Adaptive Estimation of the Mode of a Multivariate Density

Jussi Klemelä [*]
University of Mannheim
Economics Faculty, Department of Statistics
L 7 3-5 Verfügungsgebäude
68131 Mannheim, Germany
Email: klemela@rumms.uni-mannheim.de
Fax +49 621 1811931

Abstract

Estimation of the location of the mode of a multivariate density function is studied. We use the maximizer of a kernel estimator as an estimator for the location of the mode and we discuss the choice of the smoothing parameter of the kernel estimator to be maximized. This smoothing parameter is chosen in an adaptive way, that is, without assuming knowledge of the smoothness of the density function. It is proved that the proposed method has the optimal adaptive rate of convergence.

Mathematics Subject Classifications: 62G05, 62G20
Key Words: Adaptive curve estimation, Bandwidth selection, Kernel estimation, Minimax risk.
Abbreviated title: Adaptive Estimation of the Mode.

[*]Writing of this article was financed by Deutsche Forschungsgemeinschaft under project MA1026/6-2.

1 Introduction

The maximizer of a kernel estimator as an estimator for the location of the mode was proposed by Parzen (1962). For the success of this method it is essential to be able to choose the smoothing parameter of the kernel estimator in an efficient way. The methods for smoothing parameter selection which are designed for the estimation of the density function itself are not optimal when we want to estimate the location of the mode. We propose a method for smoothing parameter selection which is tailor designed for the estimation of the location of the mode.

Much of the work concerning the estimation of the location of the mode is concerned with the consistency and asymptotic normality of the estimates. For the multivariate case, see for example Konakov (1973). For work on other types of nonparametric estimators than the maximizer of a kernel estimator see for example Grenander (1965), who considered an estimator based on order statistics, or Tsybakov (1990), who considered a recursive algorithm. We will assume prior knowledge that the density has only one mode. For the problem of finding out the number of modes, see for example Müller and Sawitzki (1991) and Mammen, Marron and Fisher (1992). It is known that rates of convergence for the estimation of the location of the mode are slower than the rates for the estimation of the density itself. In fact, the rates are the same as for the estimation of the first derivative of the density. The upper bound for the rates was derived by Parzen (1962) and the lower bound by Hasminskii (1979). For minimax rates, see also Donoho and Liu (1991) and for the multivariate case Tsybakov (1990).

Work on empirical smoothing parameter selection for the kernel estimator to be maximized includes Müller (1989), who considers local smoothing parameter selection in the setting of regression estimation, and Grund and Hall (1995) who considered bootstrap methods for the choice of the smoothing parameter. One alternative option to choosing the smoothing parameter is to use the plug-in method. This method was studied by Ziegler (2000) in the connection of the estimation of the mode of a regression function. However, to apply the plug-in method one has to know beforehand the smoothness of the underlying function. Our aim is to construct an estimator which is able to adapt to the unknown smoothness of the underlying function.

We will choose the smoothing parameter with the so called Lepski method. When using this method, we do not have to know the smoothness of the underlying function. This type of method was proposed by Lepski (1992), who considered a white noise model and defined the estimator of the mode as a point of zero of a certain kernel estimator of the derivative of a univariate signal function. We develop further this approach and extend it to density estimation and multivariate functions. Note that the location of the mode of a multivariate function is a vector valued parameter, and previously the Lepski method was used only in connection with the estimation of real valued functionals. Also, the Lepski method has been until recently considered only in the Gaussian white noise model, for an exception see Butucea (2000b).

When considering adaptation to the unknown smoothness of the function, the optimal rates for the estimation of many functionals are slower than the minimax rates. This is the case also for the estimation of the location of the mode. Our first theorem states that the optimal adaptive rates of convergence are slower than the minimax rates by a logarithmic factor. The second theorem gives the construction of a rate optimal adaptive estimator. This estimator is able to adapt to the unknown smoothness on an increasing interval, that is, we will be able to estimate efficiently the location of the mode of functions, whose number of derivatives lies on the interval $[s_*, s_n]$, where $2 < s_* < s_n \to \infty$, as $n \to \infty$.

We denote with Df the gradient of function $f : \mathbf{R}^d \to \mathbf{R}$, that is, Df is the $d \times 1$-vector of partial derivatives of f. Let $D_i f(x) = (\partial/\partial x_i) f(x)$. We denote with $D^2 f$ the Hessian of function f, that is, $D^2 f$ is the $d \times d$-matrix of second partial derivatives of f. Let $D_i D_j f(x) = (\partial^2/\partial x_i \partial x_j) f(x)$. We let $\|x\|$ denote the Euclidean norm of $x = (x_1, \ldots, x_d) \in \mathbf{R}^d$, and $\|x\|_{max} = \max_{i=1,\ldots,d} |x_i|$. The L_p norm of functions $\mathbf{R}^d \to \mathbf{R}$ is denoted with $\| \cdot \|_p$, $1 \le p \le \infty$. When $\alpha = (\alpha_1, \ldots, \alpha_d)$ is a multi-index, let $|\alpha| = \alpha_1 + \cdots + \alpha_d$, $\alpha! = \alpha_1! \cdots \alpha_d!$, and, for $x \in \mathbf{R}^d$, let $x^\alpha = x_1^{\alpha_1} \cdots x_d^{\alpha_d}$. When $\lim_{n\to\infty}(a_n/b_n) = 1$, we write $a_n \sim b_n$. The support of function f is denoted with $\mathrm{supp}(f)$. For $a \in \mathbf{R}$, let $\lceil a \rceil$ be the smallest integer $\ge a$.

Results are formulated in Section 2, a simulation study is presented in Section 3, and proofs of the results are given in Section 4.

2 Results

We observe i.i.d. $\mathbf{R}^d$-valued random vectors $X_1, \ldots, X_n$, which have density function $f : \mathbf{R}^d \to \mathbf{R}$, $d \ge 1$. We assume that density $f : \mathbf{R}^d \to \mathbf{R}$ has a unique maximum and we want to estimate the location of this maximum, that is, we want to estimate the functional $x_f = \mathrm{argmax}_{x \in \mathbf{R}^d} f(x)$. We consider functional classes of the form

$$\mathcal{F}_s = \left\{ f : \mathbf{R}^d \to \mathbf{R} \,\middle|\, \rho_s(f) \le L^*, \ \ f \text{ satisfies Assumption 1} \right\}$$

where

$$\rho_s(f) = \sup_{x \ne 0} \|x\|^{-s} \left| f(x_f + x) - \sum_{i=0}^{\lceil s-1 \rceil} \sum_{|\alpha|=i} \frac{x^\alpha}{\alpha!} \left. \frac{\partial^i f(y)}{\partial^\alpha y} \right|_{y=x_f} \right|,$$

$0 < s, L^* < \infty$, and the assumptions to be needed are the following.

