PROBABILITY
AND
MATHEMATICAL STATISTICS
Vol. 24, Fasc. 2 (2004), pp. 213–235

ADAPTIVE KERNEL ESTIMATION OF THE MODE IN A NONPARAMETRIC RANDOM DESIGN REGRESSION MODEL

BY

KLAUS ZIEGLER (ILMENAU)

Abstract. In a nonparametric regression model with random design, where the regression function m is given by $m(x) = E(Y \mid X = x)$, estimation of the location θ (*mode*) and size $m(\theta)$ of a unique maximum of m is considered. As estimators, location $\hat{\theta}$ and size $\hat{m}(\hat{\theta})$ of a maximum of the Nadaraya–Watson kernel estimator $\hat{m}$ for the curve m are chosen. Within this setting, we establish joint asymptotic normality and asymptotic independence for $\hat{\theta}$ and $\hat{m}(\hat{\theta})$ (which can be exploited for constructing simultaneous confidence intervals for θ and $m(\theta)$) under mild local smoothness assumptions on m and the design density g (imposed in a neighborhood of θ). The bandwidths employed for $\hat{m}$ are data-dependent and of plug-in type. This is handled by viewing the estimators as stochastic processes indexed by a so-called scaling parameter and proving functional central limit theorems for those processes. In the same way, we obtain, as a by-product, an asymptotic normality result for the Nadaraya–Watson estimator itself at a finite number of distinct points, which improves on previous results.

AMS Subject Classification: 62G05, 62G07.

Key words and phrases: Nonparametric regression, random design, mode, kernel smoothing, Nadaraya–Watson estimator, weak convergence, functional central limit theorems.

1. INTRODUCTION

Let (X, Y) be a 2-dimensional random vector with unknown bivariate distribution such that Y is integrable. Then the dependence of Y on the (random) value x of X can be expressed by the *regression function*

$$m(x) := E(Y \mid X = x), \qquad x \in \boldsymbol{R}.$$

In this situation, we speak of the *random design regression model* and call X the *design variable* and Y the *response variable.* We will always assume X to have a marginal (Lebesgue) density g which we will refer to as the *design density.* (We do not require (X, Y) to have a bivariate Lebesgue density, however.) Our

main concern is the estimation of the location θ and size $m(\theta)$ of a unique maximum (*mode, peak*) of the (unknown) function m. Our method is indirect in the sense that the estimators for θ and $m(\theta)$ are based on a kernel estimator $\hat{m}(x)$ for the regression curve $m(x)$ (for details see Section 2 below). With $\hat{m}$ given, it suggests itself to estimate θ and $m(\theta)$ from location $\hat{\theta}$ and size $\hat{m}(\hat{\theta})$ of a global maximum of $\hat{m}$ (over some compact interval while θ is assumed to be an interior point of a slightly bigger compact interval; details will be given again in Section 2).

Within this setting, our main aim is to establish joint asymptotic normality for the estimators $\hat{\theta}$ and $\hat{m}(\hat{\theta})$ under some mild local smoothness conditions on the regression function m and the design density g (mostly imposed locally in a neighborhood of θ). Corollary 3.10 below will, in addition, reveal the asymptotic independence of $\hat{\theta}$ and $\hat{m}(\hat{\theta})$, which makes the construction of simultaneous confidence intervals for θ and $m(\theta)$ especially simple. Our results will be valid for a wide class of kernels (not necessarily having compact support) and data-dependent bandwidths of the type being generated by plug-in rules.

The above-described idea of estimating location and size of a maximum of a nonparametric curve by the corresponding functionals of a kernel estimator of the curve is not new; it stems from the closely related problem of estimating the mode of a density. In continuation of Parzen's (1962) pioneering work on density estimation and estimation of the mode, Eddy (1980), (1982) and Romano (1988a) tackled optimality questions of kernel density estimators of the mode. Romano (1988a) also seems to have been the first to consider data-dependent bandwidths in this context. In another paper (Romano (1988b)) he examined the limiting behavior of bootstrap estimators of the location of the mode, an idea picked up later by Grund and Hall (1995) in the context of bandwidth selection by minimizing the bootstrapped L_p-error for the mode estimator. See also Gasser et al. (1998) for the recent concept of estimating the mode when the data are points in a normed space. For bootstrapping in the context of the present paper, see Ziegler (2000), (2001a).

While kernel methods have quickly won recognition also in nonparametric regression models after the fundamental works of Nadaraya (1964), Watson (1964) and Gasser and Müller (1979), they have not been widely used for estimating location and size of maxima of regression functions so far. Up to now, we are only aware of the work of Müller (1985), (1988), (1989) and Ehm (1996) where this kind of "nonparametric peak estimation" is pursued. Müller's (1989) work on peak estimation culminated in a functional central limit theorem for the joint distribution of the estimated location and size of the peak. This has already led to a joint asymptotic normality and independence result for bandwidths estimating the theoretically optimal ones (which would depend on unknown quantities) by plug-in methods. However, Müller's results only apply to models with fixed design. For random design models, peak estimation based on the Nadaraya–Watson (NW) kernel estimator appears to be an ap-

propriate method, but we do not know of any results pertaining to this in the literature except those of Boularan et al. (1995) on the related concepts of estimation of zeros (the points where m attains the value 0) and nonparametric calibration in the random design model. Since the results obtained by Boularan et al. (1995) also apply to derivatives of regression functions, the estimation of maxima is contained through estimation of the zeros of the first derivative of m. However, for the validity of their approach, global existence and continuity of the first derivatives of m and g are needed, while our corresponding result gets along with continuity of m at the point θ and requires no continuity of g at all. Furthermore, Boularan et al. (1995) deal only with consistency problems and leave asymptotic normality questions for the estimator of the location of the zero out of consideration.

In all of the above-mentioned papers, unimodality, i.e. the existence of a "unique largest peak" of the density, is an overall assumption. So it is natural to ask for methods for testing the hypothesis of multimodality. This was addressed by Silverman (1981) and further examined e.g. by Mammen et al. (1992). This is closely related to the concept of 'bump hunting' where the number of modes of the curve is estimated and tests are based on this. This topic has been treated in Heckman (1992) and Harezlak and Heckman (2001), where also additional references can be found.

The outline of the present paper is as follows: Section 2 (framework) presents the precise definitions of the estimators and briefly surveys some consistency and asymptotic normality results which have been proved in other papers of the author (Ziegler (2002), (2003)). Some optimality considerations for a bandwidth choice are also given. Section 3 (functional central limit theorems and data-dependent bandwidths) then contains our main results including the announced joint asymptotic normality result for $\hat{\theta}$ and $\hat{m}(\hat{\theta})$ (Corollary 3.10). Some facts about weak convergence of stochastic processes in the sense of Hoffmann-Jørgensen needed in the proofs of Section 3 are collected in the Appendix.

2. FRAMEWORK

Let $I := [a, b]$ be a compact interval and $J := [a-\varrho, b+\varrho]$, $\varrho > 0$, a slight enlargement of I which is introduced in order to avoid boundary effects. We impose the following (very mild) overall conditions on the regression function m and the design density g which include the characterization of θ as the 'unique largest peak' of m on J:

m is bounded on J;

$\exists \theta \in I \ \forall \varepsilon > 0$: $m(\theta) > \sup_{x \in J: |x-\theta| > \varepsilon} m(x)$;

g is bounded away from zero and infinity on J.

A further discussion of these conditions is contained in Ziegler (2002). Now, in case of existence, we introduce $\hat{\theta} = \hat{\theta}_{n,h}$ as an estimator for θ through the equation

$$\hat{m}_{n,h}(\hat{\theta}_{n,h}) := \sup_{x \in I} \hat{m}_{n,h}(x), \tag{1}$$

with $\hat{m}_{n,h}$ being the Nadaraya–Watson estimator for m defined by

$$\hat{m}(x) \equiv \hat{m}_{n,h} := \frac{\sum_{i=1}^{n} Y_i K\big((x-X_i)/h\big)}{\sum_{i=1}^{n} K\big((x-X_i)/h\big)},$$

where K is a *kernel*, $h > 0$ a *bandwidth*, and (X_i, Y_i) are observations being i.i.d. copies of (X, Y). Note that $\hat{\theta}_{n,h}$ exists if K is continuous; however, it may not be unique (in fact, it is known that kernel estimators tend to produce some additional and superfluous modality). On the other hand, it is indifferent which method for choosing $\hat{\theta}_{n,h}$ from its competitors is pursued; our results apply to any choice of $\hat{\theta}_{n,h}$ satisfying (1). We emphasize that the validity of our proofs is not affected by potential non-measurability of $\hat{\theta}_{n,h}$, either.

