Density estimators for the convolution of discrete and continuous random variables

Ursula U. Müller
Texas A&M University

Anton Schick
Binghamton University

Wolfgang Wefelmeyer
Universität zu Köln

Abstract

Suppose we have independent observations of a pair of independent random variables, one with a density and the other discrete. The sum of these random variables has a density, which can be estimated by an ordinary kernel estimator. Since the two components are independent, we can write the density as a convolution and alternatively estimate it by a convolution of a kernel estimator of the continuous component with an empirical estimator of the discrete component. We show that for a given kernel and optimal bandwidth, this estimator has the same rate as the first estimator, and the same asymptotic bias, but a much smaller asymptotic variance. We also show how pointwise constraints on derivatives of the density of the continuous component can be used to improve our estimator of the convolution density.

1 Introduction

Let X and Y be independent real-valued random variables. Assume that X has a density f. Let Y be discrete with finite support T, and taking the value t with positive probability p_t for $t \in T$. Then the convolution $Z = X + Y$ has density

$$h(z) = \sum_{t \in T} f(z-t)p_t.$$

Suppose we have independent observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ and want to estimate the density h. An obvious estimator is a kernel estimator

$$\hat{h}(z) = \frac{1}{nb} \sum_{i=1}^{n} K\Big(\frac{z - X_i - Y_i}{b}\Big)$$

based on the sums $Z_i = X_i + Y_i$, where K is a kernel and b a bandwidth. Better estimators will be obtained by exploiting that X and Y are independent. One possibility is

$$\hat{h}_*(z) = \frac{1}{n^2 b} \sum_{i=1}^{n} \sum_{j=1}^{n} K\Big(\frac{z - X_i - Y_j}{b}\Big).$$

Note that these estimators do not require knowledge of T.

Let us set

$$N_t = \sum_{i=1}^{n} \mathbf{1}[Y_i = t] \quad \text{and} \quad \hat{p}_t = \frac{N_t}{n} \quad \text{for} \quad t \in \mathbb{R}.$$

Then the first estimator can be expressed as

$$\hat{h}(z) = \sum_{t \in T} \tilde{f}_t(z-t)\hat{p}_t$$

with kernel estimator $\tilde{f}_t(x) = 1/(N_t b) \sum_{i=1}^{n} \mathbf{1}[Y_i = t] K((x - X_i)/b)$ based only on the observations X_i with $Y_i = t$, while the second one can be written as

$$\hat{h}_*(z) = \sum_{t \in T} \hat{f}(z-t)\hat{p}_t$$

with kernel estimator $\hat{f}(x) = 1/(nb) \sum_{i=1}^{n} K((x - X_i)/b)$ based on all X_i. These representations suggest that $\hat{h}_*(z)$ should be better than $\hat{h}(z)$.

Let f be r times differentiable. Then so is h. Let K be of order r. Then the optimal rate of the bandwidth is $n^{-1/(2r+1)}$. We may take $b = n^{-1/(2r+1)}$, absorbing a possible positive factor as a scale parameter into K. It is known (see Lemma 1) that $n^{r/(2r+1)}(\hat{h}(z) - h(z))$ is asymptotically normal with mean

$$B = h^{(r)}(z) \frac{(-1)^r}{r!} \int u^r K(u)\, du$$

and variance

$$V = h(z) \int K^2(u)\, du = \sum_{t \in T} f(z-t) p_t \int K^2(u)\, du.$$

In Section 2 we show that $n^{r/(2r+1)}(\hat{h}_*(z) - h(z))$ is asymptotically normal with the same mean, but with variance

$$V_* = \sum_{t \in T} f(z-t) p_t^2 \int K^2(u)\, du.$$

Unless Y is a constant, V_* is strictly smaller than V because it has p_t^2 in place of p_t. The variance reduction is, in general, considerable. If Y is uniformly distributed, the variance is reduced by a factor $|T|^{-1}$.

