Adaptive Nonparametric Estimation of a Multivariate Regression Function

Y. P. MACK

Division of Statistics, University of California,
Davis, California 95616

AND

HANS-GEORG MÜLLER

University of Marburg,
Federal Republic of Germany

Communicated by P. R. Krishnaiah

We consider the kernel estimation of a multivariate regression function at a point. Theoretical choices of the bandwidth are possible for attaining minimum mean squared error or for local scaling, in the sense of asymptotic distribution. However, these choices are not available in practice. We follow the approach of Krieger and Pickands (Ann. Statist. 9 (1981) 1066–1078) and Abramson (J. Multivariate Anal. 12 (1982), 562–567) in constructing adaptive estimates after demonstrating the weak convergence of some error process. As consequences, efficient data-driven consistent estimation is feasible, and data-driven local scaling is also feasible. In the latter instance, nearest-neighbor-type estimates and variance-stabilizing estimates are obtained as special cases.

1. INTRODUCTION

Let \((X_1, Y_1),\ldots,(X_n, Y_n)\) be an i.i.d. sample such that \(X \in \mathbb{R}^d, Y \in \mathbb{R}^l\), with joint density \(p(\cdot, \cdot)\), marginal \(X\)-density \(f(\cdot)\), and regression function

\[
r(x) = E(Y | X = x) = \int y p(x, y) \, dy / f(x)
\]

\(= h(x) / f(x).
\)

(We assume \(f(x) > 0\) in this discussion.) A class of nonparametric estimates


Key words and phrases: multivariate kernel regression estimation, bias, variance, asymptotic normality, mean square error, tightness, weak convergence in \(C[a, b]\), Gaussian process, adaptation.
of $r(x)$ which has been extensively investigated are the so-called kernel estimates introduced by Nadaraya [16] and Watson [22] independently. They are of the form

$$r_n(x, b_n) = h_n(x, b_n)f_n(x, b_n),$$

with

$$h_n(x, b_n) = (nb_n^d)^{-1} \sum_j w[h_n^{-1}(x - X_j)] \cdot Y_j,$$

$$f_n(x, b_n) = (nb_n^d)^{-1} \sum_j w[h_n^{-1}(x - X_j)],$$

where $\{b_n\}$ is a sequence of positive real numbers (commonly called bandwidths) usually satisfying $b_n \to 0$, $nb_n^d \to \infty$ as $n \to \infty$, while $w(\cdot)$ is a weight function (commonly called kernel) usually taken to be a density. ($r_n(\cdot)$ will be defined as unity in the 0/0 situation.)

The quantities $h_n(x, b_n)$ and $f_n(x, b_n)$ are kernel estimates of $h(x)$ and $f(x)$, respectively. By now there is a rather substantial body of studies on these estimates. The interested reader can find extensive references in Wertz and Schneider [23] and Collomb [6]. For our purposes in this paper, it suffices to mention that kernel density estimation was initiated by Rosenblatt [18]. Subsequent works of interest include Parzen [17], Cacoullos [4], and Silverman [21]. For the study of kernel regression estimates where the predictor $X$ is random, we mention here the works of Rosenblatt [18], Johnston [8], and Mack and Silverman [13]. This is obviously a very limited listing. In the present discussion, we consider only the feasibility of data-driven local bandwidth choices. (A different approach is global bandwidth choice which can be achieved, for instance, by cross-validation.) Here the works by Krieger and Pickands [9], Abramson [1] on adaptive local bandwidth choice for density estimation, and the work of Müller and Stadtmüller [15] on adaptive local bandwidth choice for fixed-design regression estimation, are appropriate, since each of these three papers employed tightness or weak convergence arguments, as will this article.