Under appropriate regularity conditions (on the kernel, the bandwidth etc.), consistency of $\hat{\theta}$ and $\hat{m}(\hat{\theta})$ has been proved in Ziegler (2002), Theorem 2.3. Under some additional regularity conditions (including three-fold continuous differentiability of m and g in a neighborhood of θ) and for a fixed (i.e. non-data-driven) bandwidth with $nh^7 \to d \geqslant 0$, also asymptotic normality for $\hat{\theta}$ has been obtained in Ziegler (2003), Theorem 3.8; the precise statement is

$$\sqrt{nh^3}\,(\hat{\theta}-\theta) \overset{\mathscr{L}}{\to} N(\mu, \sigma^2), \tag{2}$$

where

$$\mu := -\frac{d}{2}\left[\frac{m^{(3)}(\theta)}{m^{(2)}(\theta)} + 2\frac{g^{(1)}(\theta)}{g(\theta)}\right]\int_{\mathbf{R}} z^2 K(z)\,dz \tag{3}$$

and

$$\sigma^2 := \frac{\mathrm{Var}(Y \mid X = \theta)}{g(\theta)\big(m^{(2)}(\theta)\big)^2}\int_{\mathbf{R}} [K^{(1)}(z)]^2\,dz. \tag{4}$$

In particular, the bias and variance of $\hat{\theta}$ asymptotically behave like

$$\mathrm{bias}(\hat{\theta}) \sim h^2\mu \quad \text{and} \quad \mathrm{Var}(\hat{\theta}) \sim \frac{1}{nh^3}\sigma^2.$$

Hence for the asymptotic mean-square error of $\hat{\theta}$ we obtain

$$\mathrm{MSE}(\hat{\theta}) = E(\hat{\theta}-\theta)^2 \sim \frac{1}{nh^3}\sigma^2 + h^4\mu^2. \tag{5}$$

(Grund and Hall (1995) have pointed out, in the context of estimating the mode of a density, that (5) rather describes the MSE of the asymptotic distri-

bution than the actual asymptotic mean-square error. Indeed, in order to make (5) rigorous, a uniform integrability argument would have to be added to the mere application of (2).)

Minimizing the expression (5) with respect to h yields

$$h_{\text{opt}} = s_{\text{opt}} \cdot n^{-1/7} \tag{6}$$

with

$$s_{\text{opt}} = (3\sigma^2/4\mu^2)^{1/7} \tag{7}$$

as an optimal choice for h in the sense of classical L_2-theory. (Note that there is no optimal choice at all in this sense if the asymptotic bias μ vanishes). The optimal rate of $n^{-1/7}$ (being valid for nonnegative kernels) is already well known from mode and regression peak estimation (Eddy (1980); Romano (1988a); Müller (1989)). It is also well known that the optimal constant always depends on unknowns in situations like that under consideration. However, s_{opt}, as a function of μ and σ, can be estimated consistently by the plug-in method (see Remark (i) below). It is a natural question, therefore, to ask whether the asymptotic normality result (2) is still valid when s_{opt} in (6) is replaced by a consistent estimator $\hat{s}_{\text{opt}}$, i.e. if

$$\hat{h}_{\text{opt}} = \hat{s}_{\text{opt}} \cdot n^{-1/7}$$

is chosen as a bandwidth. But observe that this is of the general form

$$h_n = S_n \cdot v_n \tag{8}$$

with v_n being a sequence of real numbers tending to zero and S_n being a statistic depending on the observations. Hence we seek for asymptotic normality results for bandwidths of the form (8). (Another motivation for considering bandwidths of this form is the so-called scale-equivariance of kernel estimators, a notion introduced and discussed at length in Romano (1988a).)

Remarks on bandwidth selection. (i) Within the present setting, the plug-in method consists in replacing $g^{(j)}(\theta)$, $j = 0, 1$, and $m^{(j)}(\theta)$, $j = 2, 3$, by $\hat{g}^{(j)}(\hat{\theta})$, $j = 0, 1$ (with $\hat{g}$ being the Rosenblatt–Parzen kernel estimator for g), and $\hat{m}^{(j)}(\hat{\theta})$, $j = 2, 3$, respectively, and $\operatorname{Var}(Y|X = \theta)$ by

$$\hat{\operatorname{Var}}(Y|X = \theta) := \frac{\sum_{i=1}^{n} (Y_i - \hat{m}(\hat{\theta}))^2 K((\hat{\theta} - X_i)/h)}{\sum_{i=1}^{n} K((\hat{\theta} - X_i)/h)} \tag{9}$$

in the expressions of μ and σ^2 (in (7)). This procedure is justified by Theorems 1.5 and 2.2 in Ziegler (2002), which ensures the consistency of these pilot estimators (under some regularity conditions); however, some care has to be taken in the choice of the 'pilot bandwidths' (see Remark 3.9 (iii) in Ziegler (2003) for details).

(ii) In the present context, bandwidth selection by the plug-in method requires the practical computation of particularly many pilot estimators. Hence, a bootstrap approach to bandwidth choice (as already known from density estimation, see Shao and Tu (1995), p. 353, and the references given there) should be considered. In Ziegler (2001a) the so-called 'smoothed paired bootstrap' (SPB), where the observation pairs (X_i, Y_i) are bootstrapped, has been proposed. We cannot go into detail here, but we mention that, in the setting of the SPB, the bootstrapped MSE (which should form a reasonable approximation for the original MSE) can be minimized directly with respect to the bandwidth, leading to a bandwidth depending on the (X_i, Y_i), but not on the bootstrap variables. Furthermore, under suitable conditions it might be shown that the bandwidth h_n obtained by this minimization algorithm can be regarded as being of the form $h_n = S_n \cdot n^{-1/7}$ with $S_n \xrightarrow{P} s_{\text{opt}}$, so that our results apply to a bandwidth chosen by this procedure. (See Ziegler (2001b), where this is proved in the case of estimation of the density at a point.)

Some sets of conditions. At the end of this section, we collect some regularity conditions (on the kernel, bandwidth etc.) of which we will make a repeated use in the sequel. On the moments of Y and the fixed sequence $v_n \to 0$ the bandwidth is based on, we impose the following:

$$E|Y|^{2+p} < \infty \quad \text{and} \quad n^p v_n^{2+p} \to \infty \quad \text{for some } p > 0, \tag{10}$$

$$nv_n^6 \to \infty, \quad nv_n^7 \to d \ (0 \leqslant d < \infty). \tag{11}$$

Further we will say that the *local smoothness condition* (S) *is satisfied* if m and g are three times continuously differentiable in some neighborhood of θ with $m^{(2)}(\theta) < 0$ and $\operatorname{Var}(Y|X = \cdot)$ is continuous at θ.

As to assumptions on the kernel, they will vary from result to result, but by *assumption* (K) we will mean from now on that K is twice continuously differentiable with K, $K^{(1)}$ and $K^{(2)}$ being of bounded variation and with

$$\lim_{|z|\to\infty} |z^3 K(z)| = 0, \quad \int_R z^2 K(z)\,dz < \infty,$$

$$\lim_{|z|\to\infty} |z^4 K^{(1)}(z)| = 0, \quad \int_R z^3 |K^{(1)}(z)|\,dz < \infty, \quad \lim_{|z|\to\infty} |z^3 K^{(2)}(z)| = 0.$$

Finally, we will speak of *condition* (Kk) with $k \geqslant 2$ if K is of bounded variation and if there is some $k \geqslant 2$ such that

$$\lim_{|z|\to\infty} |z^{k+1} K(z)| = 0,$$

$$\lim_{a\to\infty} a^{k-j} \int_{|z|\leqslant a} z^j K(z)\,dz = 0 \quad \text{for each } j = 1, \ldots, k-1$$

and

$$\int_{\boldsymbol{R}} z^k |K(z)|\, dz < \infty,$$

and we will speak of *condition* (K′) if K has a bounded and integrable first derivative $K^{(1)}$ for which $|z^2 K^{(1)}(z)|$ is bounded in $\boldsymbol{R}$.

3. FUNCTIONAL CENTRAL LIMIT THEOREMS AND DATA-DEPENDENT BANDWIDTHS

To be able to prove asymptotic normality results for data-driven bandwidths, we view $\hat{m}$ and $\hat{\theta}$ as processes in the sense

$$\hat{m}_{n,s} \equiv \hat{m}_{n,v_n s},\ s \in [s_1, s_2], \quad \text{and} \quad \hat{\theta}_{n,s} \equiv \hat{\theta}_{n,v_n s},\ s \in [s_1, s_2]$$

(where $0 < s_1 < s_2 < \infty$ should be chosen in such a manner that $s_1 < s_{\text{opt}} < s_2$) and show that

$$M_n(s) := \sqrt{n v_n s}\,(\hat{m}_{n,s}(x) - m(x)), \quad s \in [s_1, s_2], \tag{12}$$

and

$$\Theta_n(s) := \sqrt{n v_n^3 s^3}\,(\hat{\theta}_{n,s} - \theta), \quad s \in [s_1, s_2], \tag{13}$$

converge weakly to appropriately chosen Gaussian processes.

