Convergence in weighted L_1-norms of convolution estimators for the response density in nonparametric regression

Anton Schick and Wolfgang Wefelmeyer

Abstract. Consider a nonparametric regression model $Y = r(X) + \varepsilon$ with a random covariate X that is independent of the error ε. Then the density of the response Y is a convolution of the densities of ε and $r(X)$. It can therefore be estimated by a convolution of two kernel estimators for these densities, or more generally by a local von Mises statistic. If the regression function has a nowhere vanishing derivative, then these estimators are known to converge at the root-n rate, and convergence holds uniformly. We show that convergence also holds in weighted L_1-norms, and under weaker smoothness assumptions on the error density. It follows that the corresponding process obeys a functional central limit theorem in the corresponding L_1-space.

1. Introduction

We consider the nonparametric regression model $Y = r(X) + \varepsilon$ with a one-dimensional random covariate X that is independent of the unobservable error variable ε. We impose the following assumptions:

(F) The error variable ε has mean zero, a finite variance σ^2 and a density f of the form

$$f(z) = \int_{-\infty}^{z} f'(x)\,dx, \quad z \in \mathbb{R},$$

for some integrable function f' of bounded variation.

(G) The covariate X is *quasi-uniform* on the interval $[0,1]$ in the sense that its density g is bounded and bounded away from zero on the interval and vanishes outside. Furthermore, g is of bounded variation.

(R) The unknown regression function r is twice continuously differentiable on $[0,1]$, and r' is strictly positive on $[0,1]$.

We observe independent copies $(X_1, Y_1), \dots, (X_n, Y_n)$ of (X, Y), and we are interested in estimating the density h of the response Y. An obvious estimator is the kernel estimator

$$\tilde{h}(y) = \frac{1}{n}\sum_{j=1}^{n} K_b(y - Y_j), \quad y \in \mathbb{R},$$

Key words and phrases. Density estimator, local von Mises statistic, local U-statistic, local polynomial smoother, monotone regression function.

Anton Schick was supported by NSF Grant DMS 0906551.

where $K_b(t) = K(t/b)/b$ for some kernel K and some bandwidth b. Under the above assumptions on f and g, the density h has a second derivative h'' that is L_1-*Lipschitz*: For a constant L,

$$\int |h''(y+t) - h''(y)|\, dy \le L|t|, \quad t \in \mathbb{R}.$$

Thus, if the kernel is bounded, compactly supported and of order three, and the bandwidth b is chosen proportional to $n^{-1/7}$, then one can show that

$$\|\tilde{h} - h\|_1 = \int |\tilde{h}(y) - h(y)|\, dy = O_p((nb)^{-1/2} + b^3) = O_p(n^{-3/7}).$$

The kernel estimator $\tilde{h}$ neglects the structure of the regression model. We shall see that by exploiting this structure one can construct estimators that have the faster (parametric) root-n rate of convergence in the L_1-norm and obey a (functional) central limit theorem in the space L_1. For this we observe that the density h is the convolution of the error density f and the density q of $r(X)$. The latter density is given by

$$q(z) = \frac{g(r^{-1}(z))}{r'(r^{-1}(z))}, \quad z \in \mathbb{R}.$$

By our assumptions on r and g, the density q is quasi-uniform on the interval $[r(0), r(1)]$, which is the image of $[0,1]$ under r. Furthermore, q is of bounded variation.

The convolution representation $h = f * q$ suggests a plug-in estimator or *convolution estimator* $\hat{h} = \hat{f} * \hat{q}$ based on kernel estimators $\hat{f}$ and $\hat{q}$ of f and q,

$$\hat{f}(x) = \frac{1}{n}\sum_{j=1}^{n} k_b(x - \hat{\varepsilon}_j) \quad \text{and} \quad \hat{q}(x) = \frac{1}{n}\sum_{j=1}^{n} k_b(x - \hat{r}(X_j)), \quad x \in \mathbb{R},$$

with *residuals* $\hat{\varepsilon}_j = Y_j - \hat{r}(X_j)$ and $\hat{r}$ a nonparametric estimator of r. Setting $K = k * k$, the convolution estimator has the form of a *local von Mises statistic*,

$$\hat{h}(y) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} K_b(y - \hat{\varepsilon}_i - \hat{r}(X_j)), \quad y \in \mathbb{R},$$

and is thus easy to calculate.

Assuming that f also has a finite moment of order higher than $8/3$ and a bounded integrable second derivative, Schick and Wefelmeyer (2011) have shown that the estimator $\hat{h}$ is root-n consistent in the sup-norm and obeys a functional central limit theorem in the space $C_0(\mathbb{R})$ of all continuous functions on $\mathbb{R}$ that vanish at plus and minus infinity. This result was obtained under mild assumptions on the kernel k and the bandwidth b and with $\hat{r}$ an under-smoothed local quadratic smoother. We shall also work with such an estimator. As in their proof we rely on two properties of this estimator of r: its bias is uniformly very small (of order $o(n^{-1/2})$), and its precision is uniformly sufficiently good (considerably better than $n^{-1/4}$). These two criteria can be met by requiring sufficient smoothness on r and working with higher-order local polynomial smoothers. For a twice continuously differentiable r and a local quadratic smoother, the bias is of order $o(c^2)$, and the precision of order $(\log n/(nc))^{1/2}$, where c is the chosen bandwidth. Choosing c proportional to $n^{-1/4}$ results in the required uniformly small bias and a uniform precision of order $\log n^{1/2} n^{-3/8}$.

The local quadratic smoother is defined as follows. For a fixed x in $[0,1]$, the estimator $\hat{r}(x)$ is the first coordinate $\hat{\beta}_1(x)$ of the weighted least squares estimator

$$\hat{\beta}(x) = \arg\max_{\beta} \frac{1}{nc} \sum_{j=1}^{n} w\Big(\frac{X_j - x}{c}\Big)\Big(Y_j - \beta^\top \psi\Big(\frac{X_j - x}{c}\Big)\Big)^2 \tag{1.1}$$

where $\psi(x) = (1, x, x^2)^\top$. We make the following assumptions on the weight function w and the bandwidth c.

