Technische Universität Ilmenau
Institut für Mathematik

Rates of consistency for nonparametric estimation of the mode in absence of smoothness assumptions

Herrmann, Eva; Ziegler, Klaus

2003

Rates of consistency for nonparametric estimation of the mode in absence of smoothness assumptions

Eva Herrmann and Klaus Ziegler

Technical University of Darmstadt and Technical University of Ilmenau

Abstract Nonparametric estimation of the mode of a density or regression function via kernel methods is considered. It is shown that the rate of consistency of the mode estimator can be determined without the typical smoothness conditions. Only the uniform rate of the so-called stochastic part of the problem together with some mild conditions characterizing the shape or "acuteness" of the mode influence the rate of the mode estimator. In particular, outside the location of the mode, our assumptions do not even imply continuity. Overall, it turns out that the location of the mode can be estimated at a rate that is the better the "peakier" (and hence non-smooth) the mode is, while the contrary holds with estimation of the size of the mode.

1 Introduction and assumptions

An important problem in nonparametric curve estimation consists in estimation of the mode, i.e., the location of an isolated maximum of the unknown density or regression function. A number of distinguished papers deal with this topic. There are, among others, Parzen (1962), Rüschendorf (1977), Eddy (1980, 1982), Müller (1985, 1989), Romano (1988a,b), Grund and Hall (1995), Ehm (1996), and, most recently, Mokkadem and Pelletier (2003) as well as Abraham, Biau and Cadre (2003).

In the following, we will restrict to the univariate situation but extensions to the multivariate case are possible. The classical approach is as follows. Let f be the unknown real-valued curve and θ the *mode* of f, i.e.

$$f(\theta) > \sup_{|x-\theta|>\epsilon} f(x) \quad \text{for each } \epsilon > 0, \tag{1}$$

which means that θ is the location of the unique global maximum of f. Then θ is estimated from the location $\hat{\theta} \equiv \hat{\theta}_n$ of a maximum of a curve estimator $\hat{f} \equiv \hat{f}_n$ for f. Uniqueness of the maximum or even a condition like (1) for $\hat{f}$ cannot be expected here, but, in general, this does not affect the validity of asymptotic theory. It is well-known that uniform consistency of $\hat{f}_n$ for f is sufficient to ensure consistency of $\hat{\theta}_n$ for θ (see, e.g., Parzen, 1962; Rüschendorf, 1977; Nadaraya, 1989).

To obtain rates, however, one has to know some more about the "local geometry" of f around θ. Müller (1985) showed, in essence, that if the uniform consistency of $\hat{f}$ is of order β_n, i.e.

$$\sup_x |\hat{f}_n(x) - f(x)| = O(\beta_n) \tag{2}$$

(in probability or a.s.), and if there are $\rho > 0, c > 0$ such that

$$f(\theta) - f(x) \geq c|x - \theta|^\rho \quad \text{in a neighborhood of } \theta, \tag{3}$$

then

$$\hat{\theta}_n - \theta = O(\beta_n^{1/\rho}) \tag{4}$$

in probability or a.s. depending on what holds in (2). A similar concept was used by Boularan et al. (1995) for estimation of points θ with $f^{(p)}(\theta) = b$, i.e. where some p-th derivative of the curve takes a given value b.

So, (2) seems to be crucial for (4). But in order to obtain (2), global smoothness conditions have to be imposed on f. Most authors, among them Müller (1985), assume f to be twice continuously differentiable. Differentiability, however, excludes the case $\rho \leq 1$ in (3), (4), and hence his results do not apply to "cusp-shaped modes" ($\rho = 1$, see Ehm, 1996, where, however, C^p-smoothness, $p \geq 2$, of f is assumed outside of θ) or even "proper peaks" ($\rho < 1$). In fact, Müller's conditions tacitly even imply $\rho \geq 2$, and the case $\rho < 2$ doesn't seem to have been considered explicitly so far. It should be mentioned here that Härdle et al. require only some uniform local Lipschitz condition instead of differentiability in order to obtain (2), but this is still a global smoothness condition.