Assumption 1 *Density $f : \mathbf{R}^d \to \mathbf{R}$ satisfies for $i, j = 1, \ldots, d$,*

1. *f has a unique maximum at x_f, $f(x_f) \le f^*$, where $0 < f^* < \infty$,*

2. *denote* $U = \mathrm{supp}(f) \subset \mathbf{R}^d$, *the set* U *is convex and* $\sup_{x,y \in U} \|x - y\| \le \gamma^*$, $0 < \gamma^* < \infty$,

3. *for* $x \in U$, $|\det(D^2 f(x))| > 0$, $\sup_{x \in U} \left| \left([D^2 f(x)]^{-1} \right)_{i,j} \right| \le C^*$, *where* $0 < C^* < \infty$, *and* $([D^2 f(x_f)]^{-1})_{i,j}$ *are the elements of the inverse of the* $d \times d$-*matrix* $D^2 f(x_f)$,

4. $\sup_{x \in U} \sup_{y: \|y - x\| \le \delta} |D_i D_j f(y) - D_i D_j f(x)| \le \Delta(\delta)$, *where* $\lim_{\delta \to 0} \Delta(\delta) = 0$,

5. $\|D_i D_j f\|_\infty \le f^{**}$, *where* $0 < f^{**} < \infty$.

Note that smoothness condition $\rho_s(f) \le L^*$ is a local smoothness condition, imposing a smoothness condition only in a neighborhood of the location x_f of the mode. We may call ρ_s the *Taylor semi-norm*, since it measures how well f may be approximated by a Taylor polynomial in a neighborhood of x_f. On the other hand, Assumption 1 poses additional global smoothness conditions, but only on the 2nd partial derivatives of f.

We will consider a scale of classes $\mathcal{F}_s$, $s \in A \subset (2, \infty)$, and we want to construct an adaptive estimator in the sense that the estimator will be in a certain sense optimal simultaneously over the members of the scale of classes.

Define the supremum of normalized risk with normalizing factor $\psi > 0$ of estimator $x_n \in \mathbf{R}^d$ by

$$\mathcal{R}_{n,s}(x_n, \psi) = \sup_{f \in \mathcal{F}_s} E_f \left(\psi^{-p} \|x_n - x_f\|^p \right) \tag{1}$$

where we take $p \ge 1$. It is possible to prove that the minimax rate of convergence for parameter spaces $\mathcal{F}_s$, $s \in [2, \infty)$, is $\varphi_s = n^{-(s-1)/(2s+d)}$, that is, there exists an estimator $\hat{x}_n$ such that for $s \in [2, \infty)$

$$c \le \liminf_{n \to \infty} \inf_{x_n} \mathcal{R}_{n,s}(x_n, \varphi_s) \le \limsup_{n \to \infty} \mathcal{R}_{n,s}(\hat{x}_n, \varphi_s) \le C \tag{2}$$

for some positive constants c, C, where the infimum is over all $\mathbf{R}^d$ valued estimators x_n. The lower bound was proved by Hasminskii (1979) for the one dimensional case and both upper and lower bounds were proved by Tsybakov (1990) for the multivariate case, for Hölder classes. Also Donoho and Liu (1991) proved both upper and lower bounds for the one dimensional case. The property (2) may be proved for our smoothness class almost exactly the same way as the results for Hölder classes.

It was proved by Lepski (1992) for the one-dimensional Gaussian white noise model that in the estimation of the location of the mode the adaptive rate of convergence is slower than the minimax rate of convergence by a logarithmic factor. Let us consider adaptive rates of convergence in the present setting using the following definition from Tsybakov (1998).

Definition 1 *The normalizing factor $\Psi_{s,n}$ is called the ADAPTIVE RATE OF CONVERGENCE on the scale of classes $\{\mathcal{F}_s, s \in A\}$ for the risk as defined in (1) if:*

1. *for some estimator x_n^* and some positive constant C,*

$$\limsup_{n\to\infty} \sup_{s\in A} \mathcal{R}_{n,s}(x_n^*, \Psi_{s,n}) \le C, \tag{3}$$

2. *if rate of convergence $S_{s,n} > 0$ satisfies for some estimator x_n^{**} and for some positive constant C',*

$$\limsup_{n\to\infty} \sup_{s\in A} \mathcal{R}_{n,s}(x_n^{**}, S_{s,n}) \le C', \tag{4}$$

and the condition: there exists $s' \in A$ such that

$$\lim_{n\to\infty} (S_{s',n}/\Psi_{s',n}) = 0 \tag{5}$$

then there exists $s'' \in A$ such that

$$\lim_{n\to\infty} (S_{s',n}/\Psi_{s',n})\,(S_{s'',n}/\Psi_{s'',n}) = \infty. \tag{6}$$

Remark 1. In words, if rate of convergence $S_{s,n}$ is such that at some point s' it is faster than the adaptive rate of convergence $\Psi_{s',n}$, then at some other point s'', $S_{s'',n}$ is much slower than $\Psi_{s'',n}$. In fact, according to (6), $(S_{s'',n}/\Psi_{s'',n}) \to \infty$ even faster than $(S_{s',n}/\Psi_{s',n})$ approaches zero. To call rate $\Psi_{s,n}$ the adaptive rate of convergence, it is not enough that condition (3) is satisfied for some estimator x_n^*. It might happen that some other estimator has a better rate of convergence at some point $s' \in A$ of the range of smoothness indices. However, we want to call a rate an "adaptive rate of convergence" only when it is an optimal rate in some sense. Condition 2 of Definition 1 adds the requirement of optimality: if there exists an estimator x_n^{**} which has a better rate of convergence at some point in the range of smoothness indices, then there exists a point in the range of smoothness indices where the rate of estimator x_n^{**} is much worse than the adaptive rate of convergence. Adaptive rates in the setting of density estimation were considered by Butucea (2000a) and exact constants in the adaptive estimation were considered by Klemelä and Tsybakov (2001).

In the following theorem we give the adaptive rate of convergence for the present setting. We take $A = [s_*, s_n]$ in Definition 1, where $s_* > 2$ and

$$s_n = \kappa(\log n)^a \tag{7}$$

for some $\kappa, a > 0$.

Theorem 1 *The adaptive rate of convergence for the scale of classes $\{\mathcal{F}_s, s \in [s_*, s_n]\}$, where $s_* > 2$ and s_n satisfies (7), is*

$$\psi_s = \begin{cases} (\log n/n)^{(s-1)/(2s+d)} & \text{for } s \in [s_*, s_n), \\ n^{-(s-1)/(2s+d)} & \text{for } s = s_n. \end{cases} \tag{8}$$

A proof of Theorem 1 is given in Section 4.2. Next we define a rate adaptive estimator, that is, an estimator satisfying (3) with the normalizing factor ψ_s defined in (8).

We will introduce a sufficiently fine grid on $[s_*, s_n]$ and a statistic $\hat{s}$ having values on this grid. To each point of the grid we assign a linear kernel estimator and a maximizer of this estimator. The statistic $\hat{s}$ will choose one of these maximizers. We consider the grid $S = \{s_1, \ldots, s_m\}$, where the values s_i and the integer m are chosen such that: $s_* = s_1 < \cdots < s_m = s_n$. Furthermore,

$$\kappa_1(\log n)^{-a_1} \le s_{i+1} - s_i \le \kappa_2(\log n)^{-a_2}, \quad i = 1, \ldots, m-1, \tag{9}$$

where $a_1 \ge a_2 > 1$, $\kappa_2 > \kappa_1 > 0$. We introduce the notation

$$\tilde{n}(s) = \begin{cases} n/\log n & \text{for } s = s_1, \ldots, s_{m-1}, \\ n & \text{for } s = s_m. \end{cases} \tag{10}$$

For any $s \in S$ introduce a linear kernel estimator of $f(x)$:

$$f_{s,n}(x) = \frac{1}{n}\sum_{i=1}^{n} K_{s,h(s,\tilde{n}(s))}(x - X_i) \tag{11}$$

where $K_{s,h} = K_s(\,\cdot\,/h)/h^d$ and

$$h(s,n) = n^{-1/(2s+d)}. \tag{12}$$

We require the following conditions on the kernel function $K_s : \mathbf{R}^d \to \mathbf{R}$.