(W) The weight function w is a continuously differentiable symmetric density with compact support $[-1,1]$.

(C) The bandwidth c is proportional to $n^{-1/4}$.

Here we do not require w to be three times continuously differentiable as was assumed in Schick and Wefelmeyer (2011). This additional smoothness assumption was used there to show that a modification of $\hat{r}$ belongs to some Hölder space of appropriate order. We shall not need this property here. As we assumed that r is two times continuously differentiable, a bandwidth c proportional to $n^{-1/5}$ would yield optimal rates of convergence. Here we undersmooth to obtain the expansion (1.5) below. This comes at the expense of slower rates of convergence. More precisely, as in Schick and Wefelmeyer (2011), our undersmoothed local quadratic estimator possesses the following properties.

Lemma 1. *Suppose (F), (G), (R), (W) and (C) hold and ε has a finite moment of order greater than $8/3$. Then, with $\hat{a} = \hat{r} - r$, we have the following rates of convergence*

$$\sup_{0 \le x \le 1} |\hat{a}(x)| = O_p\Big(\Big(\frac{\log n}{nc}\Big)^{1/2}\Big), \tag{1.2}$$

$$\frac{1}{n}\sum_{j=1}^{n} \hat{a}^2(X_j) = O_p\Big(\frac{1}{nc}\Big), \tag{1.3}$$

$$\int \hat{a}^2(x) g(x)\,dx = O_p\Big(\frac{1}{nc}\Big), \tag{1.4}$$

and the stochastic expansion

$$\int \hat{a}(x) g(x)\,dx = \frac{1}{n}\sum_{j=1}^{n} \varepsilon_j + o_p(n^{-1/2}). \tag{1.5}$$

We are now ready to state the root-n consistency of the local von Mises statistic $\hat{h}$ in the L_1-norm. For this we set

$$H_1(y) = \frac{1}{n}\sum_{j=1}^{n} \big(q(y - \varepsilon_j) - h(y) + \varepsilon_j h'(y)\big), \quad y \in \mathbb{R},$$

$$H_2(y) = \frac{1}{n}\sum_{j=1}^{n} \big(f(y - r(X_j)) - h(y) - \varepsilon_j f'(y - r(X_j))\big), \quad y \in \mathbb{R},$$

and introduce the following assumptions on the kernel k and the bandwidth b.

(K) The kernel k has compact support, is twice continuously differentiable and of order three.

(B) The bandwidth b satisfies $nb^6 \to 0$ and $nb^4/\log^4 n \to \infty$.

THEOREM 1. *Assume that (F), (G), (R), (W), (C), (K) and (B) hold. Let ε have a finite moment of order greater than 8/3, and let the integral*

$$\int \big(1+\log(1+|x|)\big)^2(1+|x|)(f'(x))^2\,dx$$

be finite. Let $\hat r$ be the local quadratic smoother defined in (1.1). Set $H = H_1 + H_2$. Then the stochastic expansion

$$\|\hat h - h - H\|_1 = o_p(n^{-1/2}) \tag{1.6}$$

holds, and $n^{1/2}H$ converges in distribution in the space L_1 to a centered Gaussian process.

We obtain this theorem as a special case of a more general result for weighted L_1-spaces. By weighted L_1-spaces we mean L_1-spaces for measures which have a density V with respect to the Lebesgue measure. As in Schick and Wefelmeyer (2007a) we treat continuous V with special properties; see (2.1)–(2.3) below.

Our main result, Theorem 3, is presented and proved in Section 2. This result states that our convolution estimator is root-n consistency in weighted L_1-spaces and obeys a central limit theorem in these spaces. More precisely, it gives the analogue of the stochastic expansion (1.6) for the weighted L_1-norm and the convergence in distribution of $n^{1/2}H$ in the corresponding space. For the latter we utilize the central limit theorem for general L_1-spaces for σ-finite measures. We give a simple sufficient condition for the central limit theorem to hold in these spaces, see Corollary 1, and use it to derive in Lemma 2 the limiting distribution of $n^{1/2}H$. The proof of the analogue of (1.6) reduces the desired result to expansions for the density estimators $\hat f$ and $\hat q$. These expansions are proved in Sections 4 to 6. Their proofs rely on additional properties of the local quadratic smoother which are presented in Section 3.

Local von Mises statistics go back to Frees (1994). He observed that densities of certain (known) transformations $T(X_1,\dots,X_m)$ of $m \ge 2$ independent and identically distributed random variables $X_1,\dots,X_m$ can be estimated pointwise at the parametric rate by a local U-statistic. Saavedra and Cao (2000) consider the transformation $T(X_1,X_2) = X_1+\varphi X_2$ with $\varphi \ne 0$. Schick and Wefelmeyer (2004b) and (2007a) obtain this rate in the sup-norm and in L_1-norms for transformations of the form $T(X_1,\dots,X_m) = T_1(X_1)+\dots+T_m(X_m)$ and $T(X_1,X_2) = X_1+X_2$. Giné and Mason (2007) obtain such functional results in L_p-norms for $1 \le p \le \infty$ and general transformations $T(X_1,\dots,X_m)$. The results of Nickl (2007) and (2009) are also applicable in this context.

The same convergence rates have been obtained for local von Mises statistics or convolution estimators of the stationary density of linear processes. Saavedra and Cao (1999) treat pointwise convergence for a first-order moving average process. Schick and Wefelmeyer (2004a) and (2004c) consider higher-order moving average processes and convergence in L_1, and Schick and Wefelmeyer (2007b), (2008a) and (2009a) obtain parametric rates in the sup-norm and in L_1 for estimators of the stationary density of invertible linear processes. Analogous pointwise convergence results for response density estimators in nonlinear regression (with responses missing at random) and in nonparametric regression are in Müller (2010) and Støve and Tjøstheim (2010), respectively. Escanciano and Jacho-Chávez (2011) consider the nonparametric regression model and show uniform convergence on compact sets of their local U-statistic. Their results allow for a multivariate covariate X, but require the density of $r(X)$ to be bounded and Lipschitz.