Note that the process M_n can be regarded as a random element in $C[s_1, s_2]$, so that Billingsley's (1968) classical theory of weak convergence in the function space $C[s_1, s_2]$ applies, while the mapping $s \mapsto \hat{\theta}_{n,s}$ need not be continuous, so that a technical difficulty arises in proving a functional central limit theorem for $\hat{\theta}_{n,s}$ (see, however, Müller (1989), where similar mappings are viewed as random elements in a certain function space). We will overcome this difficulty by employing weak convergence in the sense of Hoffmann-Jørgensen (see the Appendix). But before doing this, we give first a functional version of Proposition 3.1 in Ziegler (2003) on which the functional central limit theorems for both the processes M_n and Θ_n will be based. The 'tightness' argument in Step 2 in the proof of Proposition 3.1 below is standard and e.g. similar to the proof of Lemma 3.1 in Müller and Stadtmüller (1987).

PROPOSITION 3.1. *Let assumption* (10) *be fulfilled. Assume that K is bounded and integrable with* $\lim_{|z| \to \infty} |zK(z)| = 0$ *and satisfies assumption* (K′). *Let x be such that m, g, and* $\operatorname{Var}(Y \mid X = \cdot)$ *are continuous at x. For fixed* $0 < s_1 < s_2 < \infty$ *consider the stochastic process given by*

$$V_{n,i}(s) \equiv V_{n,i}(s, x, K) := \frac{1}{\sqrt{n v_n s}} (Y_i - m(x))\, K\left(\frac{x - X_i}{v_n s}\right), \quad s \in [s_1, s_2]. \tag{14}$$

Then

$$W_n := \sum_{i=1}^{n} (V_{n,i} - EV_{n,i}) \overset{\mathscr{L}}{\Rightarrow} \tilde{\tilde{G}} \quad \text{in } C[s_1, s_2], \tag{15}$$

where $\tilde{\tilde{G}}$ *is a mean zero Gaussian process with paths in* $C[s_1, s_2]$ *and with covariances*

$$\operatorname{cov}(\tilde{\tilde{G}}(s), \tilde{\tilde{G}}(t)) = \operatorname{Var}(Y|X = x)\, g(x) \frac{1}{\sqrt{st}} \int_{\mathbf{R}} K\left(\frac{z}{s}\right) K\left(\frac{z}{t}\right) dz, \quad s, t \in [s_1, s_2].$$

Proof. As a first step, it is shown that for given points $t_1, \ldots, t_r \in [s_1, s_2]$, the random vector $(W_n(t_1), \ldots, W_n(t_r))$ is asymptotically normal with (mean zero and) asymptotic covariance matrix $(\sigma_{ij})_{i,j=1,\ldots,n}$, where

$$\sigma_{ij} := \operatorname{Var}(Y|X = x)\, g(x) \frac{1}{\sqrt{t_i t_j}} \int_{\mathbf{R}} K\left(\frac{z}{t_i}\right) K\left(\frac{z}{t_j}\right) dz.$$

This can be done via the Cramér–Wold device, using Proposition 3.1 in Ziegler (2003) employing (real numbers $\lambda_1, \ldots, \lambda_l$ being given) the 'auxiliary kernel'

$$\tilde{K}(z) := \sum_{l=1}^{r} \frac{\lambda_l}{\sqrt{t_l}} K\left(\frac{z}{t_l}\right).$$

Then, according to Theorem 8.1, p. 54, and Theorem 12.3, p. 95 (with $\gamma = \alpha = 2$ and $F(t) = \sqrt{At}$), in Billingsley (1968), in order to complete the proof of Proposition 3.1 it is enough to show that there is some $A > 0$ such that for every n and $s, t \in [s_1, s_2]$ it follows that

$$E\left((W_n(s) - W_n(t))^2\right) \leqslant A(s-t)^2. \tag{16}$$

Now, by the mean value theorem, we obtain

$$E\left((W_n(s) - W_n(t))^2\right) \tag{17}$$
$$\leqslant (s-t)^2 \cdot (t^*)^{-3} \cdot (v_n t^*)^{-1} E\left((Y - m(x))^2 K^*\left(\frac{x-X}{v_n t^*}\right)\right)$$

with $K^*(z) = (K(z) + zK^{(1)}(z))^2$ and some t^* between s and t.

Now, since $(t^*)^{-3} \leqslant (s_1)^{-3}$ and $v_n t^* \leqslant v_n s_2 \to 0$, an application of Theorem 1A in Parzen (1962) shows that the expression

$$(t^*)^{-3} \cdot (v_n t^*)^{-1} E\left((Y - m(x))^2 K^*\left(\frac{x-X}{v_n t^*}\right)\right)$$

in (17) has an upper bound in n which is independent of s and t, whence (16) is established. ∎

The following lemma will be needed in the proofs of our main results and is of independent interest since it leads to the asymptotic normality of the Nadaraya–Watson estimator itself with data-dependent bandwidths (Corollary 3.4 below).

LEMMA 3.2. *Assume that K (not necessarily greater than or equal to 0) is bounded and integrable with $\int_{\mathbf{R}} K(z)\,dz = 1$ and satisfies the conditions* (Kk) *(for some $k \geqslant 2$) and* (K′). *Let condition* (10) *be fulfilled and assume that, for some fixed $x \in \mathbf{R}$ with $g(x) > 0$, the function* $\mathrm{Var}(Y \mid X = \cdot)$ *is continuous at the point x. Assume further that m and g have continuous k-th derivatives in a neighborhood of x. Then the stochastic process M_n introduced in* (12) *can be written as*

$$M_n(s) = \xi_n(s) + \sqrt{n v_n^{2k+1}}\, s^{(2k+1)/2}\, \mu_{n,s}, \qquad s \in [s_1, s_2],$$

with

$$\xi_n \xrightarrow{\mathcal{L}} \tilde{G} \quad \text{in } C[s_1, s_2], \tag{18}$$

where $\tilde{G}$ is a mean zero Gaussian process with paths in $C[s_1, s_2]$ and covariances

$$\mathrm{cov}\big(\tilde{G}(s), \tilde{G}(t)\big) = \frac{\mathrm{Var}(Y|X=x)}{g(x)} \frac{1}{\sqrt{st}} \int_{\mathbf{R}} K\left(\frac{z}{s}\right) K\left(\frac{z}{t}\right) dt, \qquad s, t \in [s_1, s_2], \tag{19}$$

and with

$$\sup_{s \in [s_1, s_2]} |\mu_{n,s} - \mu(x)| \xrightarrow{P} 0,$$

where

$$\mu(x) := \frac{(-1)^k}{k!} \left[m^{(k)}(x) + \sum_{j=1}^{k-1} \binom{k}{j} \frac{m^{(k-j)}(x)\, g^{(j)}(x)}{g(x)} \right] \int_{\mathbf{R}} z^k K(z)\,dz. \tag{20}$$

Proof. Observe that

$$M_n(s) = \frac{1}{\hat{g}_{n,s}(x)} \sum_{i=1}^{n} V_{n,i}(s, x, K)$$

with

$$\hat{g}_{n,s}(x) := \frac{1}{n v_n s} \sum_{i=1}^{n} K\left(\frac{x - X_i}{v_n s}\right) \tag{21}$$

and $V_{n,i}$ being defined as in (14). First we show that

$$\xi_n(s) := \frac{1}{\hat{g}_{n,s}(x)} \sum_{i=1}^{n} \big(V_{n,i}(s, x, K) - E V_{n,i}(s, x, K)\big), \qquad s \in [s_1, s_2],$$

behaves as indicated in (18). Note that $\xi_n(s) = W_n(s)/\hat{g}_{n,s}(x)$ with $W_n(s)$ being defined as in (15). By Proposition 3.1, in order to establish (18) it suffices to

show that ξ_n and $W_n/g(x)$ are stochastically equivalent in $C[s_1, s_2]$ (see Lemma A3 in the Appendix). Now

$$\frac{W_n(s)}{\hat{g}_{n,s}(x)} - \frac{W_n(s)}{g(x)} = W_n(s)\frac{g(x)-\hat{g}_{n,s}(x)}{\hat{g}_{n,s}(x)\cdot g(x)},$$

where the right-hand side tends to 0 in probability uniformly in $s\in[s_1, s_2]$ by Proposition 3.1 (whence W_n is stochastically bounded in $C[s_1, s_2]$ by Lemma A2) and Theorem 1.5 (i) in Ziegler (2002) with $j = 0$. Consequently, (18) is proved.