The result applies in particular to spatial statistics, where we often estimate the length of a vector with independent components, using distance measures involving sums of functions of the components. Whenever only one of the components has a density and the others are discrete, our result applies. When more than one component has a density, we get different convergence rates. This holds in particular when *all* components have densities, which is the case treated extensively in the literature so far. Similar results can be obtained when we estimate the response density of a regression model $Y = r(X) + \varepsilon$ with X and ε independent, and with X discrete or ε discrete or r a step function. Here we have

independent observations (X_i, Y_i) for $i = 1, \ldots, n$ but do not observe $r(X_i)$ and $\varepsilon_i = Y_i - r(X_i)$. We must therefore estimate the regression function r by an estimator $\hat{r}$ and use $\hat{r}(X_i)$ and residuals $\hat{\varepsilon}_i = Y_i - \hat{r}(X_i)$ in place of $r(X_i)$ and ε_i. We treat this elsewhere.

Convolution estimators for $X+Y$ behave quite differently when Y also has a density, say g. Then $X+Y$ has density $h(z) = \int f(z-y)g(y)\,dy$ and can be estimated by a convolution estimator $\int \hat{f}(z-y)\hat{g}(y)\,dy$ with kernel estimators $\hat{f}$ and $\hat{g}$. If $f(X)$ and $g(Y)$ have finite second moments, such an estimator has a faster rate than the obvious estimator $\hat{h}(z)$, namely the rate $n^{-1/2}$ of an empirical estimator; see Frees (1994), Schick and Wefelmeyer (2004) and (2007), and Giné and Mason (2007). Corresponding results for nonparametric regression are in Schick and Wefelmeyer (2012a and 2013), and in Müller (2012) for nonlinear regression. If $f(X)$ or $g(Y)$ does not have a finite second moment, the convolution estimator does not attain the rate $n^{-1/2}$; see Schick and Wefelmeyer (2009a, 2009b and 2012b).

In Section 3 we assume that we have auxiliary information about f in the form of constraints on certain derivatives of f at certain points $z-t$ for $t \in T$. This requires that the support T of Y is known. From Müller and Wefelmeyer (2014) we obtain improvements of $\hat{h}_*(z)$ that reduce the asymptotic mean squared error. The main applications are to cases in which we know that certain derivatives are zero at one or several of the points $z-t$. For example, the density may have a maximum at one point, $f'(z-t) = 0$; an inflection point, $f''(z-t) = 0$; or a saddle point, $f'(z-t) = 0$ and $f''(z-t) = 0$. The density may also be known to be bimodal with the two maxima and the minimum among the points $z-t$. The proofs are in Section 4. As usual we write $X_n = o_p(1)$ if the sequence of random variables X_n converges in probability to zero, and $X_n = O_p(1)$ if it is bounded in probability, with $X_n = o_p(a_n)$ if $a_n^{-1}X_n = o_p(1)$ and analogously for $O_p(a_n)$.

2 Main result

We begin with a general lemma on the asymptotic behavior of kernel estimators. It is essentially known, also under mixing conditions and for linear processes; see Parzen (1962), Chanda (1983), Bradley (1983), Tran (1992), Hallin and Tran (1996) and Lu (2001). We obtain the asymptotic bias under a minimal differentiability assumption on the density. Let $\mathcal{K}$ denote the set of bounded measurable functions with compact support. For functions K and L in $\mathcal{K}$, we set

$$\mu_j(K) = \frac{(-1)^j}{j!} \int u^j K(u)\,du \quad \text{and} \quad \langle K, L\rangle = \int K(u)L(u)\,du.$$

We formulate our lemma abstractly for independent real-valued random variables $\xi_1, \xi_2, \ldots$ with a common density g. Set

$$H_n(z, b, K) = \frac{1}{nb} \sum_{j=1}^{n} K\Big(\frac{z-\xi_j}{b}\Big), \quad z \in \mathbb{R},\ b > 0, K \in \mathcal{K}.$$