We now give some ideas of our presentation. The point of departure is the asymptotic normality of the regression estimate at $x$ (see Rosenblatt [19] and Johnston [8] for the univariate case, the extension to the multivariate case can be treated without difficulty as in Cacoullos [4] for density estimation). Under appropriate conditions, it can be shown that $[r_n(x, b_n) - r(x)]$ is asymptotically normal (after suitable scaling) with asymptotic bias

$$\beta_n(x) = \frac{Q(rf)(x) - r(x) Q(f)(x)}{2f(x)} \cdot b_n^2(1 + o(1)).$$
where the operator $Q$ on a function $\phi$ is defined by

$$Q(\phi)(x) = \int \left[ z^T (\nabla^2 \phi(x)) z \right] w(z) \, dz,$$

$\nabla^2 \phi(x)$ is the Hessian matrix of mixed second partials of $\phi(\cdot)$ at $x$; and asymptotic variance

$$\sigma_n^2(x) = \left[ \text{var}(Y \mid X = x)/f(x) \right] \cdot \int w^2(z) \, dz \cdot (nb_n^d)^{-1}. \tag{2b}$$

A routine balancing between the contribution of $\beta_n^2(x)$ and $\sigma_n^2(x)$ towards the asymptotic mean squared error (MSE) results in the following optimal (in the sense of minimizing the asymptotic MSE) choice $h_n^*$ of the bandwidth at a point

$$h_n^* = t^* \cdot n^{1/(d+4)}, \tag{3}$$

where

$$(t^*)^{d+4} = \frac{d \cdot \text{var}(Y \mid X - x) \cdot f(x) \cdot \int w^2(z) \, dz}{\left[ Q(rf)(x) - r(x) Q(f)(x) \right]^2}. \tag{4}$$

This optimal choice requires knowledge at the point $x$ of unknown functions, and is thus not available in practice. The question arises whether one can use a pilot estimate $\hat{t}^*$ of $t^*$ form a data-driven bandwidth sequence $\hat{h}_n^*$ in such a way that $r_n(s, \hat{h}_n^*)$ is as efficient as $r_n(s, h_n^*)$. For the kernel density estimation case, Krieger and Pickands [9] and Abramson [1] both answered in the positive. Their methods of attack involved tightness and weak convergence of some error process. We show in this article that similar arguments apply also to the kernel regression case.

Another application of the weak convergence result is on “local scaling.” If we set the bandwidth

$$h_n(x) = \theta(x) \cdot n^{1/(d+4)}, \tag{5a}$$

then the asymptotic bias and variance of $r_n(x, h_n(x))$ is now given as

$$\hat{\beta}_n(x) = \left\{ \frac{Q(rf)(x) - r(x) Q(f)(x))}{2f(x)} \right\} \theta(x)^2 \cdot n^{2/(d+4)} \cdot (1 + o(1)) \tag{5b}$$

$$\hat{\sigma}_n^2(x) = \left[ \text{var}(Y \mid X = x)/f(x) \right] \cdot \int w^2(z) \, dz \frac{n^{-4/(d+4)}}{\theta(x)^d}. \tag{5b}$$

Hence, for example, if we set $\theta(x)^d = f(x)$, then the asymptotic variance (5b) is no longer sensitive to the variation of $f(x)$ as $x$ varies. The weak
convergence result will allow us to replace \( \theta(x) \) by a consistent pilot estimate \( \hat{\theta}_n(x) \) from the same data set. For instance, if we employ the following pilot estimate of \( \theta(x) \),

\[
\hat{\theta}_n(x) = \left[ k_n/n B_n^{(d)}(x) \right]^{-1/d},
\]

where \( \{k_n\} \) is a sequence of positive integers such that \( k_n \to \infty \), \( k_n/n \to 0 \) as \( n \to \infty \), \( c_d \) is the volume of the unit sphere in \( \mathbb{R}^d \), and

\[
B_n(x) = \text{distance (}\ x, k\text{th nearest neighbor of } x\text{ among } X\text{'s}),
\]

then \( \hat{\theta}(x) \) is a consistent estimate of \( \theta(x) \) (see Loftsgaarden and Quesenberry [10]). Thus adaptation is possible. In particular, for the choice \( k_n = n^{-2d/4} \), we have

\[
\hat{\theta}_n(x) = \hat{\theta}_n(x) \cdot n^{1/d + 4} = c_d B_n(x).
\]