Condition 1 *For K_s, $s \in [s_*, \infty)$, $i, j = 1, \ldots, d$,*

1. $\int_{\mathbf{R}^d} |K_s| < \infty$, $\int_{\mathbf{R}^d} K_s = 1$,
2. *K_s has compact support,* $\sup_{s\in[s_*,\infty)} \sup_{x,y\in \mathrm{supp}(K_s)} \|x - y\| < \infty$,
3. $\sup_{s_* \le s' \le s < \infty} \int \|z\|^s \, |D_i K_{s'}(z)| \, dz < \infty$,
4. $\int z^\alpha (D_i K_s)(z) dz = 0$ *when* $|\alpha| = 0, \ldots, \lceil s - 1 \rceil$, $|\alpha| \ne 1$,
5. $\sup_{s\in[s^*,\infty)} \|D_i K_s\|_m < \infty$ *for* $2 \le m < \infty$,
6. $\inf_{s\in[s^*,\infty)} \|D_i K_s\|_2 > 0$,

7. $\sup_{s\in[s_*,\infty)} \|D_iD_jK_s\|_2 < \infty$, $\sup_{s\in[s_*,\infty)} \|D_iD_jK_s\|_\infty < \infty$,

8. $\sup_{x,y\in\mathbf{R}^d} |D_iD_jK_s(x) - D_iD_jK_s(y)| \le \|x-y\|^\alpha$ *for some* $0 < \alpha \le 1$.

Define estimator $\hat{x}_s$ as a maximizer of $f_{s,n}(x)$,

$$f_{s,n}(\hat{x}_s) = \max_{y\in\mathbf{R}^d} f_{s,n}(y). \tag{13}$$

If there exists several maximizers, we choose $\hat{x}_s$ to be one of these. The adaptive estimator has the form $\hat{x}_{\hat{s}}$ where $\hat{s}$ is a suitably chosen statistic. To define $\hat{s}$ we follow the approach of Lepski (1990), (1991), (1992). That is, the statistic $\hat{s}$ is defined as the largest of those s-values in the grid for which the components of estimator $\hat{x}_s$ do not differ significantly from the components of the estimators corresponding to the smaller s-values. We choose

$$\hat{s} = \max\{s \in S : \|\hat{x}_s - \hat{x}_{s'}\|_{max} \le \eta(s') \text{ for all } s' \in S,\ s' \le s\}$$

with the threshold

$$\eta(s) = C_\eta(\log n/n)^{(s-1)/(2s+d)} \tag{14}$$

with sufficient large constant $C_\eta > 0$ (see Equation (43) on page 20). Finally, define the estimator of x_f as

$$x_n^* = \hat{x}_{\hat{s}}. \tag{15}$$

The next theorem states that estimator x_n^* is rate adaptive.

Theorem 2 *Let the normalized risk be defined by (1) and let normalising factor* ψ_s *be defined in (8). Then, for* $s_* > 2$ *and* s_n *satisfying (7),*

$$\limsup_{n\to\infty} \sup_{s\in[s_*,s_n]} \mathcal{R}_{n,s}(x_n^*, \psi_s) < \infty,$$

when the kernel function satisfies Condition 1.

A proof of Theorem 2 is given in Section 4.3.

3 Simulations

We consider 2 simulation examples: a skew density and a truncation of a skew density. When the density is symmetric, the location of the mode of the kernel estimator is robust with respect to the choice of the smoothing parameter; we get precise estimates of the mode for a large range of smoothing parameters. For skewed densities the choice of the smoothing parameter is more important.

The kernel estimator is not efficient for estimating densities with discontinuities. However, for mode estimation, the density needs to be smooth only in a

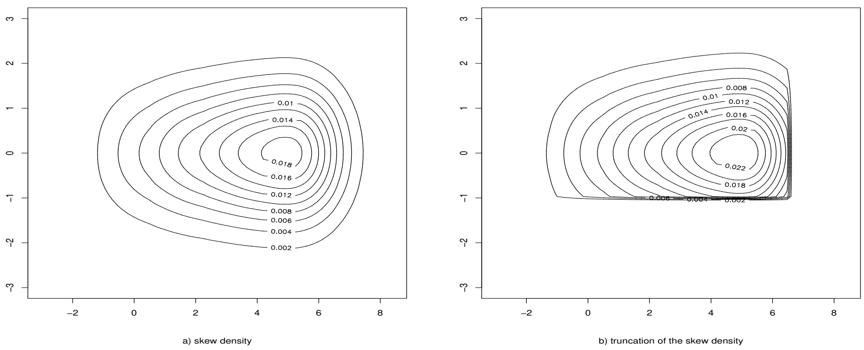

Figure 1: Contour plots of the skew density and the truncation of the skew density.

neighborhood of the mode. We show this by considering a truncation of the skew density.

We will show that in our simulation examples the empirically chosen smoothing parameter is often very close to the optimal smoothing parameter.

We studied also the effect of the threshold $\eta(s)$, defined in (14), on the accuracy of the procedure. We show that the procedure is robust with respect to the choice of factor C_η in (14). We do not report studies on the effect of the range and fineness of the grid of smoothing parameters. Based on our experience these tuning parameters have less effect than the threshold $\eta(s)$.

Skew density. The density of the first simulation example is the skew density shown in Figure 1 a). The density is in fact a mixture of four shifted standard 2-dimensional Gaussian densities; mixture centers are $(0,0)$, $(1.8,0)$, $(3.6,0)$, $(5.4,0)$. The location of the mode is approximately $(4.89,0)$.

We generated 100 samples of size $n = 1000$. We considered the grid of smoothing parameters $h \in \{0.57, 0.86, 1.11, 1.31, 1.47, 1.61, 1.73, 1.82\}$, which corresponds to the grid $s \in \{0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0\}$ of smoothness indices when we choose $h = 3(\log_e(n)/n)^{1/(2s+2)}$. We applied the Bartlett-Epanechnikov product kernel $K(x) = (3/4)^2 \Pi_{i=1}^2 (1 - x_i^2)_+$, where $(a)_+ = \max\{0, a\}$.

Figure 2 a) shows the number of times each smoothing parameter was optimal. We say that h_{opt} is optimal, for a given sample and density f, when

$$h_{opt} = h_{opt}(X_1, \ldots, X_n, f) = \mathrm{arginf}_{h>0} \|x_{\hat{f}_h} - x_f\|$$

where $\hat{f}_h$ is the kernel estimator with the Bartlett-Epanechnikov product kernel, smoothing parameter h, and based on sample $X_1, \ldots, X_n$, $x_{\hat{f}_h}$ is the mode of the kernel estimate and x_f is the mode of the true density f. We restrict h_{opt} to the grid of possible values for h. Figure 2 a) gives a barplot showing how many times each smoothing parameter in the grid was optimal for 100 samples from

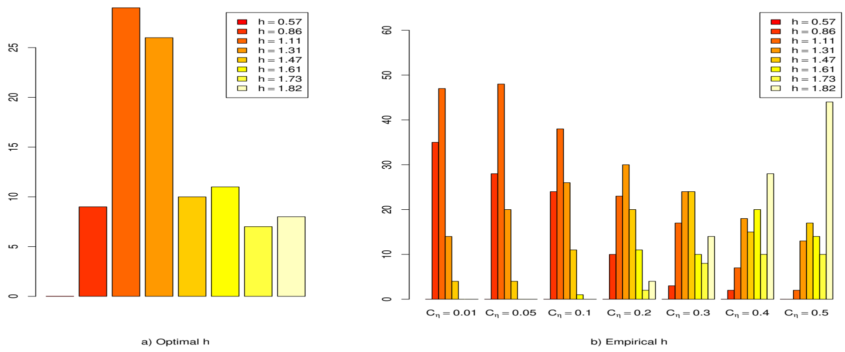

Figure 2: Simulation results; skew density, $n = 1000$.