Lemma 1. *Let b_n be positive numbers such that $b_n \to 0$ and $nb_n \to \infty$. If g is continuous at z, then we have the convergence results*

$$E[H_n(z, b_n, K)] \to g(z) \quad \textit{and} \quad nb_n \operatorname{Var} H_n(z, b_n, K) \to g(z)\langle K, K\rangle,$$

and the random variable $(nb_n)^{1/2}(H_n(z, b_n, K) - E[H_n(z, b_n, K)])$ is asymptotically normal with mean zero and variance $g(z)\langle K, K\rangle$. If g is r times differentiable at z, then

$$E[H_n(z, b_n, K)] = \sum_{j=1}^{r} b_n^j g^{(j)}(z)\mu_j(K) + o(b_n^r).$$

If g is continuous at the distinct points $z_1, \ldots, z_k$ and $K_1, \ldots, K_k$ belong to $\mathcal{K}$, then the k-dimensional random vector

$$(nb_n)^{1/2} \begin{bmatrix} H_n(z_1, b_n, K_1) - E[H_n(z_1, b_n, K_1)] \\ \vdots \\ H_n(z_k, b_n, K_k) - E[H_n(z_k, b_n, K_k)] \end{bmatrix}$$

is asymptotically normal with mean vector 0 and diagonal dispersion matrix

$$\operatorname{diag}\left(\langle K_1, K_1\rangle g(z_1), \ldots, \langle K_k, K_k\rangle\, g(z_k)\right).$$

In Section 3 we determine the asymptotic distribution of linear combinations of kernel estimators for different derivatives of a density. This is obtained from the following general result on the joint asymptotic distribution of kernel estimators. Let $\mathcal{K}_r$ denote the set of functions K in $\mathcal{K}$ that are *kernels of order r*, i.e., that satisfy

$$\mu_0(K) = 1, \quad \mu_r(K) \neq 0, \quad \text{and} \quad \mu_j(K) = 0, \quad j = 1, \ldots, r-1.$$

Let $\mathcal{L}_r$ denote the set of functions K in $\mathcal{K}$ that satisfy

$$\mu_r(K) = 1 \quad \text{and} \quad \mu_i(K) = 0, \quad i = 0, \ldots, r-1.$$

For $j = 0, \ldots, r-1$, let $\mathcal{L}_j$ denote the set of functions K in $\mathcal{K}$ that satisfy

$$\mu_i(K) = \mathbf{1}[i = j], \quad i = 0, \ldots, r-1.$$

Integration by parts shows that if $K \in \mathcal{K}_r$ is r times continuously differentiable, then $K^{(j)}$ belongs to $\mathcal{L}_j$.

For $L_j \in \mathcal{L}_j$, $j = 0, \ldots, r$, set $L = (L_0, L_1, \ldots, L_r)^\top$ and

$$\Delta_n = b_n^{-r} \left(\begin{bmatrix} H_n(z, b_n, L_0) \\ H_n(z, b_n, L_1) \\ \vdots \\ H_n(z, b_n, L_{r-1}) \\ H_n(z, b_n, L_r) \end{bmatrix} - \begin{bmatrix} g(z) \\ b_n g'(z) \\ \vdots \\ b_n^{r-1} g^{(r-1)}(z) \\ b_n^r g^{(r)}(z) \end{bmatrix} \right) - g^{(r)}(z) \begin{bmatrix} \mu_r(L_0) \\ \mu_r(L_1) \\ \vdots \\ \mu_r(L_{r-1}) \\ 0 \end{bmatrix}.$$

Since $H_n(z, b, K)$ is linear in K, we have with $b_n = n^{-1/(2r+1)}$ that

$$\Delta_n \Rightarrow N(0, g(z)\langle L, L^\top\rangle). \tag{2.1}$$

This implies that $\hat{v}^\top \Delta_n \Rightarrow N(0, g(z)\langle v^\top L, v^\top L\rangle)$ whenever the $(r+1)$-dimensional random vector $\hat{v}$ converges in probability to the constant vector v. A special case of this result is formulated in the next lemma. It is needed in the next section.

Lemma 2. *Suppose g is r times differentiable at z. Let L_j belong to $\mathcal{L}_j$ for $j = 0, \ldots, r$, and let the random coefficient $\hat{d}_k$ converge in probability to the constant d_k for $k = 1, \ldots, r$. Take $b_n = n^{-1/(2r+1)}$. Then the random variable*

$$T_n = H_n(z, b_n, L_0) - g(z) - \sum_{k=1}^{r} \hat{d}_k \big(H_n(z, b_n, L_k) - b_n^k g^{(k)}(z)\big)$$

satisfies

$$n^{r/(2r+1)} T_n \Rightarrow N(g^{(r)}(z)\nu, g(z)\langle \tilde{L}, \tilde{L}\rangle) \tag{2.2}$$

with

$$\nu = \mu_r(L_0) - \sum_{k=1}^{r-1} d_k \mu_r(L_k) \quad \text{and} \quad \tilde{L} = L_0 - \sum_{k=1}^{r} d_k L_k.$$

This result holds in particular with $L_j = K^{(j)}$, $j = 0, \ldots, r$, if K is an r times continuously differentiable member of $\mathcal{K}_r$. In this case, ν simplifies to $\mu_r(K)$, and $\tilde{L}$ equals $K - \sum_{j=1}^{r} d_j K^{(j)}$.