Results of type (2) are always proven by splitting into a *stochastic part* $\hat{f}_n(x) - f_n(x)$ and a *deterministic part* $f_n(x) - f(x)$ where in density estimation, $f_n(x) = \mathbb{E}\hat{f}_n(x)$ and the analytic part is simply the bias. A look at the proofs being available in the literature reveals that smoothness conditions are needed exclusively for handling the analytic part whereas a rate of the stochastic part

$$\sup_x |\hat{f}_n(x) - f_n(x)| = O(U_n) \tag{5}$$

(a.s. or in probability) can be determined without imposing any smoothness conditions. For example, if $\hat{f}_n$ is the Rosenblatt-Parzen kernel density estimator

$$\hat{f}_n(x) = \frac{1}{nh_n} \sum_{i=1}^n K\left(\frac{x - X_i}{h_n}\right) \tag{6}$$

based on i.i.d. observations X_i having density f, (5) can be established with a.s. rate

$$U_n = \sqrt{\frac{\log n}{nh_n}}$$

under some regularity conditions on the kernel K and the bandwidth h_n. An analogous result holds for regression estimators, see, e.g., Härdle et al. (1988) and can also be derived

2

from the results contained in Ziegler (2002). Similar results with somewhat worse rates are available in the case of dependent observations and in the multivariate case (see Györfi et al., 1989; Koshkin and Vasil'iev, 1998; Liebscher, 2001).

The aim of the present paper is to show that the knowledge of the rate of the stochastic part U_n in (6) together with some information about the shape or "acuteness" of the peak already suffice to prove a result like (4), with no smoothness conditions being required at all. Apart from θ, continuity of f will at most be required for (5), which is sometimes the case due to technical reasons, see e.g. Einmahl and Mason (1999) where even the constant in the O-term is determined.

In the sequel, we consider the slightly more general case of θ being the location of a local maximum of f, with (1) holding for x in some neighborhood I of θ. In part, the local shape of the peak is characterized by (3) (holding for x in some maybe smaller neighborhood $J \subset I$), but this gives only an upper bound for f around θ. A lower bound will be needed in addition. Therefore, we introduce the further condition that there are $\widetilde{\rho} > 0$, $d > 0$ such that

$$f(\theta) - f(x) \leq d|x - \theta|^{\widetilde{\rho}} \quad \text{in a neighborhood of } \theta. \tag{7}$$

Note, that (3) together with (7) implies $\widetilde{\rho} \leq \rho$. Indeed, $\widetilde{\rho} > \rho$ would imply $d|x - \theta|^{\widetilde{\rho}} < c|x - \theta|^{\rho}$ for $|x - \theta|$ being small enough. If, e.g., f is twice continuously differentiable in a neighborhood of θ, then, by Taylor's theorem and $f'(\theta) = 0$, the choice $\widetilde{\rho} = \rho = 2$ is possible. Even in this case, our result still improves on known ones, because this is only a *local* smoothness assumption while (2) always requires global smoothness.

If f has a "cusp-shaped" mode at θ, i.e. the one-sided derivatives exist in θ with $f'(\theta - 0) > 0$, $f'(\theta - 0) < 0$ and if f is continuously differentiable in left and right neighborhoods of θ, then $\widetilde{\rho} = \rho = 1$. For $\rho < 1$ there is necessarily $f'(\theta-0) = \infty$, $f'(\theta+0) = -\infty$. We will return to a discussion of these special cases in the remark after the corollary below and in Section 3.

In the next section, we will show for the density estimator (6) that (5), where again $f_n = \mathbb{E}\hat{f}_n$, together with (1), (3) and (7) imply (4) with $\beta_n = U_n + h_n^{\widetilde{\rho}}$. This is the same rate as being obtained in (4) from (2) from the global smoothness assumption that f is uniform local Lipschitz of order $\widetilde{\rho}$. We will also prove a rate for the estimation of the size $f(\theta)$ of the peak.

In Section 3 we compare our results to those being available in the literature and in Section 4 we indicate how to extend our techniques to regression analysis.