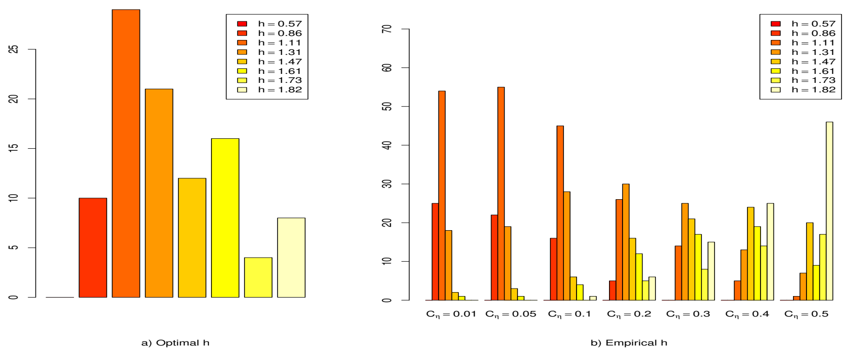

Figure 3: Simulation results; skew density, $n = 500$.

f. We have counted only the cases where the kernel estimate was unimodal. In particular, Figure 2 a) shows that the smoothing parameter $h = 1.11$ gave the best results: for 29 samples (from 100) this smoothing parameter resulted in an estimate whose mode was closest to the true mode. The average Euclidean distance of the modes of the estimates from the true mode was 0.35, when $h = 1.11$. For $h = 0.57$ all the 100 estimates were multimodal.

Figure 2 b) shows the number of times each smoothing parameter h was chosen by the method, for the choices of the factor $C_\eta = 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5$. In particular, Figure 2 b) shows that the choice $C_\eta = 0.05$ gave the best results: the optimal smoothing parameter $h = 1.11$ was chosen 48 times. Figure 2 b) shows also that the procedure is robust with respect to the choice of the parameter C_η: the results were not significantly worse when $C_\eta = 0.01$ and $C_\eta = 0.1$. The choices $C_\eta = 0.3, 0.4, 0.5$ were clearly worse. Choosing such a large factor C_η will lead to the choice of too large a smoothing parameter h.

Figure 3 shows otherwise the same situation as in Figure 2 but now we have sample size $n = 500$. Figure 3 shows that changing the sample size does not change the basic properties of the procedure.

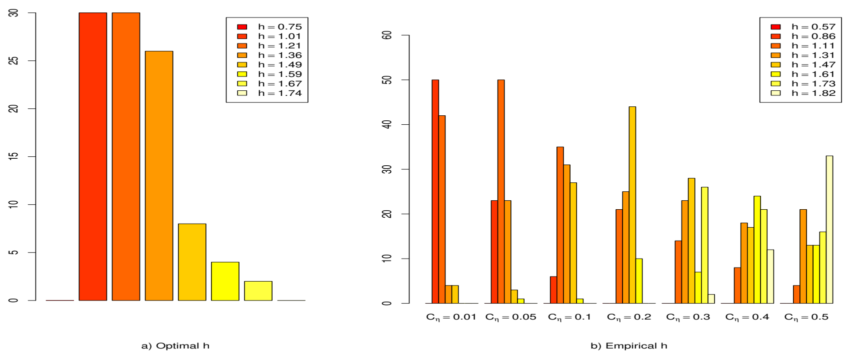

Figure 4: Simulation results; truncation of the skew density, $n = 200$.

Truncated skew density. The density of the second simulation example is shown in Figure 1 b). This is the truncation $cfI_{(-\infty,6.5]\times[-1,\infty)}$ of the skew density f shown in Figure 1 a), where c is the normalization constant and I_A is the indicator of set A: $I_A(x) = 1$ when $x \in A$ and 0 otherwise. The location of the mode is again approximately $(4.89, 0)$.

We generated 100 samples of size $n = 200$. We considered the grid of smoothing parameters $h \in \{0.75, 1.01, 1.21, 1.36, 1.49, 1.59, 1.67, 1.74\}$, which corresponds to grid $s \in \{0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0\}$ of smoothness indices when we choose $h = 2.5(\log_e(n)/n)^{1/(2s+2)}$. We applied the Bartlett-Epanechnikov product kernel.

Figure 4 a) shows the number of times each smoothing parameter was optimal. In particular, Figure 4 a) shows that the smoothing parameters $h = 1.01$ and $h = 1.21$ gave the best results: each of these smoothing parameters was optimal for 30 samples (from 100). The average Euclidean distance of the modes of the estimates from the true mode was 0.25 when $h = 1.21$, and 0.35 when $h = 1.36$. However, for $h = 1.21$ the estimate was unimodal only 51 times. For $h = 0.75$ all the 100 estimates were multimodal.

Figure 4 b) shows the number of times each smoothing parameter h was chosen by the method, for the choices of the factor $C_\eta = 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5$. In particular, Figure 4 b) shows that the choices $C_\eta = 0.01$ and $C_\eta = 0.05$ gave the best results: smoothing parameter $h = 1.01$ was chosen 50 times when $C_\eta = 0.01$ and smoothing parameter $h = 1.21$ was chosen 50 times when $C_\eta = 0.05$. The choices $C_\eta = 0.3, 0.4, 0.5$ were clearly worse. Again, the procedure is robust with respect to the choice of C_η.

Plug-in method. We record results also for a plug-in smoothing parameter selection method. The mean integrated squared error of the kernel estimator with a symmetric 2nd order kernel is $MISE(h) = \frac{1}{4}h^4\mu_2(K)\mathcal{D}(f) + (nh^2)^{-1}\|K\|_2^2$, where $\mu_2(K) = \int_{\mathbf{R}^2} x_1^2 K(x)dx$ and $\mathcal{D}(f) = \|D_1^2 f\|_2^2 + \|D_2^2 f\|_2^2 + 2\int_{\mathbf{R}^d} D_1^2 f D_2^2 f$.

See for example Wand and Jones (1994). The minimizer over $h > 0$ is $h(f) = [2\|K\|_2^2/(n\mu_2(K)\mathcal{D}(f))]^{1/(d+4)}$. We estimate $\mathcal{D}(f)$ and plug the estimate in the formula of $h(f)$. When f is the 2-dimensional Gaussian density with marginal standard deviations σ_1 and σ_2, and with covariance $\rho\sigma_1\sigma_2$, then $\|D_1^2 f\|_2^2 = 3/(2^6\pi\sigma_1^5\sigma_2(1-\rho^2)^{5/2})$ and $\int_{\mathbf{R}^d} D_1^2 f D_2^2 f = (1+2\rho^2)/(2^4\pi\sigma_1^3\sigma_2^3(1-\rho^2)^{5/2})$. We estimate σ_1, σ_2, and ρ with the sample moments and plug these estimates in the formula of $\mathcal{D}(f)$. This kind of plug-in method was studied in the univariate case by Park and Marron (1990) and Sheather and Jones (1991), for various scale estimators, and in the multivariate case by Wand and Jones (1994). We apply the Bartlett-Epanechnikov product kernel for which $\|K\|_2^2 = (3/4)^4(16/15)^2$ and $\mu_2(K) = (3/4)^2 4/15$.