With this choice, we essentially arrive at the \( k \)-nearest neighbor-type regression estimates considered in Mack [12], if the kernel \( w(\cdot) \) is replaced by

\[
w^*(\cdot) = c_d^{-1} w(c_d^{-1/d} \cdot).
\]

In this case we can “read off” the choice of the bandwidth from (5a) and (5b) to arrive at the asymptotic bias

\[
\beta_n(x) = \left[ Q(rf)(x) - r(x) Q(f)(x) \right] \cdot c_d^{-2d} \cdot f(x) \cdot n^{-2/d + 4} \cdot (1 + o(1))
\]

and the asymptotic variance

\[
\sigma_n^2(x) = \left\{ \left[ \text{var}(Y|X = x) f(x) \right] \cdot c_d \right\} \cdot w^*(z) dz \cdot f(x) \cdot n^{-4/d + 4},
\]

where the operator \( Q \) is understood to be defined w.r.t. \( w^*(\cdot) \). It is easily checked that these expressions are identical to expressions (9) and (10) in Mack [12], where the bias and variance are derived by direct computation, if we choose \( k_n = n^{-2d/4} \). (We remark here that direct evaluation of bias and variance of \( r_n(x, h_n) \) may be impossible unless some tail decay condition is assumed on the marginal \( X \)-distribution.)

Another local scaling is achieved by setting \( \theta(x)/d = \text{var}(Y|X = x)/f(x) \). In this case the asymptotic variance \( \sigma_n^2(x) \) becomes constant for all \( x \). Therefore confidence intervals based on asymptotic distribution of
$r_n(x, b_n(x))$ at the point $x$ have the same width for all $x$, if they are centered at $E r_n(x, b_n(x))$. In practical applications of function estimation this kind of "variance stabilization" can be a desirable feature.

As a final remark to this section, we would like to point out that consistent pilot estimates are available in each of the three cases considered above. For example, one can extend the technique in Schuster and Yakowitz [20] to construct consistent estimates of $Q(rf)(x)$. For the estimation of $\text{var}(Y|X=x)$, one can follow the method of Carroll [5] by first estimating $E(Y^2|X=x)$ as another regression problem from the data \{(X_i, Y_i^2), i = 1,\ldots, n\}.

2. Preliminaries

We first describe the conditions necessary for our analyses. The kernel $w(\cdot)$ will be assumed to satisfy one or more of the following collection of conditions:

(W1) $\int w(z) \, dz = 1$,
(W2) $w(z) \geq 0$ for all $z$,
(W3) $w(\cdot)$ is compactly supported,
(W4) $\int z_i w(z) \, dz = 0$ for all components $z_i$ of the vector $z$,
(W5) the first partial derivatives of $w(\cdot)$ exist and are bounded.

Some of these conditions can obviously be weakened, but are kept as stated for ease of presentation. Condition (W2) can be relaxed to allow for kernels with vanishing moments. Such kernels which take on negative values are used frequently to reduce bias in case a higher order of smoothness of $r(\cdot)$ and $f(\cdot)$ can be assumed. (See Gasser and Müller [7] and Müller [14] for detailed discussions of such kernels in the fixed-design regression estimation.) Condition (W5) is used in proving tightness of the processes defined in Section 2. A similar condition was imposed by Abramson [1] in proving tightness of the kernel density error process. This condition also guarantees that all weak convergence take place in the space $C[a,b]$. Note also that (W3) and (W5) imply that $w(\cdot)$ is bounded, continuous, and has finite total variation. These are also the usual conditions on $w(\cdot)$ one encounters in kernel estimation literature.

Next, we give two sets of conditions on the functions involved in our analysis:

(A1) the density $f(\cdot)$ is continuous at $x$;
(A2) the function $h(\cdot) = \int yp(\cdot, y) \, dy$ exists and is continuous at $x$;
(A3) the function $g(\cdot) = \int y^2 p(\cdot, y) \, dy$ exists and is continuous at $x$;
(A4) the function \( m(\cdot) = \int |y|^{2+\delta} p(\cdot, y) \, dy \) exists for some \( \delta > 0 \), is integrable w.r.t. Lebesgue measure in \( \mathbb{R}^d \), and is continuous at \( x \).