In the above applications to regression models and time series, and also in the present paper, the (auto-)regression function is assumed to have a nonvanishing derivative. This assumption is essential. Suppose there is a point x at which the regression function behaves like $r(y) = r(x) + c(y-x)^\nu + o(|y-x|^\nu)$, for y to the left or right of x, with $\nu \geq 2$. Then the density q of $r(X)$ has a strong peak at $r(x)$. This slows down the rate of the convolution density estimator or local von Mises statistic for $h = f * q$. For densities of transformations $T(X_1, X_2) = |X_1|^\nu + |X_2|^\nu$ of independent and identically distributed random variables, see Schick and Wefelmeyer (2008b) and (2009b) and the review paper by Müller et al. (2010).

Our estimator is generally not efficient as its influence function

$$I_y(X,Y) = q(y-\varepsilon) - h(y) + f(y - r(X)) - h(y) - \varepsilon\big(f'(y-r(X)) - h'(y)\big)$$

at y does not always belong to the tangent space for our nonparametric regression model. The tangent space consists of functions

$$a(X) + b(\varepsilon) + c(X)\ell(\varepsilon)$$

where the function a satisfies $\int a(x)g(x)\,dx = 0$ and $\int a^2(x)g(x)\,dx < \infty$, the function b satisfies $\int b(y)f(y)\,dy = 0 = \int yb(y)f(y)\,dy$ and $\int b^2(y)f(y)\,dy < \infty$, and the function c satisfies $\int c^2(x)g(x)\,dx < \infty$; see Schick (1993) for details. The projection of the influence function at y into the tangent space is

$$\begin{aligned} I^*_y(X,Y) &= f(y-r(X)) - h(y) + q(y-\varepsilon) - h(y) - \ell(\varepsilon)\big(f'(y-r(X)) - h'(y)\big)/J \\ &= [f(y-r(X)) - h(y)] + [q(y-\varepsilon) - h(y) - d(y)\ell(\varepsilon)] \\ &\qquad + [d(y) - \big(f'(y-r(X)) - h'(y)\big)/J]\ell(\varepsilon). \end{aligned}$$

Here $\ell = -f/f$ denotes the score function for location, J is the Fisher information which needs to be finite for efficiency considerations, and $d(y)$ is the expectation $E[q(y-\varepsilon)\varepsilon]$. Thus $I_y(X,Y) = I^*_y(X,Y)$ holds if and only if $\ell(\varepsilon)/J = \varepsilon$, which in turn holds if and only if f is a mean zero normal density. Consequently, our estimator is efficient for normal errors, but not for other errors.

We expect our result to carry over to estimation of the stationary density of the time series $X_n = r(X_{n-1}) + \varepsilon_n$. This generalization, however, will be nontrivial. In the autoregressive setup we can no longer assume that the support of the variables is compact. This makes the estimation of the function r much more complicated. Some of these difficulties were already encountered by Müller, Schick and Wefelmeyer (2009) when estimating the innovation distribution in such models.

2. The main result

Let V be a continuous function with the following properties:

$$V(0) = 1, \tag{2.1}$$

$$V(x+y) \leq V(x)V(y), \quad x, y \in \mathbb{R}, \tag{2.2}$$

$$V(x) = \sup\nolimits_{|s|\leq 1} V(sx), \quad x \in \mathbb{R}. \tag{2.3}$$

For a measurable function u, we introduce the V-norm defined by

$$\|u\|_V = \int |u(x)|V(x)\,dx.$$

We let L_V denote the collection of (equivalent classes of) measurable functions with finite V-norm. In other words, L_V equals $L_1(\mu)$ where $\mu(dx) = V(x)\,dx$. The following are easy consequences of the property (2.2) of the function V:

$$\|u(\cdot - t)\|_V \le V(t)\|u\|_V, \quad t \in \mathbb{R},\ u \in L_V, \tag{2.4}$$

$$\|u * v\|_V \le \|u\|_V \|v\|_V, \quad u, v \in L_V. \tag{2.5}$$

We say a measurable function u is L_V-*Lipschitz* if

$$\|u(\cdot - t) - u\|_V \le L|t|V(t), \quad t \in \mathbb{R},$$

holds for a constant L.

Let

$$W(x) = \big(1 + \log(1 + |x|)\big)^2 (1 + |x|)V^2(x), \quad x \in \mathbb{R}.$$

This function shares the properties (2.1)–(2.3) with V. An application of the Cauchy–Schwarz inequality and the identity

$$\int \frac{1}{\big(1 + \log(1 + |x|)\big)^2 (1 + |x|)}\,dx = 2$$

yield the inequality

$$\|u\|_V^2 \le 2\|u^2\|_W, \tag{2.6}$$

valid for all measurable u.

We are interested in a generalization of Theorem 1 to the space L_V. Toward this goal we first derive a central limit theorem for the process $n^{1/2}H = n^{1/2}(H_1 + H_2)$ in the space L_V. For this we rely on the central limit theorem in L_1-spaces; see Ledoux and Talagrand (1991, Theorem 10.10) or van der Vaart and Wellner (1996, page 92).

Theorem 2. *Let μ be a σ-finite measure on the Borel-σ-field on $\mathbb{R}$. Let $Z_1, Z_2, \dots$ be independent and identically distributed zero-mean random elements in $L_1(\mu)$. Then the sequence $n^{-1/2}\sum_{i=1}^n Z_i$ converges in distribution in $L_1(\mu)$ to a centered Gaussian process if and only if*

$$\lim_{t\to\infty} t^2 P\Big(\int |Z_1(x)|\,\mu(dx) > t\Big) = 0 \quad \text{and} \quad \int \big(E[Z_1^2(x)]\big)^{1/2}\,\mu(dx) < \infty.$$

We have the following corollary to this theorem.