For the skew density, with sample size $n = 1000$, the mean of the plug-in smoothing parameters was 0.94 and the standard deviation was 0.015, for a sequence of 1000 samples. With sample size $n = 500$, the mean of the plug-in smoothing parameters was 1.05 and the standard deviation was 0.024. Comparing these results to the optimal values in Figure 2 a) and Figure 3 a) we see that the plug-in method gave good results.

For the truncated skew density, with sample size $n = 200$, the mean of the plug-in smoothing parameters was 1.01 and the standard deviation was 0.047, for a sequence of 1000 samples. Comparing this result to the optimal values in Figure 4 a) indicates that the plug-in rule might give too small values, as is expected for the truncated density, since the plug-in method is based on the global smoothness assumptions which are not satisfied for the truncated density.

4 Proofs of the results

In the following we denote with $C, C', C_1, C_1', C_2, \ldots$ positive constants that can depend only on $L^*, f^*, f^{**}, \gamma^*, C^*, \Delta, d, p$. These constants may be different on different occasions.

4.1 Preliminary Lemmas

Let us prove that $D_i D_j f_{s,n}$ is a consistent estimator of $D_i D_j f$.

Lemma 3 *Let $s \geq s' \geq s_*$. Then,*

$$\lim_{n\to\infty} \sup_{f\in\mathcal{F}_s} \max_{i,j=1,\ldots,d} \sup_{x\in U} |E_f D_i D_j f_{s',n}(x) - D_i D_j f(x)| = 0.$$

Proof. Denote $h = h(s', \tilde{n}(s'))$. Let $\delta > 0$. We have that

$$\begin{aligned} &|E_f D_i D_j f_{s',n}(x) - D_i D_j f(x)| \\ &= \left| \int_{\mathbf{R}^d} (D_i D_j K_{s',h})(x-y) f(y) dy - D_i D_j f(x) \right| \end{aligned}$$

$$
\begin{aligned}
&= \left| \int_{\mathbf{R}^d} K_{s',h}(x-y) D_i D_j f(y) dy - D_i D_j f(x) \right| \\
&= \left| \int_{\mathbf{R}^d} K_{s',h}(x-y) (D_i D_j f(y) - D_i D_j f(x)) dy \right| \\
&\leq \sup_{y:\|y-x\|\leq\delta} |D_i D_j f(y) - D_i D_j f(x)| \int_{\mathbf{R}^d} |K_{s'}| + 2f^{**} \int_{y:\|y-x\|>\delta} K_{s',h}(x-y) dy
\end{aligned}
$$

where we used Condition 1.1 and Assumption 1.5. The lemma follows from Assumption 1.4 and Condition 1.1. □

Let us prove that the estimation error of the estimator $\hat{x}_s$ can be approximated by the estimation error of the first partial derivatives of $f_{s,n}$ at the point x_f. We may consider $Df_{s,n}(x_f)$ as an estimator for $Df(x_f) = 0$.

Lemma 4 *For $s \geq s' \geq s_*$, $f \in \mathcal{F}_s$,*

$$
P_f \left(\|\hat{x}_{s'} - x_f\|_{max} > 2C^* \sum_{i=1}^d |D_i f_{s',n}(x_f)| \right) \leq C_1 \exp\left\{ -C_2 n^{(2s_*-4)/(2s_*+d)} \right\}
$$

where C^* is from Assumption 1.3.

Proof. We have that

$$
0 = Df_{s',n}(\hat{x}_{s'}) = Df_{s',n}(x_f) + D^2 f_{s',n}(\theta_{s',n,f})(\hat{x}_{s'} - x_f)
$$

where $\theta_{s',n,f}$ is between x_f and $\hat{x}_{s'}$. Thus,

$$
\hat{x}_{s'} - x_f = -\left[D^2 f_{s',n}(\theta_{s',n,f}) \right]^{-1} Df_{s',n}(x_f). \tag{16}
$$

Note that by Assumption 1.2 and Condition 1.2, $\hat{x}_{s'} \in U$ and thus, by convexity of U, $\theta_{s',n,f} \in U$ for sufficiently large n. By Assumption 1.3, if

$$
\max_{i,j=1,\ldots,d} \sup_{x\in U} \left| \left(\left[D^2 f_{s',n}(x) \right]^{-1} \right)_{i,j} - \left(\left[D^2 f(x) \right]^{-1} \right)_{i,j} \right| \leq C^*
$$

then, because $\theta_{s',n,f} \in U$,

$$
\max_{i,j=1,\ldots,d} \left| \left(\left[D^2 f_{s',n}(\theta_{s',n,f}) \right]^{-1} \right)_{i,j} \right| \leq 2C^*. \tag{17}
$$

Equation (16) implies, that when (17) holds, then

$$
\|\hat{x}_{s'} - x_f\|_{max} \leq 2C^* \sum_{i=1}^d |D_i f_{s',n}(x_f)| \, .
$$

Thus,

$$
\begin{aligned}
& P_f\left(\|\hat{x}_{s'} - x_f\|_{max} > 2C^* \sum_{i=1}^{d} |D_i f_{s',n}(x_f)|\right) \\
\leq \; & P_f\left(\max_{i,j=1,\ldots,d} \sup_{x\in U} \left| \left(\left[D^2 f_{s',n}(x)\right]^{-1}\right)_{i,j} - \left(\left[D^2 f(x)\right]^{-1}\right)_{i,j} \right| > C^*\right).
\end{aligned}
$$

We have a further upper bound for the previous probability,

$$
\begin{aligned}
& P_f\left(\max_{i,j=1,\ldots,d} \sup_{x\in U} |D_iD_j f_{s',n}(x) - D_iD_j f(x)| > \delta(C^*)\right) \\
\leq \; & P_f\left(\max_{i,j=1,\ldots,d} \sup_{x\in U} |D_iD_j f_{s',n}(x) - E_f D_iD_j f_{s',n}(x)| > \delta(C^*)/2\right) \qquad (18)
\end{aligned}
$$

for some $\delta(C^*) > 0$, where we also used Lemma 3. Next we apply the following lemma, which follows from Lemma 5.8, Lemma 5.10, and Theorem 5.11 in van de Geer (2000).

