\top$, $\lambda = (\sigma_{10}, \dots, \sigma_{r0})^\top$ and $\Lambda = (\sigma_{jk})_{j,k}$, $j, k = 1, \dots, r$. Introduce the vector $c_- = (1, -c^\top)^\top$ and the diagonal block matrix $A_- = \operatorname{diag}(1, -A)$.

Theorem 1. *Let f be r times continuously differentiable at x. For $j = 0, \dots, r$ let $K_j \in \mathcal{K}_{j,r-j}$. Assume that $A(f^{(\cdot)}(x) - a) = 0$ holds. Then*

$$n^{r/(2r+1)}(\hat{f}_c(x) - f(x)) = c_-^\top A_- V_n$$

is asymptotically normal with mean $f^{(r)}(x)(\mu_0 - c^\top A\nu)$ and variance

$$f(x)\big(\sigma_{00} - 2c^\top A\lambda + c^\top A\Lambda A^\top c\big).$$

Remark 2. (Optimal choice of c.) It follows from Theorem 1 that the asymptotic MSE of $n^{r/(2r+1)}(\hat{f}_c(x) - f(x))$ is

$$\begin{aligned} &f^{(r)2}(x)(\mu_0 - c^\top A\nu)^2 + f(x)\big(\sigma_{00} - 2c^\top A\lambda + c^\top A\Lambda A^\top c\big) \\ &= A - 2c^\top B + c^\top C c \end{aligned}$$

with

$$\begin{aligned} A &= f(x)\sigma_{00} + f^{(r)2}(x)\mu_0^2, \\ B &= f(x)A\lambda + f^{(r)2}(x)\mu_0 A\nu, \\ C &= f(x)A\Lambda A^\top + f^{(r)2}(x)A\nu\nu^\top A. \end{aligned}$$

The matrix C is symmetric. If $C \neq 0$, the asymptotic MSE is minimised by $c = c_* = C^{-1}B$. The minimal asymptotic MSE is $A - B^\top C^{-1}B$. Both B and C depend on the density through $f(x)$ and $f^{(r)}(x)$. Write $\hat{c}_*$ for c_* with $f(x)$ and $f^{(r)}(x)$ replaced by $\hat{f}(x)$ and $\hat{f}^{(r)}(x)$. It follows that $n^{r/(2r+1)}(\hat{f}_{\hat{c}_*}(x) - f(x))$ also has asymptotic MSE $A - B^\top C^{-1}B$. The original estimator $\hat{f}(x)$ is $\hat{f}_c(x)$ with $c = 0$. Hence the asymptotic MSE of $n^{r/(2r+1)}(\hat{f}(x) - f(x))$ is A, which is strictly larger than $A - B^\top C^{-1}B$. □

Example 2. (Vanishing derivative; r=1.) Suppose the density f is continuously differentiable at x and we want to estimate $f(x)$ under the constraint $f'(x) = 0$. This is a special case of the situation in Example 1, with $r = 1$. We treat it now with Theorem 1. As bandwidth with optimal rate we take $b = n^{-1/3}$. We can improve $\hat{f}(x)$ by using modified estimators of the form $\hat{f}_c(x) = \hat{f}(x) - cn^{-1/3}\hat{f}'(x)$. The joint asymptotic distribution of $n^{1/3}(\hat{f}(x) - f(x))$ and $\hat{f}'(x) - f'(x)$ is normal with variances $f(x)\int K_0^2$ and $f(x)\int K_1'^2$ and covariance $f(x)\int K_0K_1'$, and with bias $f'(x)(-\int tK_0, \int K_1)^\top = 0$. Then $n^{1/3}(\hat{f}_c(x) - f(x))$ is asymptotically normal with mean 0 and variance

$$f(x)\Big(\int K_0^2 + c^2 \int K_1'^2 - 2c\int K_0K_1'\Big).$$

This is therefore the asymptotic MSE. It is minimal for $c = c_* = \int K_0K_1'/\int K_1'^2$. The minimal asymptotic MSE is

$$f(x)\Big(\int K_0^2 - \frac{(\int K_0K_1')^2}{\int K_1'^2}\Big).$$

This is smaller than $f(x)\int K_0^2$ unless $\int K_0K_1' = 0$.