Corollary 1. *Let $Z_1, Z_2, \dots$ be independent and identically distributed zero-mean random elements in L_V. Then the sequence $n^{-1/2}\sum_{i=1}^n Z_i$ converges in distribution in L_V to a centered Gaussian process if*

$$\|E[Z_1^2]\|_W = \int W(y)E[Z_1^2(y)]\,dy < \infty. \tag{2.7}$$

Proof. We apply the previous theorem with $\mu(dx) = V(x)\,dx$. Using (2.6) we obtain the bounds

$$\|(E[Z_1^2])^{1/2}\|_V^2 \le 2\|E[Z_1^2]\|_W < \infty$$

and

$$E[\|Z_1\|_V^2] \le 2E[\|Z_1^2\|_W] < \infty.$$

The Markov inequality and the Lebesgue dominated convergence then imply

$$\lim_{t\to\infty} t^2 P\big(\|Z_1\|_V > t\big) \le \lim_{t\to\infty} E\big[\|Z_1\|_V^2 \mathbf{1}[\|Z_1\|_V > t]\big] = 0.$$

Thus Theorem 2 yields the desired result. □

We can write $H = (1/n)\sum_{j=1}^n Z_j$, where

$$Z_j(y) = q(y-\varepsilon_j) - h(y) + f(y - r(X_j)) - h(y) - \varepsilon_j(f'(y - r(X_j)) - h'(y)), \quad y \in \mathbb{R}.$$

We have

$$\begin{aligned} E[Z_1^2] &= q^2 * f - h^2 + f^2 * q - h^2 + \sigma^2((f')^2 * q - (h')^2) \\ &\le q^2 * f + f^2 * q + \sigma^2 (f')^2 * q \end{aligned}$$

and thus

$$\|E[Z_1^2]\|_W \le \|q^2\|_W \|f\|_W + \|f^2\|_W \|q\|_W + \sigma^2 \|(f')^2\|_W \|q\|_W.$$

Note that f is bounded by assumption (F) and that q is bounded with compact support by assumptions (G) and (R). Therefore q and q^2 have finite W-norms, and f^2 has finite W-norm if f does. This shows that $\|E[Z_1^2]\|_W$ is finite if f and $(f')^2$ have finite W-norms. Thus we have the following result, with the second statement following from Remark 5 of Schick and Wefelmeyer (2007a).

LEMMA 2. *Suppose (F), (G) and (R) hold and $\|f\|_W$ and $\|(f')^2\|_W$ are finite. Then $\sqrt{n}H$ converges in distribution in the space L_V to a centered Gaussian process. Moreover, $\|n^{1/2}(H * K_b - H)\|_V = o_p(n^{-1/2})$ if K has finite V-norm.*

We are now ready to formulate our main result.

THEOREM 3. *Suppose (F), (G), (R), (W), (C), (K) and (B) hold. Let $\hat{r}$ be the local quadratic smoother defined in (1.1). Let $\|f\|_W$ and $\|(f')^2\|_W$ be finite, and let f' be L_V-Lipschitz. Then the stochastic expansion*

$$\|\hat{h} - h - H\|_V = o_p(n^{-1/2})$$

holds, and $n^{1/2}H$ converges in distribution in the space L_V to a centered Gaussian process.

Note that Theorem 1 is a direct consequence of Theorem 3 and the observation that functions of bounded variation are L_1-Lipschitz. For the latter result see Lemma 8 in Schick and Wefelmeyer (2007a).

PROOF. Set $f_b = k_b * f$ and $q_b = k_b * q$. We have the decomposition

$$\hat{f} * \hat{q} = f_b * q_b + f_b * (\hat{q} - q_b) + q_b * (\hat{f} - f_b) + (\hat{f} - f_b) * (\hat{q} - q_b).$$

Note that

$$f_b * q_b = f * q * k_b * k_b = h * K_b.$$

Let us introduce the notation

$$\bar{\varepsilon} = \frac{1}{n}\sum_{j=1}^n \varepsilon_j$$

and

$$R(y) = \frac{1}{n}\sum_{j=1}^n \big(f(y - \hat{r}(X_j)) - f(y - r(X_j)) + \varepsilon_j f'(y - r(X_j))\big), \quad y \in \mathbb{R},$$

and the kernel estimators

$$\tilde{f}(y) = \frac{1}{n}\sum_{j=1}^n k_b(y - \varepsilon_j), \quad y \in \mathbb{R},$$

and

$$\tilde{q}(z) = \frac{1}{n}\sum_{j=1}^{n} k_b(z - r(X_j)), \quad z \in \mathbb{R}.$$

Then we can write

$$\begin{aligned} q_b * (\hat{f} - f_b) &= q_b * (\tilde{f} - f_b + \bar{\varepsilon} f_b') + q_b * (\hat{f} - \tilde{f} - \bar{\varepsilon} f_b') \\ &= H_1 * K_b + q_b * (\hat{f} - \tilde{f} - \bar{\varepsilon} f_b') \end{aligned}$$

and

$$\begin{aligned} f_b * (\hat{q} - q_b) &= f_b * (\tilde{q} - q_b) + f_b * (\hat{q} - \tilde{q}) \\ &= H_2 * K_b + R * K_b. \end{aligned}$$

The above identities show that

$$\hat{h} - h - H = h * K_b - h + H * K_b - H + R * K_b + q_b * (\hat{f} - \tilde{f} - \bar{\varepsilon} f_b') + (\hat{f} - f_b) * (\hat{q} - q_b).$$