Lemma 5 *Let $\mathcal{G} \subset L_2(P_f)$, let for $g \in \mathcal{G}$, $\int g^2 dP_f \leq R^2$ and $\|g\|_\infty \leq M$. Let $H_B(\epsilon, \mathcal{G}, P_f)$ be the $L_2(P_f)$ ϵ-entropy with bracketing. For C_0', C_1' satisfying $(C_0')^2 \geq (C')^2(C_1'+1)$ and for a satisfying*

$$
\frac{\sqrt{2}C_0'}{\sqrt{n}} \max\left\{\int_{a/(2^6\sqrt{2})}^{R} H_B^{1/2}(\epsilon, \mathcal{G}, P_f)\, d\epsilon, R\right\} \leq a \leq \min\left\{8\sqrt{2}R, C_1' R^2/M\right\},
$$

we have

$$
P_f\left(\sup_{g\in\mathcal{G}} \left|\frac{1}{n}\sum_{i=1}^{n} [g(X_i) - E_f g(X_i)]\right| \geq a\right) \leq 8\exp\left\{\frac{-na^2}{(C')^2(C_1'+1)R^2}\right\}.
$$

We apply Lemma 5 for $\mathcal{G} = \{g_x : x \in U\}$ where U is as in Assumption 1.2 and

$$
g_x(y) = h^{-d-2} D_iD_j K_{s'}((x-y)/h), \;\; y \in \mathbf{R}^d,
$$

where we denoted $h = h(s', \tilde{n}(s'))$. Then,

$$
\sup_{x\in A} |D_iD_j f_{s',n}(x) - E_f D_iD_j f_{s',n}(x)| = \sup_{g\in\mathcal{G}} \left|\frac{1}{n}\sum_{i=1}^{n} [g(X_i) - E_f g(X_i)]\right|.
$$

We may take $M = h^{-d-2}\|D_iD_jK_{s'}\|_\infty$ and $R^2 = h^{-d-4} f^* \|D_iD_jK_{s'}\|_2^2$, by Condition 1.7, where f^* is from Assumption 1.1. We have, for $x_1, x_2 \in U$, using Condition 1.8,

$$
\begin{aligned}
& \int (g_{x_1} - g_{x_2})^2 dP_f \\
= \; & h^{-d-4} \int_{\mathbf{R}^d} \left(D_iD_jK_{s'}\left(y + \frac{x_2 - x_1}{h}\right) - D_iD_jK_{s'}(y)\right)^2 f(x_1 - hy)\, dy \\
\leq \; & h^{-2\alpha-2d-4} \|x_2 - x_1\|^{2\alpha}.
\end{aligned}
$$

Thus,

$$H_B(\epsilon, \mathcal{G}, P_f) \le \log\left[C_3\left(\gamma^*(\epsilon h^{\alpha+d+2})^{-1/\alpha}\right)^d\right]$$

where γ^* is from Assumption 1.2. Thus,

$$\int_{a/(2^6\sqrt{2})}^{R} H_B^{1/2}(\epsilon, \mathcal{G}, P_f) d\epsilon \le R\sqrt{\log\left[C_3\left(\gamma^*(ah^{\alpha+d+2})^{-1/\alpha}\right)^d\right]}.$$

We have that $n/R^2 \ge Cn^{(2s_*-4)/(2s_*+d)}$, by Condition 1.5, and we may apply Lemma 5 with $a = \delta(C^*)/2$ to have the required upper bound for the right hand side of (18). □

Let us prove a uniform upper bound for the bias of $D_i f_{s,n}(x_f)$, considered as an estimator of $D_i f(x_f) = 0$.

Lemma 6 *For $s_* \le s' \le s$,*

$$\sup_{f\in\mathcal{F}_s} \|E_f D f_{s',n}(x_f)\|_{max} \le C_b h^{s-1}(s', \tilde{n}(s'))$$

where constant C_b was defined in (44).

Proof. For $f \in \mathcal{F}_s$ and $i \in \{1, \ldots, d\}$, denoting $h = h(s', \tilde{n}(s'))$, and using Conditions 1.3, 1.4 on the kernel, and the fact that $D_i f(x_f) = 0$,

$$\begin{aligned}
|E_f D_i f_{s',n}(x_f)| &= \left|\int (D_i K_{s',h})(x_f - y) f(y) dy\right| = \left|\frac{1}{h}\int (D_i K_{s'})(z) f(x_f - hz) dz\right| \\
&\le L^* h^{s-1} \int \|z\|^s \left|(D_i K_{s'})(z)\right| dz
\end{aligned}$$

because, by the definition of $\mathcal{F}_s$,

$$f(x_f - hz) = \sum_{i=0}^{\lceil s-1\rceil} \sum_{|\alpha|=i} \frac{(hz)^\alpha}{\alpha!} \left.\frac{\partial^i f(y)}{\partial^\alpha y}\right|_{y=x_f} + R(hz)$$

where $|R(z)| \le L^*\|z\|^s$. □

We will need the following upper bound for the central moments of $D_i f_{s,n}(x_f)$.

Lemma 7 *Let us denote for $i = 1, \ldots, d$,*

$$Z_{s,i} = D_i f_{s,n}(x_f) - E_f D_i f_{s,n}(x_f). \tag{19}$$

Then, for sufficiently large n, for $s \ge s' \ge s_*$, $f \in \mathcal{F}_s$, $p \ge 1$,

$$E_f\left|Z_{s',i}\right|^p \le C\left(n^{-1/2} h^{-1-d/2}(s', \tilde{n}(s'))\right)^p$$

where $h(s', n)$ was defined in (12) and $\tilde{n}(s')$ in (10).

Proof. Denote in the following $h = h(s', \tilde{n}(s'))$. We have

$$Z_{s',i} = \frac{1}{n}\sum_{j=1}^{n}(Y_j - E_f Y_j)$$

where $Y_j = D_i K_{s',h}(x_f - X_j)$. By Rosenthal's inequality,

$$E_f\,|Z_{s',i}|^{2p} \le C_1\left[n^{1-2p}E_f|Y_1 - E_fY_1|^{2p} + n^{-p}(\mathrm{Var}_f(Y_1))^p\right], \tag{20}$$

see Theorem 2.9 in Petrov (1995). Now, for $m \ge 1$,

$$\begin{aligned} E_f Y_1^m &= \int_{\mathbf{R}^d}\left[D_i K_{s',h}(x_f - y)\right]^m f(y)dy \\ &= h^{-m(1+d)}\int_{\mathbf{R}^d}\left[(D_j K_{s'})((x_f - y)/h)\right]^m f(y)dy \\ &\sim h^{-m(1+d)+d} f(x_f)\int_{\mathbf{R}^d}\left[(D_j K_{s'})(y)\right]^m dy \end{aligned} \tag{21}$$

by continuity of f at x_f and by Condition 1.5. Then, by (21), $n^{-p}(\mathrm{Var}_f(Y_1))^p \le n^{-p}(E_f Y_1^2)^p \le C_2(n^{-1}h^{-2-d})^p$. Also by (21),

$$E_f|Y_1 - E_fY_1|^{2p} \le 2^{2p-1}[E_f|Y_1|^{2p} + |E_fY_1|^{2p}] \le C_3 h^{-2p(1+d)+d}.$$

Thus, $n^{1-2p}E_f|Y_1 - E_fY_1|^{2p} \le C_3(nh^d)^{1-p}(n^{-1}h^{-2-d})^p$. Finally we have $E_f\,|Z_{s',i}|^p \le (E_f\,|Z_{s',i}|^{2p})^{1/2}$ and thus the lemma follows. □

We will need the following technical lemma.

Lemma 8 *Let ψ_s be defined in (8), $h(s,n)$ be defined in (12), and $\tilde{n}(s)$ be defined in (10). Then,*

$$\frac{\psi_{s'}}{\psi_s} = \exp\left\{\frac{(d+2)(s-s')}{2(2s+d)(2s'+d)}\left(\log n - \log\log n^{1/2}\right)\right\} \tag{22}$$

and

$$\frac{h(s,\tilde{n}(s))}{h(s',\tilde{n}(s'))} = \exp\left\{\frac{4(s-s')}{2(2s+d)(2s'+d)}\left(\log n - \log\log n^{1/2}\right)\right\}. \tag{23}$$

Proof. The equations follow directly from the definitions. □

Let us prove that the probability of underestimating the value of s by the statistic $\hat{s}$ is small, uniformly over $f \in \mathcal{F}_s$.