2 Main results

For simplicity, we give the result only for fixed, i.e., non-data-driven bandwidths and compactly supported kernels but we stress that they remain valid for data-dependent bandwidths and more general kernels. This can be achieved using the techniques described in Romano (1988a), Herrmann (2000) or Ziegler (2002). Our methods also apply to the estimation of modes of derivatives of f using integration by parts as, e.g., in Ziegler (2002). And finally we remark that extensions to the multivariate case are possible, too.

Theorem 1 Let $K \geq 0$ be a bounded and symmetric kernel function with $\int K(u)du = 1$ and compact support, $h_n > 0$, $h_n \to 0$ a bandwidth sequence, $\hat{f}_n$ the density estimator (6). We assume that there are a neighborhood I of θ and a numerical sequence U_n such that

$$\sup_{x \in I} |\hat{f}_n(x) - \mathbb{E}\hat{f}_n(x)| = O(U_n) \quad a.s. \tag{8}$$

Let f and its mode θ satisfy (1) for all $x \in I$. We also assume that (3), (7) hold and $\hat{\theta}_n$ is defined by

$$\hat{f}_n(\hat{\theta}_n) = \max_{x \in I} \hat{f}_n(x) . \tag{9}$$

Then

$$\hat{\theta}_n - \theta = O((U_n + h_n^{\tilde{\rho}})^{1/\rho}) \quad a.s. \tag{10}$$

Outline of proof Write again $f_n(x)$ for $\mathbb{E}\hat{f}_n(x)$. In the proof of Theorem 2.1 in Grund and Hall (1995) it has been shown (see also, Ziegler, 2002) that for each $\epsilon > 0$ the inequality $|\hat{\theta}_n - \theta| > \epsilon$ implies

$$\sup_{x \in I} |\hat{f}_n(x) - f_n(x)| \geq \frac{1}{2}(f_n(\theta) - \sup_{|x - \theta| > \epsilon} f_n(x)) .$$

Therefore, in order to derive (10) from (8), it suffices to find for each $\eta > 0$ some $\tau > 0$ such that, for $V_n = (U_n + h_n^{\tilde{\rho}})^{1/\rho}$ it holds that

$$f_n(\theta) - \sup_{|x - \theta| > \tau V_n} f_n(x) \geq \eta U_n . \tag{11}$$

Let $x \in I$ with $|x - \theta| > \tau V_n$ be given (with τ to be specified later). According to (1) and (3), there exists $\delta > 0$ such that

$$f(\theta) - f(x - h_n u) \geq \min(c|\theta - x + h_n u|^\rho, \delta) \geq \min(c|\theta - x|^\rho, \delta)$$

for either $u > 0$ or $u < 0$, depending on the sign of $\theta - x$. Choose $M > 0$ such that $\operatorname{supp} K \subset [-M, M]$. Then, according to (7)

$$f(\theta - h_n u) - f(\theta) \geq -dh_n^{\tilde{\rho}}|u|^{\tilde{\rho}}$$

for n large enough and $u \in [-M, M]$. Hence, with $\tilde{d} = d \int_{-M}^{M} |u|^{\tilde{\rho}} K(u)du < \infty$ and $\tilde{c} = c \int_0^M K(u)du = c \int_{-M}^0 K(u)du = \frac{1}{2}c$ we have for n large enough, by $K \geq 0$, that

$$
\begin{aligned}
f_n(\theta) - f_n(x) &= \int_{-M}^{M} K(u)(f(\theta - h_n u) - f(x - h_n u))du \\
&= \int_{-M}^{M} K(u)(f(\theta - h_n u) - f(\theta))du + \int_{-M}^{M} K(u)(f(\theta) - f(x - h_n u))du \\
&\geq -\tilde{d}h_n^{\tilde{\rho}} + \tilde{c}\tau^\rho V_n^\rho \\
&= -\tilde{d}h_n^{\tilde{\rho}} + \tilde{c}\tau^\rho h_n^{\tilde{\rho}} + \tilde{c}\tau^\rho U_n .
\end{aligned}
$$

4

If we now take τ such that $\widetilde{c}\tau^\rho \geq \max(\widetilde{d}, \eta)$, then (11) will be satisfied. $\qquad\square$

Remarks (a) By quite the same proof, we see that if (8) holds in probability instead of a.s., the result (10) is also obtained in probability instead of a.s.
(b) The existence of a θ_n with (9) is e.g. automatically ensured if K is taken to be continuous, since due to the compact support, $\hat{f}_n$ is also continuous and compactly supported then.