Since the density f of X is r times differentiable at $z - t$ for $t \in T$, the density $h(z) = \sum_{t \in T} f(z - t) p_t$ of $Z = X + Y$ is r times differentiable at z. Hence Lemma 1 implies the following result.

Proposition 1. *Let f be r times differentiable at $z - t$ for $t \in T$, let $K \in \mathcal{K}_r$, and let $b \to 0$ and $nb \to \infty$. Then $b^{-r}(E[\hat{h}(z)] - h(z)) \to B$ and $nb\mathrm{Var}\,\hat{h}(z) \to V$, with B and V defined in Section 1. Set $b = n^{-1/(2r+1)}$. Then $n^{r/(2r+1)}(\hat{h}(z) - h(z))$ is asymptotically normal with mean B and variance V.*

For $t \in T$ choose a kernel K_t and set $\hat{f}_t(x) = 1/(nb) \sum_{i=1}^{n} K_t((x - X_i)/b)$ and

$$\hat{h}_*(z) = \sum_{t \in T} \hat{f}_t(z - t)\hat{p}_t.$$

Here $\hat{p}_t = N_t/n$ is the empirical estimator of p_t introduced in Section 1.

Theorem 1. *Let f be r times differentiable at $z - t$, let $K_t \in \mathcal{K}_r$ for $t \in T$, and set $b = n^{-1/(2r+1)}$. Then $n^{r/(2r+1)}(\hat{h}_*(z) - h(z))$ is asymptotically normal with mean*

$$C = \sum_{t \in T} f^{(r)}(z - t) p_t \mu_r(K_t)$$

and variance

$$W_* = \sum_{t \in T} f(z - t) p_t^2 \langle K_t, K_t\rangle.$$

In order to compare $\hat{h}$ with the convolution estimator $\hat{h}_*$, we use the same kernel $K = K_t$ for each $t \in T$. Then the asymptotic mean of $n^{r/(2r+1)}(\hat{h}_*(z) - h(z))$ is

$$C = \sum_{t \in T} f^{(r)}(z-t) p_t \mu_r(K) = h^{(r)}(z) \mu_r(K) = B,$$

and the asymptotic variance is

$$W_* = \sum_{t \in T} f(z-t) p_t^2 \langle K, K \rangle = V_*,$$

which is strictly smaller than V.

3 Pointwise constraints

As in Section 2 we assume that f is r times differentiable at $z-t$ for $t \in T$. Suppose that T is known, and that for t in a subset T_0 of T we know the values of several of the derivatives $f'(z-t), \ldots, f^{(r)}(z-t)$ at $z-t$. More precisely, we assume that there is a nonempty subset J_t of $\{1, \ldots, r\}$ and numbers $a_{t,j}$ such that

$$f^{(j)}(z-t) = a_{t,j}, \quad j \in J_t. \tag{3.1}$$

Suppose, for example, that f is twice differentiable, with a maximum at $z-t$. Then $r = 2$ and the constraint is $f'(z-t) = 0$. We will come back to this example at the end of the section.