Lemma 9 *Let $s \in [s_*, \infty)$. Then, for $s' \in S$, $s' < s^-$, where $s^- = s^-(s)$ is the largest member of the grid $\le s$,*

$$\sup_{f\in\mathcal{F}_s} P_f\,(\hat{s} = s') \le C\,\mathrm{Card}(S)\,n^{-p/2}$$

for sufficiently large n.

Proof. We have

$$\begin{aligned}\sup_{f\in\mathcal{F}_s} P_f(\hat{s}=s') &\le \sum_{s''\in S, s''\le s'} \sup_{f\in\mathcal{F}_s} P_f\left(\|\hat{x}_{\bar{s'}}-\hat{x}_{s''}\|_{max}>\eta(s'')\right)\\ &\le \operatorname{Card}(S) \max_{s''\in S, s''\le s'} \sup_{f\in\mathcal{F}_s} P_f\left(\|\hat{x}_{\bar{s'}}-\hat{x}_{s''}\|_{max}>\eta(s'')\right) \qquad (24)\end{aligned}$$

where $\bar{s'}=\bar{s'}(s')$ is the smallest element of S greater than s'. Let $f\in\mathcal{F}_s$. Then,

$$\begin{aligned}&P_f\left(\|\hat{x}_{\bar{s'}}-\hat{x}_{s''}\|_{max}>\eta(s'')\right)\\ &\le P_f\left(\|\hat{x}_{s''}-x_f\|_{max}>\eta(s'')/2\right)+P_f\left(\|\hat{x}_{\bar{s'}}-x_f\|_{max}>\eta(s'')/2\right)\\ &\overset{def}{=} p_1+p_2. \qquad (25)\end{aligned}$$

Now,

$$\begin{aligned}p_1 &\le P_f\left(2C^*\sum_{i=1}^{d}|D_i f_{s'',n}(x_f)|>\eta(s'')/2\right)\\ &\quad +P_f\left(\|\hat{x}_{s''}-x_f\|_{max}>2C^*\sum_{i=1}^{d}|D_i f_{s'',n}(x_f)|\right)\overset{def}{=}p_{11}+p_{12}. \qquad (26)\end{aligned}$$

By Lemma 4, $p_{12}\le Cn^{-p/2}$, and it is left to consider only p_{11}. For sufficiently large n, using Lemma 6, (23), and the fact that $s''\le s$ we get for $j=1,\dots,d$,

$$|E_f D_j f_{s'',n}(x_f)|\le C_b h^{s-1}(s'',\tilde{n}(s''))\le C_b h^{s-1}(s,\tilde{n}(s)). \qquad (27)$$

By definitions (8), (10) and (12) and equation (22) we get

$$h^{s-1}(s,\tilde{n}(s))=\psi_s\le\psi_{s''}. \qquad (28)$$

Thus,

$$\sum_{i=1}^{d}|D_i f_{s'',n}(x_f)|\le dC_b\psi_{s''}+\sum_{j=1}^{d}|Z_{s'',j}|$$

where $Z_{s'',j}$ is defined in (19). Thus, for sufficiently large n, using the fact following from definitions that for $s<s_n$, $\eta(s)=C_\eta\psi_s$,

$$\begin{aligned}p_{11} &\le P_f\left(\sum_{j=1}^{d}|Z_{s'',j}|>\left(C_\eta/(4C^*)-dC_b\right)\psi_{s''}\right)\\ &\le \sum_{j=1}^{d}P_f\left(|Z_{s'',j}|>\left(C_\eta/(4dC^*)-C_b\right)\psi_{s''}\right). \qquad (29)\end{aligned}$$

We will have to choose C_η so large that $C_\eta/(4C^*)-dC_b>0$ (see Equation (43)). Now, for $j=1,\dots,d$,

$$Z_{s'',j}=\frac{1}{n}\sum_{i=1}^{n}(Y_i-E_fY_i)$$

where $Y_i = D_j K_{s'',h(s'',\tilde{n}(s''))}(x_f - X_i)$. Now, for $m \geq 2$, using (21) and Condition 1.5,

$$\begin{aligned} |E_f(Y_1 - E_f Y_1)^m| &\leq 2^{2m-1}\left(E_f|Y_1|^m + |E_f Y_1|^m\right) \\ &\sim 2^{2m-1} h(s'',\tilde{n}(s''))^{-m(1+d)+d} f(x_f) \|D_j K_{s''}\|_m^m \\ &\leq 2^{2m-1} h(s'',\tilde{n}(s''))^{-m(1+d)+d} f(x_f)(K_m^*)^m. \end{aligned}$$

where $K_m^* = \max_{j=1,\ldots,d} \sup_{s \in [s_*,\infty)} \|D_j K_s\|_m$. Also, by (21) and Condition 1.6,

$$\begin{aligned} \mathrm{Var}_f(Y_1) &= E_f Y_1^2 - (E_f Y_1)^2 \sim h(s'',\tilde{n}(s''))^{-2(1+d)+d} f(x_f)\|D_j K_{s''}\|_2^2 \\ &\geq h(s'',\tilde{n}(s''))^{-2-d} f(x_f)(K_*)^2 \end{aligned}$$

where $K_* = \min_{j=1,\ldots,d} \inf_{s\in[s^*,\infty)} \|D_j K_s\|_2$. Choose H such that,

$$|E_f(Y_1 - E_f Y_1)^m| \leq \frac{1}{2} m! \,\mathrm{Var}_f(Y_1) H^{m-2}.$$

Now H can be chosen as $H = C_1 h(s'',\tilde{n}(s''))^{-1-d}$ for some positive constant C_1 and $n\mathrm{Var}_f(Y_1)/H \geq C_2 n h(s'',\tilde{n}(s''))^{-1} \to \infty$ as $n \to \infty$. Thus we can apply Bernstein's inequality as given by Theorem 2.8 in Petrov (1995). In fact, denote $a = (C_\eta/(4dC^*) - C_b)\,\psi_{s''}$. Then,

$$\begin{aligned} P_f\left(|Z_{s'',j})| > a\right) &\leq 2\exp\left\{\frac{-na^2}{4\mathrm{Var}_f(Y_1)}\right\} \leq 2\exp\left\{\frac{-na^2}{4R^2}\right\} \\ &\leq 2\exp\left\{-(p/2)\log n\right\} \end{aligned} \tag{30}$$

where $\mathrm{Var}_f(Y_1) \leq R^2 \stackrel{def}{=} h(s'',\tilde{n}(s''))^{-2-d} f^*(K_2^*)^2$. We used also the fact that, $na^2/4R^2 \geq (p/2)\log n$ which follows from the definition of C_η (Equation (43)) and from the fact that $\psi_s = \tilde{n}(s)^{-1/2} h^{-1-d/2}(s,\tilde{n}(s))$. Equations (26), (29), and (30) imply that $p_1 \leq C_3 n^{-p/2}$. The upper bound for p_2 follows similarly, with $\bar{s'}$ in the place of s'', using inequality

$$h^{-1-d/2}(\bar{s'},\tilde{n}(\bar{s'})) \leq h^{-1-d/2}(s'',\tilde{n}(s''))$$

which follows from equation (23). Equations (24) and (25) then imply the lemma. □