(B1) all mixed partial derivatives of \( f(\cdot) \) at \( x \) exist up to the second order, with the second partials continuous at \( x \);

(B2) all mixed partial derivatives of \( r(\cdot) \) at \( x \) exist up to the second order, with the second partials continuous at \( x \).

For the remainder of this article, we will consider only the bandwidth of the form given in (3).

\[ b_n(t) = t \cdot n^{-1/2(\alpha + 1)} \quad 0 < a \leq t \leq b < \infty, \]

where the point \( x \) where the curve is to be estimated is fixed, and we assume that \( f(x) > 0 \). For local bandwidth adaptation considerations, we assume the target value \( t^* \) is always contained in \( (a, b) \).

Also, from now on we will write \( r_n(t) \) in place of \( r_n(x, b_n(t)) \), \( f_n(t) \) in place of \( f_n(x, b_n(t)) \), and so on, to relieve the burden of notation.

Central to our study is the error process

\[ R_n(t) = n^{2(\alpha + 1)} [r_n(t) - r(x)], \quad a \leq t \leq b. \]

It can be rewritten as

\[ R_n(t) = \frac{1}{f(x)} \left[ \psi_n(t) + Y_n(t) \right] - \frac{r_n(t)}{f(x)} \left[ \zeta_n(t) + Z_n(t) \right], \]

where

\[ \psi_n(t) = n^{2(\alpha + 1)} [Eh_n(t) - h(x)], \]

\[ \zeta_n(t) = n^{2(\alpha + 1)} [Ef_n(t) - f(x)], \]

\[ Y_n(t) = n^{2(\alpha + 1)} [h_n(t) - Eh_n(t)], \]

\[ Z_n(t) = n^{2(\alpha + 1)} [f_n(t) - Ef_n(t)]. \]

We will show that the convergences

\[ \psi_n(t) \to \psi(t) \equiv \frac{1}{2} Q(rf)(x) \, t^2, \]

\[ \zeta_n(t) \to \zeta(t) \equiv \frac{1}{2} Q(f)(x) \, t^2; \]

are uniform on \( [a, b] \), and that \( Y_n(t) \Rightarrow Y(t) \), \( Z_n(t) \Rightarrow Z(t) \) for some appropriate Gaussian processes \( Y(\cdot) \) and \( Z(\cdot) \). Here \( \Rightarrow \) means weak convergence in the function space \( C[a, b] \) with the sup-norm (see Chap. 2 of
Billingsley [3] for a detailed discussion about weak convergence in \( C[a, b] \). Hence, by a Slutsky-type argument, we have that
\[
R_n(t) \rightarrow R(t)
\]
for some Gaussian process \( R(\cdot) \) with mean \( f(x)^{-1} [\psi(t) - r(x) \xi(t)] \) and covariance determined by the covariance of \( Y(\cdot), Z(\cdot) \) individually as well as the cross-covariance \( \text{cov}(Y(\cdot), Z(\cdot)) \). As in Krieger and Pickands [9], it will then follow that
\[
n^{2/(d+4)} [r_n(t^*) - r_n(i^*)] \rightarrow^p 0,
\]
where we recall that \( i^* \) is a consistent pilot estimate of \( t^* \). (\( \rightarrow^p \) here means convergence in probability as \( n \rightarrow \infty \)).

In order to facilitate our main discussions in Sections 3 and 4, we state the following results which are either established in standard literature or can be derived easily. (For a good reference, see Johnston [8].) We leave the details to the reader.