Suppose K_0 is supported by the bounded interval $[-s,s]$. We minimise the asymptotic MSE over K_1 by choosing K_1 with support $[-s,s]$ such that K_1' is close to K_0. We must have $\int K_1' = 0$. This holds for the choice $K_1' = K_0 - B_s$, where $B_s = (2s)^{-1}\mathbf{1}[-s,s]$ is again the box kernel on $[-s,s]$. This means that $K_1(t) = (L_0(t) - t + v)\mathbf{1}[-s,s](t)$, where L_0 is an antiderivative of K_0 and the constant v is chosen such that

$$1 = \int K_1 = \int L_0 - \int_{-s}^{s} t + \int_{-s}^{s} v = \int L_0 - s^2 + 2sv,$$

which holds for $v = s/2 - \int L_0/(2s)$. By Remark 2, with $\int K_0 = 1$,

$$c_* = \frac{\int K_0K_1'}{\int K_1'^2} = \frac{\int K_0(K_0 - B_s)}{\int (K_0 - B_s)^2} = \frac{\int K_0^2 - 1/(2s)}{\int K_0^2 - 2/(2s)) + 1/(2s)} = \frac{\int K_0^2 - 1/(2s)}{\int K_0^2 - 1/(2s)} = 1.$$

The asymptotic MSE is

$$f(x)\Big(\int K_0^2 - \frac{(\int K_0K_1')^2}{\int K_1'^2}\Big) = f(x)\Big(\int K_0^2 - \Big(\int K_0^2 - 1\Big)\Big) = f(x).$$

This is the asymptotic MSE of $n^{1/3}(\hat f(x) - f(x))$ with $K_0 = B_s$. Indeed, by Remark 1, the kernel of $\hat f_{c_*}(x) = \hat f_1(x)$ is $\tilde K = K_0 - c_*(K_0 - B_s) = K_0 - (K_0 - B_s) = B_s$, which is the optimal kernel for $\hat f(x)$ by Example 1. □

Remark 3. (Kernel choice.) It is important to choose different kernels for different derivatives. If we take K_1 equal to K_0 in Example 2 and write K for this kernel, then $\int K^2 = \int K^2(t+u)\,dt$ implies

$$0 = \partial_u \int K^2(t+u)\,dt = 2\int (KK')(t+u)\,dt = 2\int KK',$$

and there is no improvement over $\hat f(x)$. □

Now we consider two examples with constraints that do not imply that the highest derivative of f vanishes at x.

Example 3. (Nonvanishing derivative; r=1.) In the simplest example with nonvanishing highest-order derivative, the density has one continuous derivative at x, i.e. $r = 1$ as in Example 2, and the constraint is $f'(x) = a$ with $a \neq 0$. The bandwidth with optimal rate is again $b = n^{-1/3}$. Set $\hat f_c(x) = \hat f(x) - n^{-1/3}c(\hat f'(x) - a)$. The joint asymptotic distribution of $n^{1/3}(\hat{f}(x) - f(x))$ and $\hat{f}'(x) - f'(x)$ is normal with covariance matrix as above, and with bias $a(-\int tK_0, \int K_1)^\top = a(-\int tK_0, 1)^\top$. The asymptotic MSE of $n^{1/3}(\hat{f}_c(x) - f(x))$ is

$$f(x)\Big(\int K_0^2 - 2c\int K_0K_1' + c^2\int K_1'^2\Big) + a^2\Big(\int tK_0 + c\Big)^2 = A - 2cB + c^2C$$

with

$$\begin{aligned}
A &= f(x)\int K_0^2 + a^2\Big(\int tK_0\Big)^2,\\
B &= f(x)\int K_0K_1' - a^2\int tK_0,\\
C &= f(x)\int K_1'^2 + a^2.
\end{aligned}$$