Of course, the estimation of the height $f(\theta)$ of the peak is of interest, too. A natural estimator is $\hat{f}_n(\hat{\theta}_n) = \sup_{x \in \mathbb{R}} \hat{f}_n(x)$. See, e.g., Nadaraya (1989) for results under smoothness conditions. A rate for the consistency of $\hat{f}_n(\hat{\theta}_n)$ can be derived from (8) and (10) if we replace (7) by a locally uniform version.

Theorem 2 Let the conditions of Theorem 1 be fulfilled. Instead of (7), assume that

$$|f(x) - f(y)| \leq d|x - y|^{\widetilde{\rho}} \quad x, y \text{ in a neighborhood of } \theta \tag{12}$$

holds for some $\widetilde{\rho}, d > 0$. Then

$$\hat{f}_n(\hat{\theta}_n) - f(\theta) = O((U_n + h_n^{\widetilde{\rho}})^{\widetilde{\rho}/\rho}) \quad a.s.$$

Proof First we note that from (12) it follows that

$$\mathbb{E}\hat{f}_n(x) - f(x) = O(h_n^{\widetilde{\rho}}) \quad \text{uniformly in } x \text{ in a neighborhood of } \theta \tag{13}$$

since $|\mathbb{E}\hat{f}_n(x) - f(x)| \leq \int_{-M}^{M} K(u)|f(x - uh_n) - f(x)|du \leq dh_n^{\widetilde{\rho}} \int_{-M}^{M} |u|^{\widetilde{\rho}}K(u)du$ for small enough h_n so that x and $x - uh_n$ are both in the neighborhood where (12) holds if x is in a certain smaller neighborhood.

Then, from (8), (10), (12) and (13),

$$
\begin{aligned}
\hat{f}_n(\hat{\theta}_n) - f(\theta) &= \hat{f}_n(\hat{\theta}_n) - f(\hat{\theta}_n) + f(\hat{\theta}_n) - f(\theta) \\
&= O(U_n + h_n^{\widetilde{\rho}}) + O(|\hat{\theta}_n - \theta|^{\widetilde{\rho}}) \\
&= O((U_n + h_n^{\widetilde{\rho}})^{\widetilde{\rho}/\rho})
\end{aligned}
$$

where we have used $\widetilde{\rho} \leq \rho$. $\qquad\square$

Remarks (a) Again we obtain the assertion of Theorem 2 in probability if (8) is assumed to hold in probability.
(b) The condition (12) means that f is locally uniform Lipschitz of order $\widetilde{\rho}$ in a neighborhood of θ, which is stronger than (7). Indeed, it is known that for estimation of the size of the mode, stronger smoothness assumptions are required than for estimation of the location of the mode. See, e.g., Ziegler (2002).

As we have mentioned in the introduction, in case of i.i.d. observations X_i, the rate $U_n = \sqrt{\frac{\log n}{n h_n}}$ being attained a.s. for the stochastic part is familiar. Here, the rate of $U_n + h_n^{\widetilde{\rho}}$ becomes best if we choose

$$h_n = O\left(\left(\frac{\log n}{n}\right)^{\frac{1}{2\widetilde{\rho}+1}}\right),$$

which leads to

$$U_n + h_n^{\widetilde{\rho}} = O\left(\left(\frac{\log n}{n}\right)^{\frac{\widetilde{\rho}}{2\widetilde{\rho}+1}}\right).$$

Indeed, for minimization of $U_n + h_n^{\widetilde{\rho}}$ the rates of U_n and $h_n^{\widetilde{\rho}}$ must coincide, whence $h_n^{\widetilde{\rho}} = O(\sqrt{\frac{\log n}{n h_n}})$.