**Lemma 1.** (Asymptotic normality of \( Y_n(t_0) \) and \( Z_n(t_0) \) at a single point \( t_0 \)). Assume (W3) and (W5) hold in both (a) and (b):

(a) Suppose (A3), (A4) hold. Then
\[
Y_n(t_0) \xrightarrow{\mathcal{D}} N(0, \left[ E(Y^2 \mid X = x) f(x) \int w^2(z) \, dz \right] \cdot t_0^{-d})
\]
(b) Suppose (A1) holds. Then
\[
Z_n(t_0) \xrightarrow{\mathcal{D}} N(0, \left[ f(x) \int w^2(z) \, dz \right] \cdot t_0^{-d}).
\]
(\( \xrightarrow{\mathcal{D}} \) here means convergence in distribution as \( n \rightarrow \infty \).)

**Lemma 2** (Bias of \( h_n(t) \) and \( f_n(t) \)). Assume (W1)-(W5) hold in both (a) and (b):

(a) Suppose (B2) holds, then
\[
\sup_{a \leq t \leq b} |\psi_n(t) - \frac{1}{2} Q(rf)(x) \, t^2| \rightarrow 0 \quad \text{as} \quad n \rightarrow \infty.
\]
(b) Suppose (B1) holds, then
\[
\sup_{a \leq t \leq b} |\xi_n(t) - \frac{1}{2} Q(f)(x) \, t^2| \rightarrow 0 \quad \text{as} \quad n \rightarrow \infty.
\]
3. Weak Convergence of the Processes $Y_n(\cdot)$ and $Z_n(\cdot)$

The main tools to prove the functional weak convergence of these processes will be Theorems 8.1 and 12.3 of Billingsley [3]. The first of these theorems shows that we can establish weak convergence by proving convergence of the finite-dimensional distributions to the corresponding distributions of the limiting process, and tightness. The second theorem gives sufficient conditions for tightness which are applicable in the present context. Considering first the process $Y_n(t) = n^{2/(d+4)}(h_n(t) - Eh_n(t))$, we have to discuss the finite-dimensional distributions. For this purpose we derive the covariance of $Y_n(t)$.

**Lemma 3.** Assume that (A3), (W3), and (W5) are satisfied. For $t, s \in [a, b]$, 

$$\text{cov}(Y_n(t), Y_n(s)) = \frac{f(x)E(Y^2 | X = x)}{(st)^d} \int w\left(\frac{u}{t}\right)w\left(\frac{u}{s}\right)du + o(1).$$

**Proof.** Owing to the independence of $(I', Y_j)$ and $(I', I')$ for $i \neq j$, we obtain 

$$\text{cov}(Y_n(t), Y_n(s)) = \frac{1}{n^{4/(d+4)}(st)^d} \left( E \left( \sum_{i,j} w\left(\frac{x-X_i}{b_n(t)}\right) Y_i w\left(\frac{x-X_j}{b_n(s)}\right) Y_j \right) 
- \sum_{i,j} E \left( w\left(\frac{x-X_i}{b_n(t)}\right) Y_i \right) E \left( w\left(\frac{x-X_j}{b_n(s)}\right) Y_j \right) \right)$$

$$= \frac{1}{n^{4/(d+4)}(st)^d} \left( E \left( w\left(\frac{x-X_i}{b_n(t)}\right) w\left(\frac{x-X_j}{b_n(s)}\right) Y_i \right) 
- E \left( w\left(\frac{x-X_i}{b_n(t)}\right) Y_i \right) E \left( w\left(\frac{x-X_j}{b_n(s)}\right) Y_j \right) \right)$$

$$= \frac{1}{n^{4/(d+4)}(st)^d} \left( \int w\left(\frac{u}{t}\right)w\left(\frac{v}{s}\right)v^2p(x - b_n(u), v) du dv 
- n^{-4/(d+4)} \left( \int w\left(\frac{u}{t}\right)v p(x - b_n(u), v) du dv \right)^2 \right)$$

$$= \frac{1}{n^{4/(d+4)}(st)^d} \int g(x - b_n(u)) w\left(\frac{u}{t}\right)w\left(\frac{u}{s}\right)du + o(1)$$

by Fubini's theorem. The result follows by observing the continuity of $g(\cdot)$ at $x$. 

Lemma 4. Under (A3), (W3), and (W5), the sequence \( \{ Y_n(t), t \in [a, b] \} \) of random elements of \( C[a, b] \) is tight.