If $f(x) > 0$, then $C > 0$ and the asymptotic MSE is minimised by $c = c_* = B/C$. The minimal asymptotic MSE is $A - B^2/C$. Both B and C depend on the density through $f(x)$. Write $\hat{c}_*$ for c_* with $f(x)$ replaced by $\hat{f}(x)$,

$$\hat{c}_* = \frac{\big(\hat{f}(x)\int K_0K_1' - a^2\int tK_0\big)^2}{\hat{f}(x)\int K_1'^2 + a^2}.$$

Then the asymptotic MSE of $n^{1/3}(\hat{f}_{\hat{c}_*}(x) - f(x))$ is $A - B^2/C$, while $n^{1/3}(\hat{f}(x) - f(x))$ has asymptotic MSE A. □

Example 4. (Nonvanishing derivative; r=2.) For a second example with nonvanishing highest-order derivative, again take the constraint $f'(x) = 0$ as in Example 2, but now assume that f is known to be *twice* continuously differentiable at x. Then $A = \mathrm{diag}(1,0)$ and $a = (1,0)^\top$. The bandwidth $b_0 = n^{-2/5}$ gives the optimal rate. Since $f''(x)$ is not involved in the constraint, we can set $\hat{f}_c(x) = \hat{f}(x) - cn^{-1/5}\hat{f}'(x)$. By Proposition 1, $n^{2/5}(\hat{f}(x) - f(x))$ is asymptotically normal with mean $f''(x)\frac{1}{2}\int t^2K_0$ and variance $f(x)\int K_0^2$. Hence the asymptotic MSE of $n^{2/5}(\hat{f}(x) - f(x))$ is

$$A = f(x)\int K_0^2 + f''^2(x)\frac{1}{4}\Big(\int t^2K_0\Big)^2.$$

On the other hand, by Theorem 1, $n^{2/5}(\hat{f}_c(x) - f(x))$ is asymptotically normal with mean

$$f''(x)\Big(\frac{1}{2}\int t^2K_0 + c\int tK_1\Big)$$

and variance $f(x)\big(\int K_0^2 - 2c\int K_0K_1' + c^2\int K_1'^2\big)$. Hence the asymptotic MSE of $\hat{f}_c(x)$ is

$$\begin{aligned}
&f(x)\Big(\int K_0^2 - 2c\int K_0K_1' + c^2\int K_1'^2\Big)\\
&+ f''^2(x)\Big(\frac{1}{4}\Big(\int t^2K_0\Big)^2 + \frac{1}{2}c\int t^2K_0\int tK_1 + c^2\Big(\int tK_1\Big)^2\Big)\\
&= A - 2cB + c^2C
\end{aligned}$$

with

$$B = f(x) \int K_0 K_1' - \frac{1}{4} f''^2(x) \int t^2 K_0 \int t K_1,$$
$$C = f(x) \int K_1'^2 + f''^2(x) \Big(\int t K_1 \Big)^2.$$

If $C > 0$, the asymptotic MSE of $\hat{f}_c(x)$ is minimised by $c = c_* = B/C$. The minimal asymptotic MSE is $A - B^2/C$.