Since q is of bounded variation and quasi-uniform on $[r(0), r(1)]$, we may and do assume that q is of the form

$$q(x) = \int_{u \le x} \phi(u)\,\nu(du), \quad x \in \mathbb{R},$$

where ν is a finite measure with $\nu(\mathbb{R} - [r(0), r(1)]) = 0$, and ϕ is a measurable function such that $|\phi| \le 1$. As shown in Schick and Wefelmeyer (2011), this allows us to write

$$h(y) = \int f(y - x) q(x)\,dx = \int F(y - u)\phi(u)\,\nu(du),$$

where F is the distribution function corresponding to the error density f. The properties of f now yield that h is two times differentiable with bounded derivatives

$$h'(y) = \int f(y - u)\phi(u)\,\nu(du), \quad y \in \mathbb{R}, \tag{2.8}$$

$$h''(y) = \int f'(y - u)\phi(u)\,\nu(du), \quad y \in \mathbb{R}. \tag{2.9}$$

Since f' is L_V-Lipschitz, so is h''; see Lemma 6.1(2) in Schick and Wefelmeyer (2008c). As k is of order three, so is K. Thus it follows from a standard argument that

$$\|h * K_b - h\|_V = O(b^3).$$

In Sections 4–6 we shall verify the following statements:

$$\|\hat{f} - \tilde{f} - \bar{\varepsilon} f_b'\|_V = o_p(n^{-1/2}), \tag{2.10}$$

$$\|R\|_V = o_p(n^{-1/2}), \tag{2.11}$$

$$\|\hat{f} - f_b\|_V = O_p((nb)^{-1/2}), \tag{2.12}$$

$$\|\hat{q} - q_b\|_V = o_p(b^{-1}(nc)^{-1/2}). \tag{2.13}$$

Using these results we derive

$$\begin{aligned} \|\hat{h} - h - H\|_V &\le \|h * K_b - h\|_V + \|H * K_b - H\|_V + \|R\|_V \|K_b\|_V \\ &\quad + \|q_b\|_V \|\hat{f} - \tilde{f} - \bar{\varepsilon} f_b'\|_V + \|\hat{f} - f_b\|_V \|\hat{q} - q_b\|_V \\ &= O(b^3) + o_p(n^{-1/2}) + O_p(n^{-1} b^{-3/2} c^{-1/2}) = o_p(n^{-1/2}) \end{aligned}$$

by the choice of the bandwidths b and c. Here we also used Lemma 2 and the inequalities $\|q_b\|_V \le \|K_b\|_V\|q\|_V$ and $\|K_b\|_V \le \|k\|_V^2$, valid for $b \le 1$ in view of (2.3). □

3. More properties of the local quadratic smoother

In this section we collect additional properties of the local quadratic smoother $\hat{r}$ that will be needed in our proofs. For proofs of these properties we refer the reader to Schick and Wefelmeyer (2011). To simplify notation we set

$$\hat{a} = \hat{r} - r.$$

For a function u defined on $[0,1]$, we write

$$\|u\| = \sup_{0\le x\le 1} |u(x)|.$$

For $x \in [0,1]$, let $U_c(x)$ denote the matrix defined by

$$U_c(x) = \int g(x+cu)\psi(u)\psi^\top(u)w(u)\,du.$$

This matrix is invertible for $c < 1/2$. We write $D_c(x)$ for the first row of its inverse. For later use we mention the bound

$$\sup_{0\le c<1/2}\ \sup_{0\le x\le 1} |D_c(x)| < \infty. \tag{3.1}$$

Let us now set

$$\hat{\Delta}(x) = \frac{1}{nc}\sum_{j=1}^n w\Big(\frac{X_j - x}{c}\Big)\varepsilon_j D_c(x)\psi\Big(\frac{X_j-x}{c}\Big).$$

Then we have the stochastic expansion

$$\|\hat{a} - \hat{\Delta}\| = o_p(n^{-1/2}). \tag{3.2}$$

Finally, we need some results that address the dependence of $\hat{r}$ on the pairs (X_j, ε_j). To describe these results we define, for a subset C of $\{1,\dots,n\}$,

$$\hat{a}_C(x) = \frac{1}{nc}\sum_{j=1}^n w\Big(\frac{X_j-x}{c}\Big)\mathbf{1}[j\notin C]\big(\varepsilon_j + R(X_j,x)\big)D(x)\psi\Big(\frac{X_j-x}{c}\Big)$$

with

$$R(X_j,x) = r(X_j) - r(x) - r'(x)(X_j-x) - \frac{1}{2}r''(x)(X_j-x)^2.$$

We abbreviate $\hat{a}_{\{i\}}$ by $\hat{a}_i$ and $\hat{a}_{\{i,j\}}$ by $\hat{a}_{i,j}$. Then we have

$$\max_{1\le j\le n}\|\hat{a} - \hat{a}_j\| = o_p(1), \tag{3.3}$$

$$\frac{1}{n}\sum_{j=1}^n \big(\hat{a}(X_j) - \hat{a}_j(X_j)\big)^2 = O_p\Big(\frac{\log^2 n}{n^2c^2}\Big), \tag{3.4}$$

$$\frac{1}{n}\sum_{j=1}^n \int \big(\hat{a}(x) - \hat{a}_j(x)\big)^2 g(x)\,dx = O_p\Big(\frac{\log^2 n}{n^2c^2}\Big), \tag{3.5}$$

$$E\big[\big(\hat{a}_1(X_1) - \hat{a}_{1,2}(X_1)\big)^2\big] = O_p\Big(\frac{1}{n^2c}\Big). \tag{3.6}$$

4. Proof of (2.10)

Without loss of generality we assume that $c < 1/2$. Let us again set $\hat{a} = \hat{r} - r$. For a continuous function a, we set

$$\begin{aligned}\mu_{a,b}(z) &= \iint \big(k_b(z-y+a(x)) - k_b(z-y)\big) f(y)g(x)\,dy\,dx \\ &= \int f_b(z+a(x))g(x)\,dx - f_b(z), \quad z \in \mathbb{R}.\end{aligned}$$

A Taylor expansion yields

$$\mu_{\hat{a},b}(z) = f_b'(z)\int \hat{a}(x)g(x)\,dx + \int\int_0^1 (1-s)f_b''(z+s\hat{a}(x))\hat{a}^2(x)\,ds\,g(x)\,dx.$$