Proof. Observe that \( Y_n(\cdot) \in C[a, b] \) because of the continuity of the kernel function \( w \). For a fixed \( t_o \in [a, b] \), \( \{ Y_n(t_o) \} \) is tight by our Lemma 1 and Theorem 6.2 of Billingsley. If we prove that

\[
E(|Y_n(t) - Y_n(s)|^2) \leq \lambda |t - s|^2
\]

for some constant \( \lambda > 0 \), the tightness of \( \{ Y_n(t), t \in [a, b] \} \) follows from Theorem 12.3 and formula (12.51) (choosing \( \gamma = 2 \) and \( F = \text{identity} \)) of Billingsley. In order to prove (13), first set \( \hat{w}(x) = dw(x) + x^T(\nabla w(x)) \), where \( \nabla w(x) \) is the gradient vector of \( w \) at \( x \). Note that \( (d/dt)(1/t^d) w((x - X_i)/t) e^{-t} = -(1/t^{d+1}) \hat{w}((x - X_i)/t) \). Next, by the mean value theorem, there exist \( t_1, t_2 \) between \( s \) and \( t \) such that

\[
\frac{1}{t^d} w\left( \frac{x - X_i}{t} \right) - \frac{1}{s^d} w\left( \frac{x - X_i}{s} \right) = \left( -\frac{1}{t_{i+1}^d} \hat{w}\left( \frac{x - X_i}{t_i} \right) \right) (t - s), \quad i = 1, \ldots, n.
\]

Therefore, by the continuity of \( g \) at \( x \) and the fact that \( \hat{w} \) has compact support and is bounded, we have

\[
E(Y_n(t) - Y_n(s))^2 \leq (t - s)^2 n \frac{1}{a^{d+1}} \int \left( \int \hat{w}^2(z) \nu^2 p(x - z b_n(t_i), \nu) \, d\nu \right) \, dz \\
\leq (t - s)^2 n \frac{1}{a^{d+1}} g(x) \left( \int \hat{w}^2(z) \, dz + O(1) \right).
\]

whence the assertion follows.

Proposition 1. Under (A3), (A4), (W3), and (W5), we have

\( Y_n(t) \Rightarrow Y(t) \).
where $Y(t)$ is a mean-zero Gaussian process with covariance

$$\text{cov}(Y(t), Y(s)) = \int \frac{f(x) E(Y^2 | X = x)}{(s t)^d} w\left(\frac{u}{s}\right) w\left(\frac{u}{t}\right) du.$$  

**Proof.** In view of Lemma 4, it remains to show the convergence of the finite-dimensional distributions of $Y_n(t)$ to the corresponding distributions of $Y(t)$. In Lemma 1, the asymptotic normality at a fixed point $t, \in [a, b]$ was derived. The joint asymptotic normality of $Y_n$ at finitely many fixed points within $[a, b]$ is a consequence of Lemmas 1 and 3 and the Cramér–Wold device (Billingsley, Theorem 7.7). It can be easily verified that this limiting distribution agrees with the corresponding finite-dimensional distribution of $Y(\cdot)$.

For the remainder of this section, we carry out an analogous analysis for the process $Z_n(t) = n^{2d+4}(f_n(t) - Ef_n(t))$. Here, the tightness was already proven by Abramson [1].

**Lemma 5.** Assume that (A1), (W3), and (W5) are satisfied. For $t, s \in [a, b]$,

$$\text{cov}(Z_n(t), Z_n(s)) = \int \frac{f(x)}{(s t)^d} w\left(\frac{u}{s}\right) w\left(\frac{u}{t}\right) du + o(1).$$

**Proof.**

$$\text{Cov}(Z_n(t), Z_n(s)) = \frac{1}{(s t)^d} \left( \int w\left(\frac{u}{s}\right) w\left(\frac{u}{t}\right) f(x - b_n(u)) du \right) - n \left[ \int w\left(\frac{u}{t}\right) f(x - b_n(u)) du \right]^2.$$  

The result follows by the continuity of $f(\cdot)$ at $x$ in an analogous manner to the proof of Lemma 3.