For $j \in J_t$, we estimate $f^{(j)}(z-t)$, the j-th derivative of f at $z-t$, by

$$\hat{f}_t^{(j)}(z-t) = \frac{1}{n b^{j+1}} \sum_{i=1}^{n} L_{t,j}\Big(\frac{z-t-X_i}{b}\Big)$$

with $L_{t,j}$ a member of $\mathcal{L}_j$. In what follows we take $b = n^{-1/(2r+1)}$ and let K_t be a member of $\mathcal{K}_r$. In view of the constraints (3.1) we can try to improve the kernel estimator

$$\hat{f}_t(x) = \frac{1}{nb} \sum_{i=1}^{n} K_t\Big(\frac{x-X_j}{b}\Big)$$

by looking at the estimators

$$\hat{f}_{t,c}(z-t) = \hat{f}_t(z-t) - \sum_{j \in J_t} c_j b^j (\hat{f}_t^{(j)}(z-t) - a_{t,j})$$

for vectors $c = (c_j)_{j \in J_t}$. By Lemma 2, the random variable $n^{r/(2r+1)}\big(\hat{f}_{t,c}(z-t) - f(z-t)\big)$ is asymptotically normal with mean

$$f^{(r)}(z-t)\Big(\mu_r(K_t) - \sum_{j \in J_t} c_j \mu_r(L_{t,j}) \mathbf{1}[j \neq r]\Big)$$

and variance

$$f(z-t)\Big(\langle K_t, K_t\rangle - 2\sum_{j\in J_t} c_j\langle K_t, L_{t,j}\rangle + \sum_{j,k\in J_t} c_j\langle L_{t,j}, L_{t,k}\rangle c_k\Big).$$

The asymptotic MSE is $M_t - 2c^\top B_t + c^\top C_t c$ with nonnegative number M_t, vector B_t and symmetric matrix C_t given by

$$\begin{aligned} M_t &= f(z-t)\langle K_t, K_t\rangle + f^{(r)2}(z-t)\mu_r^2(K_t),\\ B_t &= f(z-t)\lambda_t + f^{(r)2}(z-t)\mu_r(K_t)\nu_t,\\ C_t &= f(z-t)\Lambda_t + f^{(r)2}(z-t)\nu_t\nu_t^\top, \end{aligned}$$

with $\lambda_{t,j} = \langle K_t, L_{t,j}\rangle$, $\Lambda_{t,jk} = \langle L_{t,j}, L_{t,k}\rangle$ and $\nu_{t,j} = \mathbf{1}[j \neq r]\mu_r(L_{t,j})$ for $j,k \in J_t$. Suppose that the matrix Λ_t is invertible. Then so is C_t provided $f(z-t) \neq 0$. In this case the asymptotic MSE is minimized by $c = c_t = C_t^{-1}B_t$.

The vector c_t depends on the unknown density f and must be replaced by an estimator. Write $\hat{B}_t$ and $\hat{C}_t$ for B_t and C_t with $f(z-t)$ and $f^{(r)}(z-t)$ replaced by the estimators $\hat{f}_t(z-t)$ and $\hat{f}_t^{(r)}(z-t)$. Set $\hat{c}_t = \hat{C}_t^{-1}\hat{B}_t$. Then $\hat{c}_t - c_t = o_p(1)$. Thus it follows from Lemma 2 that $n^{r/(2r+1)}(\hat{f}_{t,\hat{c}_t}(z-t) - f(z-t))$ has minimal asymptotic MSE $M_t - B_t^\top C_t^{-1}B_t$.

Using $\hat{f}_{t,\hat{c}_t}(z-t)$ instead of $\hat{f}_t(z-t)$ for $t \in T_0$ in the definition of $\hat{h}_*(z)$, we now obtain new estimators for $h(z) = \sum_{t\in T} f(z-t)p_t$ as

$$\hat{h}_0(z) = \sum_{t\in T_0} \hat{f}_{t,\hat{c}_t}(z-t)\hat{p}_t + \sum_{t\in T-T_0} \hat{f}_t(z-t)\hat{p}_t.$$

This estimator can be written as

$$\hat{h}_0(z) = \hat{h}_*(z) - \sum_{t\in T_0}\hat{p}_t\sum_{j\in J_t}\hat{c}_{t,j}b^j(\hat{f}_t^{(j)}(z-t) - a_{t,j})$$

Theorem 2. *For $t \in T$, let f be r times differentiable at $z-t$ and let K_t belong to $\mathcal{K}_r$. For $t \in T_0$, let the constraints (3.1) hold, let $f(z-t) \neq 0$, and let Λ_t be positive definite. Set $b = n^{-1/(2r+1)}$. Then $n^{r/(2r+1)}(\hat{h}_0(z) - h(z))$ is asymptotically normal with mean*

$$\sum_{t\in T} p_t f^{(r)}(z-t)\big(\mu_r(K_t) - d_t\nu_t\big)$$

and variance

$$\sum_{t\in T} p_t^2 f(z-t)\big(\langle K_t, K_t\rangle - 2d_t^\top\lambda_t + c_t^\top\Lambda_t c_t\big),$$

where d_t equals c_t for $t \in T_0$ and 0 otherwise.