As $(f')^2$ has finite W-norm, the inequality (2.6) yields that f' has finite V-norm. Using this and the fact that k and k' are bounded with compact support, we derive the bounds

$$\|f_b'\|_V = \|f' * k_b\|_V \le \|f'\|_V \|k_b\|_V = O(1)$$

and

$$\|f_b''\|_V = \|f' * k_b'\|_V \le \|f'\|_V \|k_b'\|_V = O(b^{-1}).$$

Using these bounds, relations (1.2), (1.4), (1.5), and $b^2nc^2 \to \infty$, we obtain the rate

$$\begin{aligned}\|\mu_{\hat{a},b} - f_b'\bar{\varepsilon}\|_V \le \|f_b'\|_V &\Big|\int \hat{a}(x)g(x)\,dx - \bar{\varepsilon}\Big| \\ &+ \|f_b''\|_V V(\|\hat{a}\|)\int \hat{a}^2(x)g(x)\,dx = o_p(n^{-1/2}).\end{aligned}$$

The desired result (2.10) follows from this if we show

$$\|\hat{f} - \tilde{f} - \mu_{\hat{a},b}\|_V = o_p(n^{-1/2}). \tag{4.1}$$

To this end we introduce

$$\bar{T}(z) = \frac{1}{n}\sum_{j=1}^n T_j(z, \tau(\hat{a})) \quad \text{and} \quad \bar{T}_*(z) = \frac{1}{n}\sum_{j=1}^n T_j(z, \tau(\hat{a}_j)), \quad z \in \mathbb{R},$$

where

$$T_j(z,a) = k_b(z - \varepsilon_j + a(X_j)) - k_b(z - \varepsilon_j) - \mu_{a,b}(z),$$

and where τ is the function defined by

$$\tau(x) = -\mathbf{1}[x < -1] + x\mathbf{1}[|x| \le 1] + \mathbf{1}[x > 1].$$

Note that τ is bounded by 1 and satisfies $|\tau(x) - \tau(y)| \le |x - y|$.

The probability of the event $\{\|\hat{a}\| > 1\}$ tends to zero. On its complement we have the identity $\tau(\hat{a}) = \hat{a}$ and thus $\bar{T} = \hat{f} - \tilde{f} - \mu_{\hat{a},b}$. In view of this, relation (4.1) follows if we show

$$\|\bar{T} - \bar{T}_*\|_V = o_p(n^{-1/2}) \tag{4.2}$$

and

$$\|\bar{T}_*\|_V = o_p(n^{-1/2}). \tag{4.3}$$

For real numbers $x_1, \dots, x_m$ and $y_1, \dots, y_m$, the function χ_b defined by

$$\begin{aligned}\chi_b(z) &= \frac{1}{m}\sum_{i=1}^m \big(k_b(z - x_i - y_i) - k_b(z - x_i)\big)\\ &= -\frac{1}{m}\sum_{i=1}^m \int_0^1 y_i k_b'(z - x_i - s y_i)\, ds, \quad z \in \mathbb{R},\end{aligned}$$

satisfies the inequalities

$$\|\chi_b\|_V \le \|k_b'\|_V \frac{1}{m}\sum_{i=1}^m V(x_i)V(y_i)|y_i| \tag{4.4}$$

and

$$\|\chi_b^2\|_W \le \|(k_b')^2\|_W \frac{1}{m}\sum_{i=1}^m W(x_i)W(y_i)y_i^2. \tag{4.5}$$

Note also that $\|k_b'\|_V = O(b^{-1})$ and $\|(k_b')^2\|_W = O(b^{-3})$. The inequality (4.4) and the statements (1.2), (3.3), (3.4) and (3.5) yield the rate

$$\|\bar{T} - \bar{T}_*\|_V = O_p\Big(\frac{\log n}{bnc}\Big) = o_p(n^{-1/2}).$$

The last step used the fact that $nc^2b^2/\log^2 n = n^{1/2}b^2/\log^2 n \to \infty$. This proves (4.2). To prove (4.3), we write

$$nE[\bar{T}_*^2(z)] = E[T_1^2(z, \tau(\hat{a}_1))] + (n-1)E[T_1(z, \tau(\hat{a}_1))T_2(z, \tau(\hat{a}_2))].$$

Conditioning on $\xi = (\varepsilon_2, X_2, \dots, \varepsilon_n, X_n)$, we see that the first expectation on the right-hand side is bounded by

$$\Xi_{1,b}(z) = E\big[\big(k_b(z - \varepsilon_1 + \tau(\hat{a}_1(X_1))) - k_b(z - \varepsilon_1)\big)^2\big].$$

Moreover, we calculate

$$E[T_1(z, \tau(\hat{a}_1))T_2(z, \tau(\hat{a}_{1,2}))] = E[T_2(z, \tau(\hat{a}_{1,2}))E(T_1(z, \tau(\hat{a}_1))|\xi)] = 0.$$

Similarly one verifies

$$E[T_1(z, \tau(\hat{a}_{1,2}))T_2(z, \tau(\hat{a}_2))] = 0 \quad \text{and} \quad E[T_1(z, \tau(\hat{a}_{1,2}))T_2(z, \tau(\hat{a}_{1,2}))] = 0.$$

Thus the expectation $e(z) = E[T_1(z, \tau(\hat{a}_1))T_2(z, \tau(\hat{a}_2))]$ equals

$$E\big[\big(T_1(z, \tau(\hat{a}_1)) - T_1(z, \tau(\hat{a}_{1,2}))\big)\big(T_2(z, \tau(\hat{a}_2)) - T_2(z, \tau(\hat{a}_{1,2}))\big)\big].$$

An application of the Cauchy–Schwarz inequality shows that $e(z)$ is bounded by $E\big[\big(T_1(z, \tau(\hat{a}_1)) - T_1(z, \tau(\hat{a}_{1,2}))\big)^2\big]$ which in turn is bounded by

$$\Xi_{2,b}(z) = E\big[\big(k_b(z - \varepsilon_1 - \tau(\hat{a}_1(X_1))) - k_b(z - \varepsilon_1 - \tau(\hat{a}_{1,2}(X_1)))\big)^2\big].$$

With the help of (3.6) and (4.5) we obtain the bounds

$$\|\Xi_{2,b}\|_W \le \|(k_b')^2\|_W W(1)E[W(\varepsilon_1)]E[(\hat{a}_1(X_1) - \hat{a}_{1,2}(X_1))^2] = O\Big(\frac{1}{n^2b^3c}\Big)$$

and

$$\|\Xi_{1,b}\|_W \le \|(k_b')^2\|_W E[W(\varepsilon_1)]E[(\hat{a}_1(X_1))^2] = O\Big(\frac{1}{nb^3c}\Big).$$

The above bounds show that $\|(\bar{T}_*)^2\|_W = O_p(n^{-2}b^{-3}c^{-1}) = o_p(n^{-1})$, which implies relation (4.3).