Now, in order to conclude the proof of Lemma 3.2, we have only to show that

$$\text{(22)}\quad \mu_{n,s} := \big(\hat{g}_{n,s}(x)\big)^{-1}\frac{1}{\sqrt{n(v_n s)^{2k+1}}}\sum_{i=1}^{n} EV_{n,i}(s) \xrightarrow{P} \mu(x) \quad \text{in } C[s_1, s_2].$$

Again by Theorem 1.5 (i) in Ziegler (2002) it suffices to show that

$$\text{(23)}\quad \sup_{s\in[s_1,s_2]}\left|\frac{1}{\sqrt{n(v_n s)^{2k+1}}}\sum_{i=1}^{n} EV_{n,i}(s) - \mu(x)g(x)\right| \to 0.$$

But note that

$$\frac{1}{\sqrt{n(v_n s)^{2k+1}}}\sum_{i=1}^{n} EV_{n,i}(s) = \frac{1}{(v_n s)^{k+1}} E\big(Y-m(x)\big)K\left(\frac{x-X}{v_n s}\right),$$

while it can be readily inferred from Proposition 3.2 in Ziegler (2003) that

$$\frac{1}{h^{k+1}} E\left(\big(Y-m(x)\big)K\left(\frac{x-X}{h}\right)\right) \to \mu(x)g(x) \quad \text{as } h\to 0,$$

whence (23) is established. ∎

THEOREM 3.3. *If, in addition to the assumptions of Lemma* 3.2, *we have* $p > 1/k$ *and*

$$\text{(24)}\quad nv_n^{2k+1} \to c^2,$$

then for the process M_n *introduced in* (12) *it follows that*

$$M_n \xrightarrow{\mathcal{L}} G \quad \text{in } C[s_1, s_2],$$

where G *is a Gaussian process taking its paths in* $C[s_1, s_2]$ *with means*

$$E\big(G(s)\big) = c\cdot s^{(2k+1)/2}\cdot\mu(x), \quad s\in[s_1, s_2],$$

with $\mu(x)$ *being defined as in* (20) *and the same covariance structure as* $\tilde{G}$ (see (19)).

Proof. By (22) and (24),

$$\sup_{s\in[s_1,s_2]} |\sqrt{nv_n^{2k+1}}\cdot s^{(2k+1)/2}\cdot \mu_{n,s} - c\cdot s^{(2k+1)/2}\cdot \mu(x)| \xrightarrow{P} 0.$$

Hence Theorem 3.3 follows from Lemma 3.2 by virtue of a simple application of Slutsky in $C[s_1, s_2]$ (Lemma A4). ∎

The subsequent corollary is our first asymptotic normality result for data-dependent band-widths.

COROLLARY 3.4 (Asymptotic normality of the NW estimator taken at a single point – data-dependent bandwidths). *If h_n fulfills* (8) *with $S_n \xrightarrow{P} s_0 > 0$, then under the assumptions of Theorem* 3.3 *(put $\hat{m}_n(x) := \hat{m}_{n,h_n}(x)$)*

$$\sqrt{nh_n}\big(\hat{m}_n(x) - m(x)\big) \xrightarrow{\mathcal{L}} N\big(c\cdot s_0^{(2k+1)/2}\cdot \mu(x),\ \sigma^2(x)\big)$$

with $\mu(x)$ as in (20) *and*

$$\sigma^2(x) := \frac{\mathrm{Var}(Y|X=x)\int_R K^2(z)\,dz}{g(x)}. \tag{25}$$

Proof. Since

$$\sqrt{nh_n}\big(\hat{m}_n(x) - m(x)\big) = M_n(S_n) = M_n(s_0) + M_n(S_n) - M_n(s_0),$$

with $M_n(s_0)$ having a limit distribution as desired (Theorem 3.3), we have only to show that

$$M_n(S_n) - M_n(s_0) \xrightarrow{P} 0.$$

But this is standard employing the asymptotic equicontinuity of M_n, which in turn is a consequence of Theorem 3.3 (see Theorem A5). ∎

Remark 3.5. For estimating m at a fixed point x, an asymptotically optimal bandwidth is

$$h_{\mathrm{opt}} = t_{\mathrm{opt}}\cdot n^{-1/(2k+1)}, \quad \text{where } t_{\mathrm{opt}} = \left(\frac{\sigma^2(x)}{2k\mu^2(x)}\right)^{1/(2k+1)}$$

with $\sigma^2(x)$ and $\mu(x)$ as in (25) and (20). This can be obtained from Corollary 3.4 by the same reasoning as described at the beginning of the chapter. Furthermore, t_{opt} can be estimated consistently if, in particular, $EY^4 < \infty$ and $\mathrm{Var}(Y|X=\cdot)$ is continuous in a neighborhood of x (compare Remark 3.9 (iii) in Ziegler (2003)). For a *globally* optimal choice of h, see Nadaraya (1989), Theorem 1.5, p. 121, and the remark thereafter.

Our next step towards the main result (which will consist in Theorem 3.9 together with Corollary 3.10) is a functional central limit theorem for the peak estimator $\hat{\theta}_{n,s}$ scaled by $s\in[s_1, s_2]$.

THEOREM 3.6. *Let the assumptions* (10), (11), *and* (S) *be fulfilled, and let* K *be a symmetric probability density satisfying* (K). *Then for the process* Θ_n *introduced in* (13) *it follows that*

$$\Theta_n \xrightarrow{\mathscr{L}} H \quad \text{in } l^\infty[s_1, s_2]$$

in the sense of Hoffmann-Jørgensen (see Definition A1), *where* H *is a Gaussian process taking its paths in* $C[s_1, s_2]$ *and with means*

$$EH(s) = s^{7/2} \cdot \mu$$

(μ *as in* (3)) *and covariances* ($s, t \in [s_1, s_2]$)

$$\operatorname{cov}(H(s), H(t)) = \frac{\operatorname{Var}(Y \mid X = \theta)}{g(\theta)(m^{(2)}(\theta))^2} \frac{1}{\sqrt{st}} \int_{\mathbf{R}} K^{(1)}\left(\frac{z}{s}\right) K^{(1)}\left(\frac{z}{t}\right) dz.$$

Proof. By the mean value theorem, we have for each $s \in [s_1, s_2]$

$$\Theta_n(s) = -\frac{\sqrt{nv_n^3 s^3}\, \hat{m}_{n,s}^{(1)}(\theta)}{\hat{m}_{n,s}^{(2)}(\theta_s^*)}, \tag{26}$$

with some θ_s^* lying between $\hat{\theta}_{n,s}$ and θ. For the numerator, we utilize the identity (suppressing indices)

$$\hat{m}^{(1)}(\theta) = \mathrm{I} - \mathrm{II}, \tag{27}$$

where

$$\mathrm{I} := \frac{\hat{r}^{(1)}(\theta) - m(\theta)\hat{g}^{(1)}(\theta)}{\hat{g}(\theta)} \tag{28}$$

and

$$\mathrm{II} := \frac{\hat{g}^{(1)}(\theta)}{\hat{g}(\theta)}(\hat{m}(\theta) - m(\theta)) \tag{29}$$

with $\hat{g}$ defined as in (21) and

$$\hat{r}(x) = \hat{r}_{n,s}(x) = \frac{1}{nv_n s} \sum_{i=1}^{n} Y_i K\left(\frac{x - X_i}{v_n s}\right). \tag{30}$$

As to II, it follows from Lemma 3.2 ($k = 2$) together with Theorem 1.5 from Ziegler (2002) ($j = 0, 1$) that

$$\sup_{s \in [s_1, s_2]} \left| \sqrt{nv_n^3 s^3}\, \mathrm{II} - \frac{d}{2} \cdot s^{7/2} \cdot \frac{g^{(1)}(\theta)}{g(\theta)} m^{(2)}(\theta) \int_{\mathbf{R}} z^2 K(z)\, dz \right| \xrightarrow{P} 0. \tag{31}$$

As to I, an application of Proposition 3.1 with $K^{(1)}$ instead of K (note that $K^{(1)}$ fulfills the assumptions needed there for K) together with two Slutsky arguments in $C[s_1, s_2]$ (Lemmas A3 and A4) like in the proofs of Lemma 3.2 and Theorem 3.3, respectively (with the second one employing Proposition 3.2 in

Ziegler (2003) with $k = 3$ and observing that $m^{(1)}(\theta) = 0$, $K^{(1)}(-z) = -K^{(1)}(z)$ and $\frac{1}{3}\int z^3 K^{(1)}(z)\,dz = -\int K(z) z^2\,dz$) yield

$$(\sqrt{nv_n^3 s^3}\, I)_{s\in[s_1,s_2]} \xrightarrow{\mathscr{L}} \tilde{H} \quad \text{in } C[s_1, s_2], \tag{32}$$

where

$$E\tilde{H}(s) = \frac{d\cdot\mu'}{g(\theta)}\cdot s^{7/2}$$

with

$$\mu' := \frac{1}{2}[m^{(3)}(\theta) g(\theta) + 3m^{(2)}(\theta) g^{(1)}(\theta)] \int_{\mathbf{R}} z^2 K(z)\,dz$$

and

$$\operatorname{cov}(\tilde{H}(s), \tilde{H}(t)) = \frac{\operatorname{Var}(Y \mid X = \theta)}{g(\theta)} \frac{1}{\sqrt{st}} \int_{\mathbf{R}} K^{(1)}\left(\frac{z}{s}\right) K^{(1)}\left(\frac{z}{t}\right) dz.$$

Hence, by (27), (31) and (32),

$$\sqrt{nv_n^3 s^3}\, \hat{m}_{n,s}^{(1)}(\theta) \xrightarrow{\mathscr{L}} -m^{(2)}(\theta) H \quad \text{in } C[s_1, s_2]. \tag{33}$$

Now according to Theorem 2.3 in Ziegler (2002) we obtain

$$\sup_{s\in[s_1,s_2]} |\theta_s^* - \theta| \xrightarrow{P} 0,$$

and hence, by Theorem 1.5 (ii) in Ziegler (2002),

$$\sup_{s\in[s_1,s_2]} |\hat{m}_{n,s}^{(2)}(\theta_s^*) - m^{(2)}(\theta)| \to 0.$$

Then, recalling (26), a Slutsky argument (Lemma A3) for weak convergence in the space $l^\infty[s_1, s_2]$ (again being analogous to the first step in the proof of Lemma 3.2) shows that the processes $\Theta_n(s)$ and

$$-\frac{\sqrt{nv_n^3 s^3}\, \hat{m}_{n,s}^{(1)}(\theta)}{m^{(2)}(\theta)}$$

are stochastically equivalent, and this together with (33) yields the assertion. ∎

Remark 3.7. The proof of Theorem 3.6 shows that the weak convergence result (33) for the denominator in (26) holds without assuming $K \geqslant 0$ and with requiring the second derivative $K^{(2)}$ of the kernel only to be integrable with $|z^2 K^{(2)}(z)|$ being bounded.