The following result was given by Abramson [1]. He required existence of second order partial derivatives of the marginal density $f(\cdot)$ at $x$, whereas we require here only continuity.

**Lemma 6.** Under (A1), (W3), and (W5), the sequence $\{ Y_n(t), t \in [a, b] \}$ of random elements of $C[a, b]$ is tight.

**Proof.** Following the lines of the proof of Lemma 4, we obtain

$$E(Y_n(t) - Y_n(s))^2 \leq (t - s)^2 \frac{1}{d^{d+1}} f(x) \left( \int w^2(z) dz + o(1) \right).$$
We now state the second central result of this section. The proof is similar to that of Proposition 1 and is therefore omitted. It is based on Lemmas 1, 5, and 6.

**Proposition 2.** Under (A1), (W3), and (W5), we have

\[ Z_n(t) \Rightarrow Z(t). \]

where \( Z(t) \) is a mean-zero Gaussian process with covariance

\[ \text{cov}(Z(t), Z(s)) = \frac{f(x)}{(st)^{1/2}} \int w\left(\frac{t}{s}\right) w\left(\frac{u}{x}\right) du. \]

4. **Weak Convergence of the Process \( R_n(\cdot) \)**

We first establish uniform convergence in probability of the factor \( c_n(t) = r_n(t)/f(x) \) appearing in (11) towards the limit \( c(x) = r(x)/f(x) \).

**Lemma 7.** Under (A1), (A3), (A4), (B1), (B2), (W1)-(W5),

\[ \sup_{a \leq t \leq b} |r_n(t) - r(x)| \overset{p}{\longrightarrow} 0 \quad \text{as } n \to \infty. \]

**Proof.** From Propositions 1 and 2 we infer by the continuous mapping theorem (cf. Billingsley, Theorem 5.1 and Corollary 1) that

\[ \sup_{a \leq t \leq b} |Z_n(t)| \overset{\mathcal{D}}{\longrightarrow} \sup_{a \leq t \leq b} |Z(t)| \]

and

\[ \sup_{a \leq t \leq b} |Y_n(t)| \overset{\mathcal{D}}{\longrightarrow} \sup_{a \leq t \leq b} |Y(t)|. \]

It follows that

\[ \sup_{a \leq t \leq b} |f_n(t) - Ef_n(t)| - n^{3d+4} \sup_{a \leq t \leq b} |Z_n(t)| \overset{p}{\longrightarrow} 0 \quad \text{as } n \to \infty \]

by Slutsky's theorem. A similar argument applies to the uniform convergence of \( h_n(t) \). Finally, an application of Lemma 2 yields the result.

In order to establish our main result, we still need the covariance between the processes \( Y_n(\cdot) \) and \( Z_n(\cdot) \).
Lemma 8. Under (A2), (W3), and (W5), we have for $t, s \in [a, b]$,
\[
\text{cov}(Y_n(t), Z_n(s)) = \frac{r(x) f(x)}{(st)^d} \left\{ \int w\left( \frac{u}{t} \right) w\left( \frac{u}{s} \right) du + o(1) \right\}.
\]

Proof. We calculate
\[
\text{cov}(Y_n(t), Z_n(s))
\]
\[
= \frac{1}{(st)^d} \left( \int \int w\left( \frac{u}{t} \right) w\left( \frac{u}{s} \right) dp(x - b_n(u), v) du dv \right.
\]
\[
- \int w\left( \frac{u}{t} \right) \int w\left( \frac{u}{s} \right) dp(x - b_n(u), v) du dv
\]
\[
\times \left( \int w\left( \frac{u}{s} \right) f(x - b_n(u)) du \right)
\]
whence the result follows by the continuity of $h(\cdot)$ at $x$, compare the proof of Lemma 3.