Hence we have the following corollary:

Corollary Let the density estimator be based on i.i.d. samples. Then, we have under the assumptions of Theorem 1

$$\hat{\theta}_n - \theta = O\left(\left(\frac{\log n}{n}\right)^{\frac{\widetilde{\rho}}{(2\widetilde{\rho}+1)\rho}}\right) \tag{14}$$

(a.s.), while under the assumptions of Theorem 2 it holds that

$$\hat{f}_n(\hat{\theta}_n) - f(\theta) = O\left(\left(\frac{\log n}{n}\right)^{\frac{\widetilde{\rho}^2}{(2\widetilde{\rho}+1)\rho}}\right) \tag{15}$$

(a.s.) $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

Remarks (a) Assume $\rho = \widetilde{\rho}$ which will cover most situations anyway. Then, we see from (14) that the rate of $\hat{\theta}_n - \theta$ is $O((\frac{\log n}{n})^{1/(2\rho+1)}) = O(h_n)$ which improves as ρ decreases, i.e., as the peak gets "acuter". Instead, the rate of $\hat{f}_n(\hat{\theta}_n) - f(\theta)$ is $O(\frac{\log n}{n})^{\rho/(2\rho+1)})$ which worsens with ρ getting smaller. Naturally, the latter is the same rate at which f can be estimated uniformly if it is uniformly local Lipschitz of order ρ. This behavior is roughly what we should have expected before. A "high and slim" peak is met more exactly by the estimator, while its height will be "abraded" by the smoothing process. Furthermore, it is natural that the bandwidth should be chosen the smaller the acuter the peak is.

(b) Note that our assumptions in Theorem 1 do not even imply continuity except in θ itself. Indeed, the case $\widetilde{\rho} < \rho$ may allow f to jump and oscillate quite heavily outside θ. In Theorem 2, the condition (12) implies continuity of f in a neighborhood of θ. However, the proof shows that Theorem 2 holds under the weaker condition (13) which might be valid without continuity in special situations.

6

3 Brief discussion

(a) In the twice differentiable case $\rho = \widetilde{\rho} = 2$ our rates coincide with the known ones in Müller (1985; see also Vieu, 1996). However, our results are still a slight improvement in this case since we need to impose differentiability only locally in a small neighborhood of θ. On the other hand, under some additional requirements, the rate has been slightly improved by Leclerc and Pierre-Loti-Viaud (2000). In the degenerate case, i.e. $f''(\theta) = 0$ with some higher differentiability, the exact rate has been recently determined by Mokkadem and Pelletier (2003). In the case $\rho < 2$, no results seem to have been available so far. This has also been pointed out by Abraham et al. (2003), p.7. See, however, Ehm, 1996, where f is assumed to be C^p-smooth, $p \geq 2$ except at θ itself, where it has a "kink" ($\rho = \widetilde{\rho} = 1$). In this case, our rate (14), i.e. $O(n^{-1/3})$ can be improved by the very sophisticated construction of another estimator for θ.

(b) In Abraham et al. (2003), for computational reasons, a different estimator is considered which maximizes $\hat{f}_n$ only over the values of $X_1, \ldots, X_n$ (instead of maximizing over $x \in \mathbb{R}$ or an interval which the classical mode estimate does). The authors compare the performance of their estimator to that of the classical one in the smooth case, and state that it would be desirable to do the like in the non-smooth case. Now even as we have results for the classical estimator in non-smooth cases, such a comparison is difficult since the conditions given in Abraham et al. (2003) do not directly correspond to ours. However, their β clearly equals our $\widetilde{\rho}$ since they employ our condition (13), while their α should correspond to our $1/\rho$. Furthermore, we are in the univariate case $d = 1$. With i.i.d. observations being available, the rate obtained from (14) is $O((\frac{\log n}{n})^{\widetilde{\rho}/(2\widetilde{\rho}+1)\rho}) = O((\frac{\log n}{n})^{\alpha\beta/(2\beta+d)})$. Since $\alpha\beta \leq 1$, this is slightly better than the rate given in Cor. 2.1 of Abraham et al. (2003) which is $O(\frac{(\log n)^{2/(2\beta+d)}}{n^{\alpha\beta/(2\beta+d)}})$. Hence, even in the non-smooth case, the classical estimator still seems to perform slightly superior to the computationally advantageous one. However, we do not know if one of those rates can still be improved.