The following corollary for data-dependent bandwidths is derived from Theorem 3.6 in a completely analogous manner as Corollary 3.4 has been inferred from Theorem 3.3.

COROLLARY 3.8 (Asymptotic normality of the estimator of the location of the peak – data-dependent bandwidths). *If h_n fulfills* (8) *with* $S_n \xrightarrow{P} s_0 > 0$, *then under the assumptions of Theorem* 3.5 (*put* $\hat{\theta}_n := \hat{\theta}_{n,h_n}$))

$$\sqrt{nh_n^3}(\hat{\theta}_n - \theta) \xrightarrow{\mathscr{L}} N(s_0^{7/2} \cdot \mu, \sigma^2),$$

with μ *and* σ^2 *defined as in* (3) *and* (4), *respectively.* ∎

In order to get an idea of the necessary amount of bias correction (in constructing a confidence interval for θ), put $s_0 = s_{\text{opt}}$ (from (7)) in Corollary 3.8. Observe that, in this case, the asymptotic bias becomes

$$s_{\text{opt}}^{7/2} \cdot \mu = \pm \frac{\sqrt{3}}{2} \sigma,$$

depending on the sign of μ, i.e. on the sign of

$$-\left[\frac{m^{(3)}(\theta)}{m^{(2)}(\theta)} + 2\frac{g^{(1)}(\theta)}{g(\theta)}\right].$$

To compare with the estimation of m at a point x, put $c = 1$ and $s_0 = t_{\text{opt}}$ (from Remark 3.5) in the situation of Corollary 3.4, leading to an asymptotic bias of $\pm\sigma(x)/\sqrt{2k}$ (depending on the sign of $\mu(x)$). This shows that the bias correction needed for peak estimation is much more extensive than that being sufficient for estimation of the regression function itself, and that the latter bias correction is getting smaller with increasing order of the kernel being used.

Now, in the context of functional limit theorems and data-dependent bandwidths, we tackle the problem of asymptotic normality of the estimated sizes of the peaks. In Ziegler (2003) it has been explained why it is necessary to consider kernels of order $k > 2$ (which take negative values!) here in order to achieve a nondegenerate limit distribution. Furthermore, if we take ordinary, i.e. second-order kernels ($k = 2$, $K \geqslant 0$) for estimating the location of the peak, i.e. for deriving $\hat{\theta}_n$, and plug this estimator into a curve estimator $\hat{m}_{2,n}$ using a third-order kernel (whence $\hat{m}_{2,n}$ is in particular different from the regression curve estimator $\hat{m}_{1,n}$ from which $\hat{\theta}_n$ had been derived), $h_n \sim n^{-1/7}$ will be an optimal rate for both $\hat{\theta}_n$ and $\hat{m}_{2,n}(\hat{\theta}_n)$.

To be more precise, let

$$\hat{m}_{l,n,s}(x) := \frac{\sum_{i=1}^n Y_i K_l\big((x - X_i)/v_n s\big)}{\sum_{i=1}^n K_l\big((x - X_i)/v_n s\big)}, \qquad l = 1, 2,$$

with some second-order kernel (i.e. a symmetric probability density) K_1 satisfying the assumption (K), and some (third-order) kernel K_2 satisfying (K3) (i.e. (Kk) with $k = 3$) and sufficiently many additional conditions to ensure (33), see Remark 3.7.

Now choose $\hat{\theta}_{1,n,s} \equiv \hat{\theta}_{n,s}$ such that

$$\hat{m}_{1,n,s}(\hat{\theta}_{1,n,s}) = \max_{x \in I} \hat{m}_{1,n,s}(x)$$

and $0 < s_1 < s_{\text{opt}} < s_2 < \infty$, $0 < t_1 < t_{\text{opt}} < t_2 < \infty$ (for $k = 3$; see Remark 3.5).

Within this setting, we have the following joint functional central limit theorem for the estimators of the location and the size of the peak. The limit is a 2-dimensional Gaussian process. Note, in particular, that the limiting means and covariances of the estimator $\hat{m}_{2,n,t}(\hat{\theta}_{n,s})$ of the size of the peak do not depend on the scaling parameter s of the estimator for the location of the peak. Note further that the estimators for the location and the size of the peak are asymptotically uncorrelated. Weak convergence of 2-dimensional processes in the space $l^\infty(S) \times l^\infty(T)$ (with S and T being metric spaces) will be treated by Corollary A6 in the Appendix.

THEOREM 3.9. *Let, in addition to the above, the assumptions* (10), (11) *and* (S) *be fulfilled. Define the stochastic processes* $G_{1,n}$ *and* $G_{2,n}$ *by*

$$G_{1,n}(s) := \sqrt{n v_n^3 s^3}\,(\hat{\theta}_{n,s} - \theta), \qquad s \in [s_1, s_2],$$

and

$$G_{2,n}(s, t) := \sqrt{n v_n t}\,(\hat{m}_{2,n,t}(\hat{\theta}_{n,s}) - m(\theta)), \qquad (s, t) \in [s_1, s_2] \times [t_1, t_2].$$

Then

$$\begin{pmatrix} G_{1,n} \\ G_{2,n} \end{pmatrix} \xrightarrow{\mathscr{L}} \begin{pmatrix} G_1 \\ G_2 \end{pmatrix} \quad \text{in } l^\infty[s_1, s_2] \times l^\infty([s_1, s_2] \times [t_1, t_2]),$$

where (G_1, G_2) *is a 2-dimensional Gaussian process with paths in* $C[s_1, s_2] \times C([s_1, s_2] \times [t_1, t_2])$ *and with means*

$$EG_1(s) = -\frac{d \cdot s^{7/2}}{2}\left[\frac{m^{(3)}(\theta)}{m^{(2)}(\theta)} + 2\frac{g^{(1)}(\theta)}{g(\theta)}\right]\int_{\mathbf{R}} z^2 K_1(z)\,dz, \qquad s \in [s_1, s_2],$$

$$EG_2(s, t) = -\frac{d \cdot t^{7/2}}{6}\left[m^{(3)}(\theta) + 3\frac{m^{(2)}(\theta) g^{(1)}(\theta)}{g(\theta)}\right]\int_{\mathbf{R}} z^3 K_2(z)\,dz,$$
$$(s, t) \in [s_1, s_2] \times [t_1, t_2],$$

and covariances $(u_1, u_2 \in [s_1, s_2]$, $v, v_1, v_2 \in [t_1, t_2])$

$$\operatorname{cov}(G_1(u_1), G_1(u_2)) = \frac{\operatorname{Var}(Y \mid X = \theta)}{g(\theta)\left(m^{(2)}(\theta)\right)^2}\frac{1}{\sqrt{u_1 u_2}}\int_{\mathbf{R}} K_1^{(1)}\left(\frac{z}{u_1}\right) K_1^{(1)}\left(\frac{z}{u_2}\right) dz,$$

$$\operatorname{cov}(G_1(u_1), G_2(u_2, v)) = 0,$$

$$\operatorname{cov}(G_2(u_1, v_1), G_2(u_2, v_2)) = \frac{\operatorname{Var}(Y \mid X = \theta)}{g(\theta)}\frac{1}{\sqrt{v_1 v_2}}\int_{\mathbf{R}} K_2\left(\frac{z}{v_1}\right) K_2\left(\frac{z}{v_2}\right) dz.$$

Proof. By Corollary A6 it suffices to show that

$$(34)\quad \big(G_{1,n}(u_1), \ldots, G_{1,n}(u_r), G_{2,n}(w_1, v_1), \ldots, G_{2,n}(w_p, v_p)\big) \xrightarrow{\mathcal{L}} \big(G_1(u_1), \ldots, G_1(u_r), G_2(w_1, v_1), \ldots, G_2(w_p, v_p)\big)$$

for any choice of $u_1, \ldots, u_r \in [s_1, s_2]$, $(w_1, v_1), \ldots, (w_p, v_p) \in [s_1, s_2] \times [t_1, t_2]$ and that (we denote, in contrast to the notation of the Appendix, outer probability again by P)

$$(35)\quad \limsup_{n\to\infty} P\Big(\sup_{|u_1-u_2|\leq\delta} |G_{1,n}(u_1)-G_{1,n}(u_2)| > \varepsilon\Big) \to 0 \quad \text{as } \delta \to 0$$

and

$$(36)\quad \limsup_{n\to\infty} P\Big(\sup_{\substack{|u_1-u_2|\leq\delta \\ |v_1-v_2|\leq\delta}} |G_{2,n}(u_1, v_1)-G_{2,n}(u_2, v_2)| > \varepsilon\Big) \to 0 \quad \text{as } \delta \to 0.$$

Observe, by Theorem A5, that (35) is an immediate consequence of Theorem 3.6.