Suppose that K_1 is of order 2. Then $\int t K_1 = 0$, and we have $B = f(x) \int K_0 K_1'$ and $C = f(x) \int K_1'^2$. Assume that K_1' does not vanish. This means that K_1 is not a box kernel. Then $\int K_1'^2 \neq 0$, and c_* simplifies to $\int K_0 K_1' / \int K_1'^2$, and the asymptotic MSE simplifies to

$$f(x) \Big(\int K_0^2 - \frac{\big(\int K_0 K_1'\big)^2}{\int K_1'^2} \Big) + \frac{1}{4} f''^2(x) \Big(\int t^2 K_0 \Big)^2.$$

By the Cauchy inequality, $\big(\int K_0 K_1'\big)^2 \leq \int K_0^2 \int K_1'^2$. Note again that K_1' cannot be proportional to K_0 since $\int K_0 = 1$ but $\int K_1' = 0$. Hence the variance term of the asymptotic MSE is always positive. It may however happen that $\int K_0 K_1' = 0$. Then $\hat{f}_{c_*}(x)$ has the same asymptotic MSE as $\hat{f}(x)$. This is in particular the case if K_0 and K_1 are symmetric, so K_1' is antisymmetric and hence orthogonal to K_0. □

3 Proofs

Proof of Lemma 1. For $j = 0, \dots, r$ we have the Taylor expansion

$$f^{(j)}(x - b_j t) - f^{(j)}(x) = \sum_{k=1}^{r-j} \frac{(-b_j t)^k}{k!} f^{(j+k)}(x)$$
$$+ \frac{(-b_j t)^{r-j}}{(r-j-1)!} \int_0^1 (1-t)^{r-j-1} \big(f^{(r-j)}(x - b_j t) - f^{(r-j)}(x)\big)\, dt.$$

For appropriately differentiable f and g we have $(f * g)' = f' * g = f * g'$ and therefore $(f * g)^{(j)} = f^{(j)} * g = f * g^{(j)}$. In particular,

$$E[\hat{f}^{(j)}(x)] = E[K_{jb_j}^{(j)}(x)] = K_{jb_j}^{(j)} * f(x) = K_{jb_j} * f^{(j)}(x) = \int K_j(t) f^{(j)}(x - b_j t)\, dt.$$

Since K_j is of order j and $f^{(r-j)}$ is j times continuously differentiable, we obtain the asserted expansion of the bias of $\hat{f}^{(j)}(x)$.

The variance of $\hat{f}^{(j)}(x)$ is

$$n \mathrm{Var}\, \hat{f}^{(j)}(x) = \mathrm{Var}\, K_{jb_j}^{(j)}(x - X)$$
$$= E[K_{jb_j}^{(j)2}(x - X)] - (E[K_{jb_j}^{(j)}(x - X)])^2$$
$$= b_j^{-2j-1} \int K_j^{(j)2}(t) f(x - b_j t)\, dt - \Big(b_j^{-j} \int K_j^{(j)}(t) f(x - b_j t)\, dt \Big)^2.$$

Since $f^{(j)}$ is continuous at x, we obtain the asserted approximation of the variance of $\hat{f}^{(j)}(x)$ similarly as for the bias.

Proof of Proposition 1. Recall that we have set $b_j = b = n^{-1/(2r+1)}$. Lemma 1 gives expansions for the bias and variance of $\hat{f}^{(j)}(x)$. For the covariances we obtain by a similar argument

$$
\begin{aligned}
E[K_{jb}^{(j)}(x-X)K_{kb}^{(k)}(x-X)] &= \frac{1}{b^{j+1}b^{k+1}} E\Big[K_j^{(j)}\Big(\frac{x-X}{b}\Big)K_k^{(k)}\Big(\frac{x-X}{b}\Big)\Big] \\
&= \frac{1}{b^{j+k+2}} \int K_j^{(j)}\Big(\frac{x-u}{b}\Big)K_k^{(k)}\Big(\frac{x-u}{b}\Big) f(u)\,du \\
&= n^{(j+k+2)/(2r+1)} \int K_j^{(j)}(t)K_k^{(k)}(t) f(x-b_j t)\,dt.
\end{aligned}
$$

Hence

$$
n^{(2r-j-k)/(2r+1)} \operatorname{Cov} \hat{f}^{(j)}(x)\hat{f}^{(k)}(x) \to f(x)\sigma_{jk}.
$$