(c) Note that our assumptions, even in the case $\rho = \widetilde{\rho}$, do not imply any local symmetry of f around θ. The situation changes dramatically as soon as we want to construct a confidence interval for θ. For asymptotic normality of $\hat{\theta}_n - \theta$ the conditions $f'(\theta - 0) = -f'(\theta + 0)$ and $f''(\theta - 0) = f''(\theta + 0)$, i.e., local symmetry of f around θ up to order 2, seem to be crucial. This will be shown in a forthcoming paper of the second author. However, even in non-symmetric situations of this kind, the mode can still be estimated asymptotically normal using a different estimator. The construction of such an estimator is described in Ehm (1996).

4 Extensions to regression functions

Now we turn to regression analysis. For fixed design regression

$$Y_i = f(x_i) + \epsilon_i$$

with design points $x_1, \ldots, x_n \in I = [a, b]$ satisfying $|x_i - x_{i-1} - \frac{1}{n}| = o(\frac{1}{n})$ uniformly in $2 \le i \le n$ and i.i.d. error variables with zero mean and variance $\mathbb{E}\epsilon_i^2 = \sigma^2 < \infty$, the Gasser-Müller estimator

$$\hat{f}_n(x) = \frac{1}{h_n} \sum_{i=1}^{n} \int_{s_{i-1}}^{s_i} K(\frac{x-u}{h_n}) du$$

(with $s_{i-1} = \frac{1}{2}(x_i + x_{i-1}), \; i = 2, \ldots, n, \; s_0 = a, s_n = b$) is known to fulfill

$$\mathbb{E}\hat{f}_n(x) = \int K(u) f(x - h_n u) du + O(\frac{1}{nh_n})$$

(Müller, 1985). Therefore, our results take over to this case quite straightforwardly as long as $\frac{1}{nh_n}$ is of smaller order than both U_n and $h_n^{\tilde{\rho}}$ which is the typical case. In the case of random design, with

$$f(x) = \mathbb{E}(Y|X = x)$$

to be estimated via the Nadaraya-Watson estimator

$$\hat{f}_n(x) = \frac{\hat{r}_n(x)}{\hat{g}_n(x)} \;,$$

where

$$\hat{r}_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} Y_i K(\frac{x - X_i}{h_n})$$

and

$$\hat{g}_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K(\frac{x - X_i}{h_n}) \;,$$

the quantity $\hat{f}_n$ is compared to in the stochastic part (5) is

$$f_n(x) = \frac{\mathbb{E}\hat{r}_n(x)}{\mathbb{E}\hat{g}_n(x)}$$

rather than $\mathbb{E}\hat{f}_n(x)$. Note, however, that the asymptotic equivalence of the two quantities is shown in Ziegler (2001a). For estimation from i.i.d. pairs of observations, (5) can still be proven with

$$U_n = \sqrt{\frac{\log n}{nh_n}}$$

8

(see Härdle et al., 1988). For U_n in the case of dependent observations, see Györfi et al. (1989) or Liebscher (1998) among others.

To prove an analogue to our theorem for the Nadaraya-Watson estimator, one has to find an appropriate lower bound of $f_n(\theta) - f_n(x) = \frac{\mathbb{E}\hat{r}_n(\theta)}{\mathbb{E}\hat{g}_n(\theta)} - \frac{\mathbb{E}\hat{r}_n(x)}{\mathbb{E}\hat{g}_n(x)}$. If X has a design density g being bounded away from zero and infinity on I, i.e.

$$0 < C_1 \leq g(x) \leq C_2 < \infty \quad \text{for } x \in I,$$

we have

$$
\begin{aligned}
&f_n(\theta) - f_n(x) \\
=\ & \frac{\int K(u)f(\theta - uh)g(\theta - uh)du}{\int K(u)g(\theta - uh)du} - \frac{\int K(u)f(x - uh)g(x - uh)du}{\int K(u)g(x - uh)du} \\
=\ & \frac{\int K(u)(f(\theta - uh) - f(\theta))g(\theta - uh)du}{\int K(u)g(\theta - uh)du} + \frac{\int K(u)(f(\theta) - f(x - uh))g(x - uh)du}{\int K(u)g(x - uh)du} \\
\geq\ & \frac{C_2}{C_1}\int K(u)(f(\theta - uh) - f(\theta))du + \frac{C_1}{C_2}\int K(u)(f(\theta) - f(x - uh))du\ ,
\end{aligned}
$$

and from now on one may proceed as in the proof of the theorem.