Next we show (36). Now, with

$$(37)\quad M_{2,n}(t) := \sqrt{nv_n}\, t\big(\hat{m}_{2,n,t}(\theta) - m(\theta)\big)$$

it follows that

$$(38)\quad \begin{aligned} G_{n,2}(u_1, v_1) - G_{n,2}(u_2, v_2) &= M_{2,n}(v_1) - M_{2,n}(v_2) + \sqrt{nv_n v_1}\big(\hat{m}_{2,n,v_1}(\hat{\theta}_{n,u_1}) - \hat{m}_{2,n,v_1}(\theta)\big) \\ &\quad - \sqrt{nv_n v_2}\big(\hat{m}_{2,n,v_2}(\hat{\theta}_{n,u_2}) - \hat{m}_{2,n,v_2}(\theta)\big). \end{aligned}$$

By Theorem 3.3 (and, again, Theorem A5) we obtain

$$\limsup_{n\to\infty} P\Big(\sup_{|v_1-v_2|\leq\delta} |M_{2,n}(v_1)-M_{2,n}(v_2)| > \varepsilon\Big) \to 0 \quad \text{as } \delta \to 0,$$

so that in order to prove (36) it suffices to show that

$$(39)\quad \sup_{\substack{s_1\leq s\leq s_2 \\ t_1\leq t\leq t_2}} \Big|\sqrt{nv_n}\, t\big(\hat{m}_{2,n,t}(\hat{\theta}_{n,s}) - \hat{m}_{2,n,t}(\theta)\big)\Big| \xrightarrow{P} 0.$$

Now, suppressing the dependence on n in the sequel, we get

$$\begin{aligned} \hat{m}_{2,t}(\hat{\theta}_s) - \hat{m}_{2,t}(\theta) &= \hat{m}^{(1)}_{2,t}(\theta^*_s)(\hat{\theta}_s - \theta) \\ &= \hat{m}^{(1)}_{2,t}(\theta)(\hat{\theta}_s - \theta) + \big(\hat{m}^{(1)}_{2,t}(\theta^*_s) - \hat{m}^{(1)}_{2,t}(\theta)\big)(\hat{\theta}_s - \theta) \\ &= \hat{m}^{(1)}_{2,t}(\theta)(\hat{\theta}_s - \theta) + \hat{m}^{(2)}_{2,t}(\theta^{**}_s)(\theta^*_s - \theta)(\hat{\theta}_s - \theta) \end{aligned}$$

with θ_s^* between $\hat{\theta}_s$ and θ and with θ_s^{**} between θ_s^* and θ, and hence

$$(40)\quad \sqrt{nv_n}\left(\hat{m}_{2,t}(\hat{\theta}_s)-\hat{m}_{2,t}(\theta)\right)=\sqrt{nv_n^3}\,\hat{m}_{2,t}^{(1)}(\theta)\frac{1}{\sqrt{nv_n^5}}\sqrt{nv_n^3}\,(\hat{\theta}_s-\theta)$$
$$+\hat{m}_{2,t}^{(2)}(\theta_s^{**})\frac{1}{\sqrt{nv_n^5}}\sqrt{nv_n^3}\,(\theta_s^*-\theta)\sqrt{nv_n^3}\,(\hat{\theta}_s-\theta).$$

Now, by Theorem 3.6 and Lemma A2 (and by the fact that $s^{-3/2}$ is bounded for $s\in[s_1, s_2]$), $\sqrt{nv_n^3}\,(\hat{\theta}_s-\theta)$ is stochastically bounded in $l^\infty[s_1, s_2]$, and hence $\sqrt{nv_n^3}\,(\theta_s^*-\theta)$ is also stochastically bounded, since $\sup_s|\theta_s^*-\theta|\leqslant\sup_s|\hat{\theta}_s-\theta|$. By Remark 3.7, $\sqrt{nv_n^3\,t^3}\,\hat{m}_{2,t}^{(1)}(\theta)$ is weakly convergent in $l^\infty[t_1, t_2]$ (note that K_2 fulfills all the assumptions needed for this), whence, by Lemma A2, $\sqrt{nv_n^3}\,\hat{m}_{2,t}^{(1)}(\theta)$ is stochastically bounded in $l^\infty[t_1, t_2]$ since $t^{-1/2}$ is bounded for $t\in[t_1, t_2]$. Furthermore, by Theorems 1.5 and 2.3 in Ziegler (2002) (recall that $nv_n^6\to\infty$) it follows that

$$\sup_{\substack{s_1\leqslant s\leqslant s_2\\ t_1\leqslant t\leqslant t_2}}|\hat{m}_{2,t}^{(2)}(\theta_s^{**})-m^{(2)}(\theta)|\xrightarrow{\mathbf{P}}0.$$

Considering all this and $nv_n^5\to\infty$ in (40), we can infer that (39), and hence (36) is proved.

The proof of (34) utilizes the Cramér–Wold device. Consider (for fixed $u_l, w_q\in[s_1, s_2]$, $v_q\in[t_1, t_2]$ and real numbers λ_l, μ_q) the expression

$$\sum_{l=1}^{r}\lambda_l\,G_{n,1}(u_l)+\sum_{q=1}^{p}\mu_q\,G_{n,2}(w_q, v_q),$$

which is, by (38) and (39), stochastically equivalent to

$$\sum_{l=1}^{r}\lambda_l\,G_{n,1}(u_l)+\sum_{q=1}^{p}\mu_q\,M_{2,n}(v_q).$$

This expression is, in turn, stochastically equivalent (see (27), (28), (31) and, again, Theorems 1.5 and 2.3 in Ziegler (2002)) to the following one, where the quantities $\hat{g}_l$, $\hat{r}_l$ are defined in an analogous fashion as in (21), (30), the index l referring to the kernel K_l:

$$\frac{1}{g(\theta)}\left(\sum_{l=1}^{r}\lambda_l\left[-\frac{\sqrt{nv_n^3\,u_l^3}}{m^{(2)}(\theta)}\left(\hat{r}_{1,n,u_l}^{(1)}(\theta)-m(\theta)\,\hat{g}_{1,n,u_l}^{(1)}(\theta)\right)+\frac{d}{2}u_l^{7/2}\,g^{(1)}(\theta)\int_{\boldsymbol{R}}z^2\,K_1(z)\,dz\right]\right.$$
$$\left.+\sum_{q=1}^{p}\mu_q\sqrt{nv_n\,v_q}\left(\hat{r}_{2,n,v_q}(\theta)-m(\theta)\,\hat{g}_{2,n,v_q}(\theta)\right)\right)$$
$$=\frac{1}{g(\theta)}\left(-\frac{1}{\sqrt{nv_n}}\sum_{i=1}^{n}\left(Y_i-m(\theta)\right)\tilde{K}\left(\frac{\theta-X_i}{v_n}\right)+\frac{d}{2}\sum_{l=1}^{r}\lambda_l u_l^{7/2}\,g^{(1)}(\theta)\int_{\boldsymbol{R}}z^2\,K_1(z)\,dz\right)$$

with

$$\tilde{K}(z) = \frac{1}{m^{(2)}(\theta)} \sum_{l=1}^{r} \frac{\lambda_l}{\sqrt{u_l}} K_1^{(1)}\left(\frac{z}{u_l}\right) - \sum_{q=1}^{p} \frac{\mu_q}{\sqrt{v_q}} K_2\left(\frac{z}{v_q}\right).$$

Now the proof of (34) can be concluded by Propositions 3.1 and 3.2 in Ziegler (2003) and taking into account that $\int_R z^3 K_1^{(1)}(z)\,dz = -3\int_R z^2 K_1(z)\,dz$ and

$$\int_R K_1^{(1)}\left(\frac{z}{u}\right) K_2\left(\frac{z}{v}\right) dz = 0 \quad \text{for any } u, v > 0$$