We are now ready to state our main result. Recall that
\[
R_n(t) = \frac{1}{f(x)} (\psi_n(t) + Y_n(t)) - \frac{r_n(t)}{f(x)} (\zeta_n(t) + Z_n(t)).
\]

Theorem. Under (A1), (A4), (B1), (B2), (W1)-(W5), we have
\[
R_n(t) \Rightarrow R(t),
\]
where $R(\cdot)$ is a Gaussian process with
\[
ER(t) = \frac{(Q(f) - r(x) Q(f))}{2f(x)} \cdot t^2
\]
and
\[
\text{cov}(R(t), R(s)) = \var(\frac{Y|X=x}{f(x)}) \frac{1}{(st)^d} \left\{ \int w\left( \frac{u}{t} \right) w\left( \frac{u}{s} \right) du \right\}.
\]

Remark. If $d = 1$, $ER(t)$ simplifies to
\[
ER(t) = \left( \frac{r'(x)f'(x)}{f(x)} + \frac{r''(x)}{2} \right) \int w(z) z^2 dz \cdot t^2.
\]
Proof. First we show that
\[ c_n(t) Z_n(t) \Rightarrow c(x) Z(t). \]  
(14)

By Theorem 8.1 of Billingsley, we have to show weak convergence of the finite-dimensional distributions and tightness. Convergence of the finite-dimensional distributions follows from Proposition 2, Lemma 7, the Cramér–Wold device and Slutsky’s theorem. Since for a fixed \( t_n \in [a, b] \),
\[ c_n(t_n) Z_n(t_n) \rightarrow c(x) Z(t_n), \]
the sequence \( \{c_n(t_n) Z_n(t_n)\} \) is tight by Theorem 6.2 of Billingsley. By Theorem 8.3 of Billingsley in the form of (8.12), it then suffices to prove the following statement in order to infer tightness of \( \{c_n(\cdot) Z_n(\cdot)\} \) on \( C[a, b] \) and therefore (14):

For each positive \( \varepsilon \) and \( \eta \), there exist a \( \delta, 0 < \delta < 1 \), and an integer \( n_\eta \) such that for all \( n > n_\eta \) and \( t \in [a, b] \),
\[ \frac{1}{\delta} P \left( \sup_{t \leq s \leq t + \delta} |c_n(s) Z_n(s) - c_n(t) Z_n(t)| > \varepsilon \right) \leq \eta. \]  
(15)

In order to prove this, we observe that there exists a positive \( \eta = \eta(\varepsilon) \) such that
\[ P \left( \sup_{t \leq s \leq t + \delta} |c_n(s) Z_n(s) - c(x) Z_n(x)| \geq \eta \right) \]
\[ \leq P \left( \sup_{t \leq s \leq t + \delta} |c_n(s) Z_n(s) - c(x) Z_n(s)| \geq \eta \right) \]
\[ + P \left( \sup_{t \leq s \leq t + \delta} |c(x) Z_n(s) - c(x) Z_n(t)| \geq \eta \right) \]
\[ + P \left( |c_n(t) Z_n(t) - c(x) Z_n(t)| \geq \eta \right). \]

The second term on the right can be made arbitrarily small by appropriate choices of \( \delta \) and \( n_\eta \) because of the tightness of \( \{Z_n(\cdot)\} \). The third term is dominated by the first term. The first term is bounded by
\[ P \left( \sup_{a \leq s \leq b} |c_n(s) - c(x)| \cdot \sup_{a \leq s \leq b} |Z_n(s)| \geq \eta \right). \]

Since by the continuous mapping theorem,
\[ \sup_{a \leq s \leq b} |Z_n(s)| \xrightarrow{\mathcal{L}} \sup_{a \leq s \leq b} |Z(s)|, \]
Lemma 7 implies that this term can be made arbitrarily small and therefore (15) is satisfied.

The weak convergence of \( \{R_n(\cdot)\} \) on \( C[a, b] \) to a Gaussian process now follows from (14) and Proposition 1. The covariance structure of the
limiting process is derived by simple algebra from Lemmas 3, 5, and 8. The expectation of the limiting process is found by Lemma 2 and Slutsky's theorem.

REFERENCES