Further extensions to local polynomial smoothers are possible in a similar way.

Finally we remark that our method should take over to the estimation of points θ with $f^{(p)} = b$ as mentioned in the introduction by modifying assumptions (3) and (7) appropriately. This will be shown in a forthcoming paper of the second author.

References

Abraham, C., Biau, G. and Cadre, B. (2003). Simple estimation of the mode of a multivariate density. Preprint.

Boularan, J., Ferré, L. and Vieu, P. (1995). Location of particular points in non-parametric regression analysis. *Aust. J. Stat.* **37**, 161-168.

Eddy, W. (1980). Optimal kernel estimators of the mode. *Ann. Statist.* **8** 870-882.

Eddy, W. (1982). The asymptotic distributions of kernel estimators of the mode. *Z. Wahrsch. Verw. Gebiete* **59** 279-290.

Ehm, W. (1996). Adaptive kernel estimation of a cusp-shaped mode. In: Fischer, Herbert (ed.) et al.: *Applied mathematics and parallel computing. Festschrift for Klaus Ritter*, 109-120, Physica-Verlag, Heidelberg.

Einmahl, U. and Mason, D.M. (2000). An empirical process approach to the uniform consistency of kernel-type function estimators. *J. Theoret. Probab.* **13**, 1 - 37.

Grund, B. and Hall, P. (1995). On the minimisation of L^p error in mode estimation. *Ann. Statist.* **23** 2264-2284.

Györfi, L., Härdle, W., Sarda, P. and Vieu, P. (1989). *Nonparametric Estimation from Time Series.* Lecture Notes in Statistics **60**, Springer-Verlag.

Härdle, W., Janssen, P. and Serfling, R. (1988). Strong uniform consistency rates for estimators of conditional functionals. *Ann. Statist.* **16** 1428-1449.

Koshkin, G.M. and Vasil'iev, V.A. (1998). Nonparametric estimation of derivatives of a multivariate density from dependent observations. *Mathematical Methods Statist.* **7** 361-400.

Herrmann, E. (2000) *Data adaptive kernel regression estimation.* Habilitationsschrift, Darmstadt Technical University.

Leclerc, J. and Pierre-Loti-Viaud, D. (2000). Vitesse de convergence presque sure de l'estimateur a noyau du mode. *Comptes rendus de l'Academie des Sciences de Paris* **331** 637-640.

Liebscher, E. (2001). Estimation of the density and the regression function under mixing condition. *Statistics and Decisions* **19**, 9-26.

Mokkadem, A. and Pelletier, M. (2003): The law of the iterated logarithm for the multivariate kernel mode estimator. *ESAIM: Probability and Statistics* **7** 1-21.

Müller, H.-G. (1985). Kernel estimators of zeros and of location and size of extrema of regression functions. Scand. J. Statist. **12** 221-232.

Müller, H.-G. (1989). Adaptive nonparametric peak estimation. *Ann. Statist.* **17** 1053-1069.

Nadaraya, E.A. (1989). *Nonparametric Estimation of Probability Densities and Regression Curves.* Kluver Academic Publishers, Dordrecht.

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

Romano, J.P. (1988a). On weak convergence and optimality of kernel density estimates of the mode. *Ann. Statist.* **16** 629-647.

Romano, J.P. (1988b). Bootstrapping the mode. *Ann. Inst. Statist. Math.* **40** 565-586.

Rüschendorf, L. (1977). Consistency of estimators for multivariate density functions and for the mode. *Sankhyā Ser. A* **39** 243-250.

Vieu, P. (1996). A note on density mode estimation. *Statistics and Probability Letters* **26** 297-307.

Ziegler, K. (2001): On approximations to the bias of the Nadaraya-Watson regression estimator. *J. Nonparametric Statistics* **13** 583-589.

Ziegler, K. (2002): On nonparametric kernel estimation of the mode of the regression function in the random design model. *J. Nonparametric Statistics* **14** 749-774.