(note that $K_1^{(1)}$ is odd, while K_2 is even). ∎

COROLLARY 3.10 (Joint asymptotic normality of the estimators for location and size of the peak – data-dependent bandwidths). *Let* $S_n \xrightarrow{P} s_0$, $T_n \xrightarrow{P} t_0$ *and* $h_{1,n} := S_n \cdot v_n$, $h_{2,n} := T_n \cdot v_n$. *Then under the assumptions of Theorem* 3.9, *with the notation* $\hat{\theta}_n := \hat{\theta}_{1,n,S_n}$ *and* $\hat{m}_{2,n} := \hat{m}_{2,n,T_n}$, *it follows that*

$$\begin{pmatrix} (nh_{1,n}^3)^{1/2}(\hat{\theta}_n - \theta) \\ (nh_{2,n})^{1/2}(\hat{m}_{2,n}(\hat{\theta}_n) - m(\theta)) \end{pmatrix} \xrightarrow{\mathcal{L}} N\left(\begin{pmatrix} \mu_1(s_0) \\ \mu_2(t_0) \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}\right)$$

with

$$\mu_1(s_0) = EG_1(s_0), \qquad \mu_2(t_0) = EG_2(s_0, t_0)$$

(see Theorem 3.9; note again that $EG_2(s_0, t_0)$ does not depend on s_0) *and*

$$\sigma_1^2 = \frac{\operatorname{Var}(Y|X=\theta)}{g(\theta)\left(m^{(2)}(\theta)\right)^2} \int_R [K_1^{(1)}(z)]^2\,dz, \qquad \sigma_2^2 = \frac{\operatorname{Var}(Y|X=\theta)}{g(\theta)} \int_R [K_2(z)]^2\,dz.$$

Remarks 3.11. (i) For the Gasser–Müller estimator in the fixed design model, Müller (1985), (1989) has established results similar to our Theorem 3.9 and Corollary 3.10. He employs compactly supported kernels and asymptotically equidistant (nonrandom) design points and imposes a smoothness assumption on the regression function globally on some compact interval. Since the Gasser–Müller estimator is well-behaved also with respect to its higher derivatives, Müller's results are also valid for the higher derivatives of this estimator. However, his approach, in contrast to ours, consists in taking different growth rates of bandwidth for location and size of the peak, but using the same (higher order) kernel for both.

(ii) For the estimator of the mode of a density based on the kernel density estimator, Romano (1988a) has proved results similar to our Theorem 3.6 and Corollary 3.8, but not a result corresponding to Theorem 3.9 which, however, may also be established in the context of estimating the mode of a density using the same methods as presented here.

In Ziegler (2003) it has been mentioned that Theorem 3.6 *loc. cit.* generalizes to the case of distinct points (instead of a single point x). With the methods

that have just been presented, it is possible to prove even a functional central limit theorem for estimating the regression function at distinct points $x_1, \ldots, x_r$, whose corollary can be used for constructing simultaneous confidence intervals for $m(x_1), \ldots, m(x_r)$ (with data-dependent bandwidths). For details see Ziegler (2000), Theorem 4.13 and Corollary 4.14. We will only state the corollary here which improves on Theorem 4.2.1 in Härdle (1990) and Corollary 1.4, Chapter 4, in Nadaraya (1989):

THEOREM 3.12 (Joint asymptotic normality of the NW estimator at a finite number of distinct points – data-dependent bandwidths). *Under the assumptions of Lemma* 3.2 *and* (24), *and with* m *and* g *having* k*-th continuous derivatives in appropriate neighborhoods of* $x_1, \ldots, x_r$, $\mathrm{Var}(Y \mid X = \cdot)$ *being continuous at* $x_1, \ldots, x_r$ *and* $g(x_i) > 0$ *for all* $x_1, \ldots, x_r$, *we have with* $h_{i,n} = S_{i,n} \cdot v_n$, $\hat{m}_{i,n} := \hat{m}_{n,S_{i,n}}$ *and* $S_{i,n} \xrightarrow{P} s_{i,0}$ *for certain* $s_{i,0} \in [s_1^{(i)}, s_2^{(i)}]$

$$\begin{pmatrix} (nh_{1,n})^{1/2}(\hat{m}_{1,n}(x_1) - m(x_1)) \\ \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\ (nh_{r,n})^{1/2}(\hat{m}_{r,n}(x_r) - m(x_r)) \end{pmatrix}$$

$$\xrightarrow{\mathcal{L}} N\left(\begin{pmatrix} c \cdot s_{1,0}^{(2k+1)/2} \cdot \mu(x_1) \\ \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\ c \cdot s_{r,0}^{(2k+1)/2} \cdot \mu(x_r) \end{pmatrix}, \begin{pmatrix} \sigma^2(x_1) & & 0 \\ & \ddots & \\ 0 & & \sigma^2(x_r) \end{pmatrix} \right)$$

with $\mu(\cdot)$ *and* $\sigma^2(\cdot)$ *defined as in* (20) *and* (25). ∎

APPENDIX: WEAK CONVERGENCE OF STOCHASTIC PROCESSES

In order to avoid measurability problems, Hoffmann-Jørgensen (1984) introduced a notion of weak convergence in metric spaces where measurability is required for the limit only. Here we compile the definition and some useful properties. We skip the proofs since they can be found in the literature, see e.g. van der Vaart and Wellner (1996) or Gaenssler and Rost (1999). Since the only metric spaces occurring in the present work are the function spaces C and l^∞ and finite products thereof, we confine ourselves to the case of normed vector spaces.

A1. DEFINITION. Let $(\Omega, \mathfrak{A}, P)$ be a p-space and $(V, \|\cdot\|)$ a normed vector space. Consider mappings $X_n: \Omega \to V$ (not necessarily measurable) and let $X_0: \Omega \to V$ be $\mathfrak{A}$-$\mathfrak{B}(V)$-measurable (where $\mathfrak{B}(V)$ denotes the Borel σ-field on $(V, \|\cdot\|)$). Then we define *weak convergence* in V as follows:

$$X_n \xrightarrow{\mathcal{L}} X_0 \text{ (in } V) :\Leftrightarrow \lim_{n\to\infty} E^*(f \circ X_n) = E(f \circ X_0) \ \forall f \in C^b(V),$$

where $C^b(V)$ is the space of all bounded and continuous functions on $(V, \|\cdot\|)$ and E^* denotes outer expectation. ∎

Note that Definition A1 is an extension of classical weak convergence in function spaces such as C and D (see Billingsley (1968)) in the sense that if the X_n can be regarded as random elements in one of those spaces, then our definition coincides with the classical one; moreover, most of the properties of the latter carry over to weak convergence in the sense of Definition A1. One of them is the so-called Portmanteau theorem which, in particular, implies the following lemma:

A2. LEMMA. *If $X_n \overset{\mathcal{L}}{\Rightarrow} X_0$ in V (in the sense of Definition* A1*), then X_n is (P^*-)* stochastically bounded *in $(V, \|\cdot\|)$, i.e. for each $\varepsilon > 0$ there is some $M > 0$ such that*

$$P^*(\|X_n\| > M) < \varepsilon \quad \text{for } n \text{ large enough,}$$

where P^* denotes outer probability. ∎

(Cramér)-Slutsky-type results carry over as well:

A3. LEMMA. *Let $X_n \overset{\mathcal{L}}{\Rightarrow} X_0$ in V and assume that the sequence of mappings X_n and Y_n: $\Omega \to V$ are* stochastically equivalent *in $(V, \|\cdot\|)$, i.e. for each $\varepsilon > 0$*

$$\lim_{n\to\infty} P^*(\|X_n - Y_n\| > \varepsilon) = 0,$$

then

$$Y_n \overset{\mathcal{L}}{\Rightarrow} X_0 \text{ in } V,$$

i.e. Y_n is also weakly convergent with the same weak limit as X_n. ∎

The following lemma is often called also (Cramér)-Slutsky:

A4. LEMMA. *Assume that $X_n \overset{\mathcal{L}}{\Rightarrow} X_0$ in V and $Y_n \overset{\mathcal{L}}{\Rightarrow} a \in V$ (i.e. that Y_n has a nonrandom weak limit, which is, by the way, equivalent to $\|Y_n - a\| \overset{P^*}{\to} 0$), then $X_n + Y_n \overset{\mathcal{L}}{\Rightarrow} X_0 + a$.* ∎

The next theorem is a characterization of weak convergence for stochastic processes indexed by some totally bounded pseudometric parameter space (T, d) and with bounded sample paths. In this situation, we can take $V \subset l^\infty(T)$, where $l^\infty(T)$ is the space of all bounded functions on T endowed with the sup norm. The limiting process will have paths in the separable subspace

$$U^b(T) \equiv U^b(T, d) := \{x \in l^\infty(T): x \text{ uniformly } d\text{-continuous}\}.$$

In the applications in Section 3 of the present paper, T is always a compact interval so that T is totally bounded and $U^b(T)$ coincides with $C(T)$.

A5. THEOREM. *Let (T, d) be totally bounded and let $(X_n(t))_{t\in T}$ $(n\in N)$ be stochastic processes having paths in $V \subset l^\infty(T)$. Assume that there is a stochastic process $(\bar{X}_0(t))_{t\in T}$ such that for any choice of $t_1, \ldots, t_p \in T$*

$$(X_n(t_1), \ldots, X_n(t_p)) \xrightarrow{\mathcal{L}} (\bar{X}_0(t_1), \ldots, \bar{X}_0(t_p))$$

('fidi'-convergence) and assume that for each $\varepsilon > 0$

$$\limsup_{n\to\infty} P^*\Big(\sup_{\substack{s,t\in T\\ d(s,t)\leqslant\delta}} |X_n(s)-X_n(t)| > \varepsilon\Big) \to 0 \quad \text{as } \delta\to 0 \tag{41}$$

('asymptotic equicontinuity', AEC). Then there is a stochastic process $(X_0(t))_{t\in T}$ with paths in $U^b(T)$ such that

$$X_n \xrightarrow{\mathcal{L}} X_0 \text{ in } V, \tag{42}$$

where X_0 has the same 'fidi'-convergence as $\bar{X}_0$.