The asymptotic MSE of $n^{r/(2r+1)}(\hat{h}_0(z) - h(z))$ is

$$\begin{aligned} M_{(d_t)} = &\sum_{s,t\in T} p_s p_t f^{(r)}(z-s)f^{(r)}(z-t)\big(\mu_r(K_s) - d_s^\top\nu_s\big)\big(\mu_r(K_t) - d_t^\top\nu_t\big)\\ &+ \sum_{t\in T} p_t^2 f(z-t)\big(\langle K_t, K_t\rangle - 2d_t^\top\lambda_t + d_t^\top\Lambda_t d_t\big). \end{aligned}$$

In particular, setting all $d_t = 0$, we obtain the asymptotic MSE of $\hat{h}_*(z)$ as $M = M_{(0)}$.

The most important special case is that of T_0 being a singleton, say $T_0 = \{u\}$. Then the constraints can be written as $f^{(j)}(z-u) = a_{u,j}$ for j in some subset $J = J_u$ of $\{1, \dots, r\}$. These constraints can be used to construct the estimator

$$\hat{h}_{*,c} = \hat{h}_*(z) - \hat{p}_u \sum_{j \in J} c_j b^j (\hat{f}_u^{(j)}(z-u) - a_{u,j})$$

with $M - 2c^\top B + c^\top C c$ as asymptotic MSE, where the vector B and the symmetric matrix C are

$$B = p_u \nu_u f^{(r)}(z-u) \sum_{t \in T} p_t f^{(r)}(z-t) \mu_r(K_t) + p_u^2 \lambda_u f(z-u),$$
$$C = p_u^2 f^{(r)2}(z-u) \nu_u \nu_u^\top + p_u^2 f(z-u) \Lambda_u.$$

We require that Λ_u is positive definite and $f(z-u)$ is positive. Then C is invertible and the asymptotic MSE is minimized by the choice $c = c^* = C^{-1}B$. Write $\hat{c}$ for c with $f(z-u)$ and $f^{(r)}(z-t)$ replaced by the estimators $\hat{f}_u(z-u)$ and $\hat{f}_t^{(r)}(z-t)$, and p_t replaced by the estimator $\hat{p}_t$ for $t \in T$. It follows that $n^{r/(2r+1)}(\hat{h}_{*,\hat{c}}(z) - h(z))$ has minimal asymptotic MSE $M - B^\top C^{-1} B$.

For example, let $r = 2$. Then the bandwidth $b = n^{-1/5}$ has the optimal rate. The asymptotic MSE of $n^{2/5}(\hat{h}_*(z) - h(z))$ is

$$M = \sum_{s,t \in T} p_s p_t f''(z-s) f''(z-t) \mu_2(K_s) \mu_2(K_t) + \sum_{t \in T} p_t^2 f(z-t) \langle K_t, K_t \rangle.$$

Let us return to the example from the beginning of this section in which the location of a maximum is known, i.e. $f'(z-u) = 0$ for some $u \in T$. New estimators for $f(z-u)$ are of the form $\hat{f}_{u,c}(z-u) = \hat{f}_u(z-u) - cn^{-1/5} \hat{f}'_u(z-u)$. Our proposed estimators for $h(z)$ are therefore

$$\hat{h}_{*,c}(z) = \hat{h}_*(z) - cn^{-1/5} \hat{f}'_u(z-u) \hat{p}_u.$$

The asymptotic MSE of $n^{2/5}(\hat{h}_{*,c}(z) - h(z))$ is $M - 2cB + c^2 C$, where B and C are the real numbers

$$B = p_u \mu_2(L_{u,1}) f''(z-u) \sum_{t \in T} p_t f''(z-t) \mu_2(K_t) + p_u^2 \langle K_u, L_{u,1} \rangle f(z-u),$$
$$C = p_u^2 \mu_1^2(L_{u,1}) f''^2(z-u) + p_u^2 \langle L_{u,1}, L_{u,1} \rangle f(z-u).$$

The asymptotic MSE of $n^{2/5}(\hat{h}_{*,c} - h(z))$ is therefore minimized if c is replaced by a consistent estimator $\hat{c}$ of $c_* = B/C$, and the constraint $f'(z-u) = 0$ leads to a reduction of the asymptotic MSE by $2c_*B - c_*^2 C = B^2/C$.