5. Proof of (2.11)

We set

$$J(y) = \frac{1}{n}\sum_{j=1}^{n} \varepsilon_j f'(y - r(X_j)), \quad y \in \mathbb{R},$$

and, for a continuous function a,

$$S_a(y) = \frac{1}{n}\sum_{j=1}^{n} f(y - r(X_j) - a(X_j)), \quad y \in \mathbb{R},$$

$$\nu_a(y) = \int f(y - r(x) - a(x))g(x)\,dx, \quad y \in \mathbb{R},$$

$$\dot\nu_a(y) = \int a(x) f'(y - r(x))g(x)\,dx, \quad y \in \mathbb{R}.$$

Then, with $\hat a = \hat r - r$, we can express R as the sum

$$(S_{\hat a} - S_0 - \nu_{\hat a} + \nu_0) + (\nu_{\hat a} - \nu_0 + \dot\nu_{\hat a}) - (\dot\nu_{\hat a} - J).$$

Therefore the desired result (2.11) is implied by the statements

$$\|S_{\hat a} - S_0 - \nu_{\hat a} + \nu_0\|_V = o_p(n^{-1/2}), \tag{5.1}$$

$$\|\nu_{\hat a} - \nu_0 + \dot\nu_{\hat a}\|_V = o_p(n^{-1/2}), \tag{5.2}$$

$$\|\dot\nu_{\hat a} - J\|_V = o_p(n^{-1/2}). \tag{5.3}$$

The first statement, (5.1), is verified by arguing as in the proof of (4.1); we skip the details. The identity

$$\begin{aligned}
&\nu_{\hat a}(y) - \nu_0(y) + \dot\nu_{\hat a}(y) \\
&\qquad = -\frac{1}{n}\sum_{j=1}^{n} \int \hat a(x) \int_0^1 \big(f'(y - r(x) - s\hat a(x)) - f'(y - r(x))\big)\,ds\,g(x)\,dx,
\end{aligned}$$

the inequality (2.4), and the L_V-Lipschitz property of f' let us bound the left-hand side of (5.2) by a constant times

$$\int V(r(x))V(\hat a(x))\hat a^2(x)g(x)\,dx.$$

The second statement, (5.2), now follows from (1.2) and (1.4). It remains to prove the third statement, (5.3). Relation (3.2) implies

$$\|\dot\nu_{\hat a} - \dot\nu_{\hat\Delta}\|_V = o_p(n^{-1/2}).$$

We can write

$$\begin{aligned}
\dot\nu_{\hat\Delta}(y) &= \frac{1}{n}\sum_{j=1}^{n} \varepsilon_j \int D(x)\psi\Big(\frac{X_j - x}{c}\Big)\frac{1}{c}w\Big(\frac{X_j - x}{c}\Big)f'(y - r(x))g(x)\,dx \\
&= \frac{1}{n}\sum_{j=1}^{n} \varepsilon_j \int D(X_j - cu)\psi(u)w(u)f'(y - r(X_j - cu))g(X_j - cu)\,du \\
&= J(y) + R_1 y) + R_2(y),
\end{aligned}$$

where

$$R_1(y) = \frac{1}{n}\sum_{j=1}^{n} \varepsilon_j f'(y - r(X_j))(t_c(X_j) - 1)$$

and

$$R_2(y) = \frac{1}{n}\sum_{j=1}^{n} \varepsilon_j \int \big(f'(y - r(X_j - cu)) - f'(y - r(X_j))\big)\phi_c(X,u)\,du$$

with

$$\phi_c(x,u) = g(x-cu)D(x-cu)\psi(u)w(u)$$

and

$$t_c(x) = \int \phi_c(x,u)\,du.$$

It was shown in Schick and Wefelmeyer (2011) that

$$E[(t_c(X)-1)^2] \to 0.$$

We have

$$nE[R_1^2(y)] = \sigma^2 E\big[\big(f'(y-r(X))(t_c(X)-1)\big)^2\big]$$

and

$$nE[R_2^2(y)] \le \sigma^2 E\Big[\int \phi_c^2(X,u)\,du \int \big(f'(y-r(X-cu)) - f'(y-r(X))\big)^2\,du\Big].$$

Since $(f')^2$ has finite W-norm, we have

$$m(c) = \sup_{|s|\le c} \int W(y)\big(f'(y-s) - f'(y)\big)^2\,dy \to 0.$$

With

$$B = \sup_{0<c<1/2}\ \sup_{0\le x\le 1}\ \sup_{|u|\le 1} |\phi_c(x,u)| < \infty$$

we find that

$$nE[\|R_1^2\|_W] = n\int W(y)E[R_1^2(y)]\,dy \le \sigma^2\|(f')^2\|_W W(\|r\|)E[(t_c(X)-1)^2] \to 0$$

and

$$nE[\|R_2^2\|_W] = n\int W(y)E[R_2^2(y)]\,dy \le \sigma^2 W(\|r\|)Bm(c) \to 0.$$

This proves the third statement, (5.3), and we are done.