Set $Y_{ni} = b^r K^{(\cdot)}((x-X_i)/b)$ with $K^{(\cdot)} = (K_0^{(0)}, \dots, K_r^{(r)})^\top$. Then

$$
\begin{pmatrix} n^{r/(2r+1)}\hat{f}(x) \\ \vdots \\ \hat{f}^{(r)}(x) \end{pmatrix} = \sum_{i=1}^n Y_{ni}.
$$

With $nb^{2r+1} = 1$ we have

$$
\begin{aligned}
nE\|Y_n\|^2 \mathbf{1}(\|Y_n\| > \varepsilon) &= nb^{2r} E\Big\|K^{(\cdot)}\Big(\frac{x-X}{b}\Big)\Big\|^2 \mathbf{1}\Big(\Big\|K^{(\cdot)}\Big(\frac{x-X}{b}\Big)\Big\| > b^{-r}\varepsilon\Big) \\
&= \int \|K^{(\cdot)}(t)\|^2 \mathbf{1}\big(\|K^{(\cdot)}(t)\| > b^{-r}\varepsilon\big) f(x-bt)\,dt \to 0.
\end{aligned}
$$

The assertion now follows from the central limit theorem of Lindeberg and Feller. See e.g. van der Vaart (1998), Proposition 2.27, for a multivariate version.

Proof of Theorem 1. Write

$$
n^{r/(2r+1)}(\hat{f}_c(x) - f(x)) = c_-^\top A_- V_n.
$$

This is asymptotically normal by Proposition 1. The mean is

$$
f^{(r)}(x) c_-^\top A_- \mu = f^{(r)}(x)(\mu_0 - c^\top A \nu),
$$

and the variance is

$$
f(x) c_-^\top A_- \Sigma A_-^\top c_- = f(x)\big(\sigma_{00} - 2c^\top A\lambda + c^\top A \Lambda A^\top c\big).
$$

Acknowledgments

The authors thank the reviewer for a number of suggestions which they believe have improved the paper.

References

[1] Bhattacharya, P. K. (1967). Estimation of a probability density function and its derivatives. *Sankhyā Ser. A* **29**, 373–382.

[2] Bradley, R. C. (1983). Asymptotic normality of some kernel-type estimators of probability density. *Statist. Probab. Lett.* **1**, 295–300.

[3] Fan, J. and Yao, Q. (2003). *Nonlinear Time Series. Nonparametric and Parametric Methods.* Springer Series in Statistics. Springer-Verlag, New York.

[4] Levit, B. Y. (1975). Conditional estimation of linear functionals. *Probl. Inf. Transm.* **11**, 291–301.

[5] Lu, Z. (2001). Asymptotic normality of kernel density estimators under dependence. *Ann. Inst. Statist. Math.* **53**, 447–468.

[6] Masry, E. and Fan, J. (1997). Local polynomial estimation of regression functions for mixing processes. *Scand. J. Statist.* **24**, 165–179.

[7] Müller, U. U. and Wefelmeyer, W. (2002). Estimators for models with constraints involving unknown parameters. *Math. Methods Statist.* **11**, 221–235.

[8] Parzen, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.* **33**, 1065–1076.

[9] Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.* **27**, 832–837.

[10] Schuster, E. F. (1969). Estimation of a probability density function and its derivatives. *Ann. Math. Statist.* **40**, 1187–1195.

[11] Singh, R. S. (1977). Improvement on some known nonparametric uniformly consistent estimators of derivatives of a density. *Ann. Statist.* **5**, 394–399.

[12] Singh, R. S. (1981). On the exact asymptotic behavior of estimators of a density and its derivatives. *Ann. Statist.* **9**, 453–456.

[13] van der Vaart, A. W. (1998). *Asymptotic Statistics.* Cambridge Series in Statistical and Probabilistic Mathematics 3, Cambridge University Press.

[14] Zhang, B. (1998). A note on kernel density estimation with auxiliary information. *Comm. Statist. Theory Methods* **27**, 1–11.