As pointed out in Remark 3 of Müller and Wefelmeyer (2014), it is important to use different kernels for different derivatives. For example, there is no improvement if we take $L_{u,1} = K'_u$ as then $\mu_2(K'_u) = 0$ and $\langle K_u, K'_u \rangle = 0$.

4 Proofs

Proof of Lemma 1. The first three results follow from Parzen (1962). The fourth conclusion follows from the identity

$$E[H_n(z,b,K)] = \frac{1}{b}\int K\Big(\frac{z-x}{b}\Big)g(x)\,dx = \int g(z-bu)K(u)\,du$$

and the Taylor expansion

$$\begin{aligned} g(z-bu) = &\sum_{j=0}^{r}\frac{(-bu)^j}{j!}g^{(j)}(z) \\ &+ (-bu)^{r-1}\int \big(g^{(r-1)}(z-buv) - g^{(r-1)}(z) - buvg^{(r)}(z)\big)\,F_{r-1}(dv), \end{aligned}$$

where F_0 denotes the point mass at 1 and F_k denotes the measure with density

$$f_k(v) = \frac{(1-v)^{k-1}}{(k-1)!}\mathbf{1}[0<v<1] \quad \text{for} \quad k>0.$$

These identities allow us to derive the inequality

$$\Big|E[H_n(z,b,K)] - \sum_{j=0}^{r}\frac{(-b)^j}{j!}g^{(j)}(z)\int u^jK(u)\,du\Big| \le b^r w(bC)\int |v|\,dF_{r-1}(v)\int |u^rK(u)|\,du$$

with C a constant such that $[-C,C]$ contains the support of K, and

$$w(\delta) = \sup_{|t|\le\delta}\frac{\big|g^{(r-1)}(z-t) - g^{(r-1)}(z) - tg^{(r)}(z)\big|}{|t|}$$

for small $\delta>0$. Since $g^{(r-1)}$ is differentiable at z, we have $w(\delta)\to 0$ as $\delta\to 0$.

For the fifth result one shows that the random vectors

$$U_{ni} = \frac{1}{\sqrt{b_n n}}\begin{bmatrix} K_1((z_1-\xi_j)/b_n) - E[K_1((z_1-\xi_j)/b_n)] \\ \vdots \\ K_k((z_k-\xi_j)/b_n) - E[K_k((z_k-\xi_j)/b_n)] \end{bmatrix}, \quad i=1,\ldots,n,$$

are independent and centered, satisfy the Lindeberg condition in view of the bound $\|U_{ni}\|^2 \le 4kB^2/(nb_n) = o(1)$ with B a bound for $K_1,\ldots,K_k$, and their common dispersion matrix multiplied by n converges to the diagonal matrix diag $(\langle K_1,K_1\rangle g(z_1),\ldots,\langle K_k,K_k\rangle g(z_k))$ in view of the continuity of g at the points $z_1,\ldots,z_k$ and the fact that the functions $K_1,\ldots,K_k$ have compact supports.

Proof of Theorem 1. By the properties of f and the kernels K_t, Lemma 1 lets us conclude that

$$T_n = n^{r/(2r+1)}\sum_{t\in T}p_t(\hat{f}_t(z-t) - f(z-t))$$

is asymptotically normal with mean μ and variance σ^2, where

$$\mu = \sum_{t\in T} p_t f^{(r)}(z-t)\mu_r(K_t) \quad \text{and} \quad \sigma^2 = \sum_{t\in T} p_t^2 f(z_t)\langle K_t, K_t\rangle.$$

Since $\hat{p}_t - p_t = O_p(n^{-1/2})$ for all $t \in T$, we have

$$\hat{h}_*(z) = \sum_{t\in T} \hat{f}_t(z-t)\hat{p}_t = \sum_{t\in T} \hat{f}_t(z-t)p_t + O_p(n^{-1/2}).$$

Using the representation $h(z) = \sum_{t\in T} f(z-t)p_t$ we see that $n^{r/(2r+1)}(\hat{h}_*(z) - h(z))$ equals $T_n + o_p(1)$ and is therefore asymptotically normal with mean μ and variance σ^2 by Slutsky's Theorem. This is the desired result as μ and σ^2 are as asserted.