6. Proofs of (2.12) and (2.13)

The statement (2.12) follows from (2.10), the property $\|\bar\varepsilon f'\|_V = O_p(n^{-1/2})$, and the property

$$\|\tilde f - f_b\|_V = O_p((nb)^{-1/2}).$$

The latter follows in turn from the inequality (2.6) and the fact that

$$nbE[\|\tilde f - f_b\|_W^2] \le \int W(x)\int bk_b^2(x-y)f(y)\,dy\,dx \le \|bk_b^2\|_W\|f\|_W = O(1).$$

Similarly, one verifies

$$\|\tilde q - q_b\|_V = O((nb)^{-1/2}).$$

Thus (2.13) follows from

$$\|\hat q - \tilde q\|_V = o_p(b^{-1}(nc)^{-1/2}). \tag{6.1}$$

Let $\hat a = \hat r - r$. In view of the identity

$$\hat q(z) - \tilde q(z) = -\frac{1}{n}\sum_{j=1}^{n}\int_0^1 k_b'\big(z - r(X_j) - s\hat a(X_j)\big)\hat a(X_j)\,ds$$

we have

$$\|\hat{q} - \tilde{q}\|_V \le \|k_b'\|_V V(M_n) \frac{1}{n} \sum_{j=1}^{n} |\hat{a}(X_j)|$$

with

$$M_n = \max_{1 \le j \le n} \big(|r(X_j)| + |\hat{a}(X_j)| \big) = O_p(1).$$

We have $\|k_b'\|_V = O(b^{-1})$, and the desired (6.1) follows from (1.3).

References

[1] Escanciano, J. C. and Jacho-Chávez, D. T. (2011). $\sqrt{n}$-uniformly consistent density estimation in nonparametric regression models. To appear in: *J. Econometrics*.
[2] Frees, E. W. (1994). Estimating densities of functions of observations. *J. Amer. Statist. Assoc.* **89**, 517–525.
[3] Giné, E. and Mason, D. (2007). On local U-statistic processes and the estimation of densities of functions of several sample variables. *Ann. Statist.* **35**, 1105–1145.
[4] Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces. Isoperimetry and Processes.* Ergebnisse der Mathematik und ihrer Grenzgebiete (3) 23, Springer, Berlin.
[5] Müller, U. U. (2010). Estimating the density of a possibly missing response variable in nonlinear regression. Submitted.
[6] Müller, U.U., Schick, A. and Wefelmeyer, W. (2009). Estimating the innovation distribution in nonparametric autoregression. *Probab. Theory Related Fields* **144** 53–77.
[7] Müller, U. U., Schick, A. and Wefelmeyer, W. (2010). Non-standard behavior of density estimators for functions of independent observations. Submitted.
[8] Nickl, R. (2007). Donsker-type theorems for nonparametric maximum likelihood estimators. *Probab. Theory Related Fields* **138**, 411–449. Erratum **141** (2008), 331-332.
[9] Nickl, R. (2009). On convergence and convolutions of random signed measures. *J. Theoret. Probab.* **22**, 38–56.
[10] Saavedra, A. and Cao, R. (1999). Rate of convergence of a convolution-type estimator of the marginal density of an MA(1) process. *Stochastic Process. Appl.* **80**, 129–155.
[11] Saavedra, A. and Cao, R. (2000). On the estimation of the marginal density of a moving average process. *Canad. J. Statist.* **28**, 799–815.
[12] Schick, A. (1993). On efficient estimation in regression models. *Ann. Statist.* **21**, 1486–1521. Correction and Addendum **23** (1995), 1862–1863.
[13] Schick, A. and Wefelmeyer, W. (2004a). Root n consistent and optimal density estimators for moving average processes. *Scand. J. Statist.* **31**, 63–78.
[14] Schick, A. and Wefelmeyer, W. (2004b). Root n consistent density estimators for sums of independent random variables. *J. Nonparametr. Statist.* **16**, 925–935.
[15] Schick, A. and Wefelmeyer, W. (2004c). Functional convergence and optimality of plug-in estimators for stationary densities of moving average processes. *Bernoulli* **10**, 889–917.
[16] Schick, A. and Wefelmeyer, W. (2007a). Root-n consistent density estimators of convolutions in weighted L_1-norms. *J. Statist. Plann. Inference* **37**, 1765–1774.
[17] Schick, A. and Wefelmeyer, W. (2007b). Uniformly root-n consistent density estimators for weakly dependent invertible linear processes. *Ann. Statist.* **35**, 815–843.
[18] Schick, A. and Wefelmeyer, W. (2008a). Root-n consistency in weighted L_1-spaces for density estimators of invertible linear processes. *Stat. Inference Stoch. Process.* **11**, 281–310.
[19] Schick, A. and Wefelmeyer, W. (2008b). Non-standard behavior of density estimators for sums of squared observations. *Statist. Decisions* **27**, 55–73.
[20] Schick, A. and Wefelmeyer, W. (2008c). Convergence rates in weighted L_1 spaces of kernel density estimators for linear processes. *ALEA* **4**, 117–129.
[21] Schick, A. and Wefelmeyer, W. (2009a). Improved density estimators for invertible linear processes. In: *Recent Advances in Theory and Applications of Statistics. Festschrift for Shelemyahu Zacks* (N. Mukhopadhyay, ed.), *Comm. Statist. Theory Methods* **38**, 3123–3147.
[22] Schick, A. and Wefelmeyer, W. (2009b). Convergence rates of density estimators for sums of powers of observations. *Metrika* **69** 249–264.

[23] Schick, A. and Wefelmeyer, W. (2011). Uniform convergence of convolution estimators for the response density in nonparametric regression. Submitted.

[24] Støve, B. and Tjøstheim, D. (2010). A convolution estimator for the density of nonlinear regression observations. Submitted.

[25] van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes. With Applications to Statistics.* Springer Series in Statistics, Springer, New York.

Anton Schick, Department of Mathematical Sciences, Binghamton University, Binghamton, NY 13902-6000, USA

Wolfgang Wefelmeyer, Mathematical Institute, University of Cologne, Weyertal 86–90, 50931 Köln, Germany