Conversely, (42) with X_0 having paths in $U^b(T)$ implies (41). ∎

In the present paper, the convergence of 2-dimensional stochastic processes also occurs. The characterization Theorem A5 essentially carries over to that case (see also Problem 2, p. 42, in van der Vaart and Wellner (1996)):

A6. COROLLARY. *Let (T_i, d_i), $i = 1, \ldots, r$, be totally bounded pseudometric parameter spaces and let $V_i \subset l^\infty(T_i)$. Let*

$$(X_{1,n}, \ldots, X_{r,n}) \equiv \big((X_{1,n}(t))_{t\in T_1}, \ldots, (X_{r,n}(t))_{t\in T_r}\big) \quad (n\in N)$$

and $(\bar{X}_{1,0}, \ldots, \bar{X}_{r,0})$ be r-dimensional stochastic processes such that, for any choice of $t_{i,1}, \ldots, t_{i,p_i} \in T_i$, $i = 1, \ldots, r$,

$$(X_{i,n}(t_{i,l}))_{l=1,\ldots,p_i, i=1,\ldots,r} \xrightarrow{\mathcal{L}} (\bar{X}_{i,0}(t_{i,l}))_{l=1,\ldots,p_i, i=1,\ldots,r}.$$

Assume further that

$$\limsup_{n\to\infty} P^*\Big(\sup_{\substack{s,t\in T_i\\ d_i(s,t)\leqslant\delta}} |X_{i,n}(s)-X_{i,n}(t)| > \varepsilon\Big) \to 0 \quad \text{as } \delta\to 0,\ i = 1, \ldots, r. \tag{43}$$

Then there is an r-dimensional stochastic process $(X_{1,0}, \ldots, X_{r,0})$ with paths in $U^b(T_1, d_1)\times\ldots\times U^b(T_r, d_r)$ and having the same 'fidi'-convergence as $(\bar{X}_{1,0}, \ldots, \bar{X}_{r,0})$ in the sense that, for any choice of $t_{i,1}, \ldots, t_{i,p_i} \in T_i$, $i = 1, \ldots, r$,

$$(X_{i,0}(t_{i,l}))_{l=1,\ldots,p_i, i=1,\ldots,r} \xrightarrow{\mathcal{L}} (\bar{X}_{i,0}(t_{i,l}))_{l=1,\ldots,p_i, i=1,\ldots,r}$$

and satisfying

$$(X_{1,n}, \ldots, X_{r,n}) \xrightarrow{\mathcal{L}} (X_{0,1}, \ldots, X_{0,n}) \text{ in } V_1\times\ldots\times V_r, \tag{44}$$

where $V_1\times\ldots\times V_r$ is equipped with the norm

$$\|(x_1, \ldots, x_r)\| := \max_{i=1,\ldots,r} \|x_i\|.$$

Conversely, if (44) *holds with* $(X_{0,1}, \ldots, X_{0,n})$ *having sample paths in* $U^b(T_1, d_1) \times \ldots \times U^b(T_r, d_r)$, *then* (43) *holds true.*

Acknowledgements. The author would like to thank Professors Peter Gaenssler and Helmut Pruscha and Dr. Daniel Rost for their constant interest in this work and for many valuable discussions and helpful comments and suggestions.

REFERENCES

[1] P. Billingsley, *Convergence of Probability Measures*, Wiley, New York 1968.
[2] J. Boularan, L. Ferré and P. Vieu, *Location of particular points in non-parametric regression analysis*, Austral. J. Statist. 37 (1995), pp. 161–168.
[3] W. Eddy, *Optimal kernel estimators of the mode*, Ann. Statist. 8 (1980), pp. 870–882.
[4] W. Eddy, *The asymptotic distributions of kernel estimators of the mode*, Z. Wahrsch. Verw. Gebiete 59 (1982), pp. 279–290.
[5] W. Ehm, *Adaptive kernel estimation of a cusp-shaped mode*, in: *Applied Mathematics and Parallel Computing. Festschrift for Klaus Ritter*, H. Fischer et al. (Eds.), Physica-Verlag, Heidelberg 1996, pp. 109–120.
[6] P. Gaenssler and D. Rost, *Empirical and Partial-sum Processes Revisited as Random Measure Processes*, MaPhySto Lecture Notes No. 5, Department of Mathematical Sciences, University of Aarhus, Aarhus, Denmark, 1999.
[7] Th. Gasser, P. Hall and B. Presnell, *Nonparametric estimation of the mode of a distribution of random curves*, J. Roy. Statist. Soc. Ser. B 60 (1998), pp. 681–691.
[8] Th. Gasser and H.-G. Müller, *Kernel estimation of regression functions*, Lecture Notes in Math. 757, Springer, Berlin–New York 1979, pp. 23–68.
[9] B. Grund and P. Hall, *On the minimisation of L^p error in mode estimation*, Ann. Statist. 23 (1995), pp. 2264–2284.
[10] W. Härdle, *Applied Nonparametric Regression*, Cambridge University Press, Cambridge 1990.
[11] J. Harezlak and N. Heckman, *CriSP: A tool in bump hunting*, J. Computational and Graphical Statist. 10 (2001), pp. 713–729.
[12] N. Heckman, *Bump hunting in regression analysis*, Statist. Probab. Lett. 14 (1992), pp. 141–152.
[13] J. Hoffmann-Jørgensen, *Stochastic Processes on Polish Spaces*, 1984, unpublished.
[14] E. Mammen, J. S. Marron and N. I. Fisher, *Some asymptotics for multimodality tests based on kernel estimates*, Probab. Theory Related Fields 91 (1992), pp. 115–132.
[15] H.-G. Müller, *Kernel estimators of zeros and of location and size of extrema of regression functions*, Scand. J. Statist. 12 (1985), pp. 221–232.
[16] H.-G. Müller, *Nonparametric Regression Analysis of Longitudinal Data*, Lecture Notes in Statist. 46, Springer, 1988.
[17] H.-G. Müller, *Adaptive nonparametric peak estimation*, Ann. Statist. 17 (1989), pp. 1053–1069.
[18] H.-G. Müller and U. Stadtmüller, *Variable bandwidth kernel estimators of regression functions*, Ann. Statist. 15 (1987), pp. 182–201.
[19] E. A. Nadaraya, *On estimating regression*, Theory Probab. Appl. 10 (1964), pp. 186–190.
[20] E. A. Nadaraya, *Nonparametric Estimation of Probability Densities and Regression Curves*, Kluwer Academic Publishers, Dordrecht 1989.
[21] E. Parzen, *On estimation of a probability density function and mode*, Ann. Math. Statist. 33 (1962), pp. 1065–1076.

[22] J. P. Romano, *On weak convergence and optimality of kernel density estimates of the mode*, Ann. Statist. 16 (1988a), pp. 629–647.
[23] J. P. Romano, *Bootstrapping the mode*, Ann. Inst. Statist. Math. 40 (1988b), pp. 565–586.
[24] J. Shao and D. Tu, *The Jackknife and Bootstrap*, Springer, New York 1995.
[25] B. W. Silverman, *Using kernel density estimates to investigate multimodality*, J. Roy. Statist. Soc. Ser. B 43 (1981), pp. 97–99.
[26] A. W. van der Vaart and J. A. Wellner, *Weak Convergence and Empirical Processes with Applications to Statistics*, Springer, New York 1996.
[27] G. S. Watson, *Smooth regression analysis*, Sankhyā Ser. A 26 (1964), pp. 359–372.
[28] K. Ziegler, *Nonparametric estimation of location and size of maxima of regression functions in the random design case based on the Nadaraya–Watson estimator with data-dependent bandwidths*, Habilitationsschrift, Univ. of Munich, 2000.
[29] K. Ziegler, *On bootstrapping the mode in the nonparametric regression model with random design*, Metrika 53 (2001a), pp. 151–170.
[30] K. Ziegler, *On local bootstrap bandwidth choice in kernel density estimation*, submitted for publication (2001b); available under http://www.mathematik.tu-ilmenau.de/~ziegler/papers.html
[31] K. Ziegler, *On nonparametric kernel estimation of the mode of the regression function in the random design model*, J. Nonparametr. Statist. 14 (2002), pp. 749–774.
[32] K. Ziegler, *On the asymptotic normality of kernel regression estimators of the mode in the nonparametric random design model*, J. Statist. Plann. Inference 115 (2003), pp. 123–144.

Technical University of Ilmenau
Institute for Mathematics
Postfach 100565
D-98984 Ilmenau, Germany
http://www.mathematik.tu-ilmenau.de/~ziegler/kziegler.html

Received on 6.8.2003