Proof of Theorem 2. Let $L_t = (L_{t,j})_{j\in J_t}$ and set

$$S_n = \sum_{t\in T} p_t\Big(\hat{f}_t(z-t) - f(z-t) - \sum_{j\in J_t} d_{t,j}b^j\big(\hat{f}_t^{(j)}(z-t) - f^{(j)}(z-t)\big)\Big)$$

with d_t as in Theorem 2. It follows from Lemma 1 that $n^{r/(2r+1)}S_n$ is asymptotically normal with mean

$$\mu = \sum_{t\in T} p_t f^{(r)}(z-t)(\mu_r(K_t) - d_t^\top \nu_t)$$

and variance

$$\sigma^2 = \sum_{t\in T} p_t^2 f(z-t)\langle K_t - d_t^\top L_t, K_t - d_t^\top L_t\rangle.$$

It follows from $\hat{p}_t - p_t = O_p(n^{-1/2})$ and the results in Section 3 that

$$\hat{h}_0(z) = S_n + o_p(n^{-r/(2r+1)}).$$

The above and Slutsky's theorem yield that $n^{r/(2r+1)}(\hat{h}_0(z) - h(z))$ is asymptotically normal with mean and variance as asserted.

Acknowledgment

We thank the referee for several suggestions that improved the presentation and the results.

References

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

[2] Chanda, K. C. (1983). Density estimation for linear processes. *Ann. Inst. Statist. Math.* **35**, 439–446.

[3] Frees, E. W. (1994). Estimating densities of functions of observations. *J. Amer. Statist. Assoc.* **89**, 517–525.

[4] Giné, E. and and Mason, D. M. (2007). On local U-statistic processes and the estimation of densities of functions of several variables. *Ann. Statist.* **35**, 1105–1145.

[5] Hallin, M. and Tran, L. T. (1996). Kernel density estimation for linear processes: asymptotic normality and optimal bandwidth derivation. *Ann. Inst. Statist. Math.* **48**, 429–449.

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

[7] Müller, U. U. (2012). Estimating the density of a possibly missing response variable in nonlinear regression. *J. Statist. Plann. Inference* **142**, 1198–1214.

[8] Müller, U. U. and Wefelmeyer, W. (2014). Estimating a density under pointwise constraints on the derivatives. *Math. Meth. Statist.* **23**, 201–209.

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

[10] Schick, A. and Wefelmeyer, W. (2004). Root n consistent density estimators for sums of independent random variables. *J. Nonparametr. Statist.* **16**, 925–935.

[11] Schick, A. and Wefelmeyer, W. (2007). Root-n consistent density estimators of convolutions in weighted L_1-norms. *J. Statist. Plann. Inference* **137**, 1765–1774.

[12] Schick, A. and Wefelmeyer, W. (2009a). Convergence rates of density estimators for sums of powers of observations. *Metrika* **69**, 249–264.

[13] Schick, A. and Wefelmeyer, W. (2009b). Non-standard behavior of density estimators for sums of squared observations. *Statist. Decisions* **27**, 55–73.

[14] Schick, A. and Wefelmeyer, W. (2012a). Convergence in weighted L_1-norms of convolution estimators for the response density in nonparametric regression. *J. Indian Statist. Assoc.* **50**, 241–261.

[15] Schick, A. and Wefelmeyer, W. (2012b). On efficient estimation of densities for sums of squared observations. *Statist. Probab. Lett.* **82**, 1637–1640.

[16] Schick, A. and Wefelmeyer, W. (2013). Uniform convergence of convolution estimators for the response density in nonparametric regression. *Bernoulli* **19**, 2250–2276.

[17] Tran, L. T. (1992). Kernel density estimation for linear processes. *Stochastic Process. Appl.* **41**, 281–296.