4.2 Proof of Theorem 2

Consider first $s \in [s_*, s_n)$ and let $s^- = s^-(s)$ be the largest gridpoint $\leq s$. We have

$$\mathcal{R}_{n,s}(x_n^*,\psi_s) = \sup_{f\in\mathcal{F}_s} E_f\left(\psi_s^{-p}\|x_n^* - x_f\|^p\right) = R_{n,s}^- + R_{n,s}^+$$

where

$$R_{n,s}^- = \sup_{f\in\mathcal{F}_s} E_f\left(\psi_s^{-p}\|x_n^* - x_f\|^p I\left(\hat{s} < s^-\right)\right), \tag{31}$$

$$R_{n,s}^+ = \sup_{f\in\mathcal{F}_s} E_f\left(\psi_s^{-p}\|x_n^* - x_f\|^p I\left(\hat{s} \geq s^-\right)\right). \tag{32}$$

Lemma 10 *For $R^-_{n,s}$ defined in (31),* $\lim_{n\to\infty}\sup_{s\in[s_*,s_n)} R^-_{n,s} = 0$.

Proof. Let $s \in [s_*, s_n)$, $s' \in S$, $s' < s^-$, and $f \in \mathcal{F}_s$. We have by Assumption 1.2 and by Condition 1.2, that $x_f, \hat{x}_{s'} \in U$ for sufficiently large n. Thus $\|\hat{x}_{s'} - x_f\|$ is bounded,

$$\|\hat{x}_{s'} - x_f\|^{2p} \le (\gamma^*)^{2p} \tag{33}$$

where γ^* is from Assumption 1.2. Then, by Equation (33) and Lemma 9,

$$\begin{aligned} R^-_{n,s} &\le \sum_{s'\in S, s'<s^-} \sup_{f\in\mathcal{F}_s} E_f\left(\psi_s^{-p}\|\hat{x}_{s'} - x_f\|^p I(\hat{s} = s')\right) \\ &\le C_1\psi_s^{-p} \sum_{s'\in S, s'<s^-} \sup_{f\in\mathcal{F}_s} P_f(\hat{s} = s') \\ &\le C_2\psi_s^{-p} n^{-p/2}\left[\mathrm{Card}(S)\right]^2. \end{aligned}$$

We have from (7) and (9) that $\mathrm{Card}(S) \le (\log n)^{a+a_1}\kappa/\kappa_1$, and the lemma follows. □

Lemma 11 *For $R^+_{n,s}$ defined in (31),* $\lim_{n\to\infty}\sup_{s\in[s_*,s_n)} R^+_{n,s} < \infty$.

Proof. Let $s \in [s_*, s_n)$ and $f \in \mathcal{F}_s$. Definitions, application of (22), and (9) yield

$$\eta(s^-) = C_\eta \psi_{s^-} \le C_1\psi_s. \tag{34}$$

Let $\hat{s} = s' \ge s^-$. Then, using the definition of $\hat{s}$ and (34),

$$\begin{aligned} \|x_n^* - x_f\| &= \|\hat{x}_{s'} - x_f\| \le \|\hat{x}_{s'} - \hat{x}_{s^-}\| + \|\hat{x}_{s^-} - x_f\| \\ &\le \eta(s^-) + \|\hat{x}_{s^-} - x_f\| \le C_1\psi_s + \|\hat{x}_{s^-} - x_f\|. \end{aligned} \tag{35}$$

Because $\|\hat{x}_{s^-} - x_f\|$ is bounded by Assumption 1.2 and Condition 1.2, then, denoting

$$\mathcal{A} = \left\{\|\hat{x}_{s^-} - x_f\| > 2C^* \sum_{i=1}^d |D_i f_{s^-}(x_f)|\right\}, \tag{36}$$

we have

$$E_f\left(\psi_s^{-p}\|\hat{x}_{s^-} - x_f\|^p\right) \le E_f\left(\psi_s^{-p}\left(2C^*\sum_{i=1}^d |D_i f_{s^-}(x_f)|\right)^p\right) + C_2\psi_s^{-p} P_f(\mathcal{A}). \tag{37}$$

Similarly as in (27) and in (28), because $s^- \le s$, for $j = 1, \ldots, d$,

$$|E_f D_j f_{s^-,n}(x_f)| \le C\psi_s.$$

Then

$$\sum_{i=1}^d |D_i f_{s^-}(x_f)| \le \sum_{i=1}^d \left(|E_f D_i f_{s^-}(x_f)| + |Z_{s^-,i}|\right) \le C_3\psi_s + \sum_{i=1}^d |Z_{s^-,i}|. \tag{38}$$

By (35), (37), and (38),

$$
\begin{aligned}
& E_f\left(\psi_s^{-p}\left\|x_n^*-x_f\right\|^p I\left(\hat{s} \geq s^-\right)\right) \\
\leq\ & C_4\left(1+\psi_s^{-p} \sum_{i=1}^d E_f\left|Z_{s^-, i}\right|^p\right)+C_2 \psi_s^{-p} P_f(\mathcal{A}) \\
\leq\ & C_5\left(1+\psi_s^{-p}\left(n^{-1/2} h^{-1-d/2}\left(s^-, \tilde{n}\left(s^-\right)\right)\right)^p\right)+C_2 \psi_s^{-p} P_f(\mathcal{A}) \qquad (39)
\end{aligned}
$$

where we used also Lemma 7. By Lemma 4,

$$
\lim_{n \rightarrow \infty} \sup_{s \in\left[s_*, s_n\right)} \psi_s^{-p} \sup_{f \in \mathcal{F}_s} P_f(\mathcal{A})=0. \tag{40}
$$

Definition (8) entails that $\psi_s=\tilde{n}(s)^{-1/2} h^{-1-d/2}(s, \tilde{n}(s))$ and thus

$$
\psi_{s^-}^{-1} n^{-1/2} h^{-1-d/2}\left(s^-, \tilde{n}\left(s^-\right)\right)=(\log n)^{-1/2}. \tag{41}
$$

Equations (34) and (41) entail that

$$
\psi_s^{-1} n^{-1/2} h^{-1-d/2}\left(s^-, \tilde{n}\left(s^-\right)\right) \leq C_6(\log n)^{-1/2} \longrightarrow 0, \quad \text{as } n \longrightarrow \infty. \tag{42}
$$

It remains to note that the lemma follows from (39), (40), and (42). □

It is left to consider the case $s=s_n$.

Lemma 12 $\limsup_{n \rightarrow \infty} \mathcal{R}_{n, s_n}\left(x_n^*, \psi_{s_n}\right)<\infty$.

Proof. Similarly as in the proof of Lemma 10 we can prove using Lemma 9 that

$$
\lim_{n \rightarrow \infty} \sup_{f \in \mathcal{F}_{s_n}} E_f\left(\psi_{s_n}^{-p}\left\|x_n^*-x_f\right\|^p I\left(\hat{s}<s_n\right)\right)=0.
$$

On the other hand,

$$
\limsup_{n \rightarrow \infty} \sup_{f \in \mathcal{F}_{s_n}} E_f\left(\psi_{s_n}^{-p}\left\|x_n^*-x_f\right\|^p I\left(\hat{s}=s_n\right)\right)<\infty.
$$

Indeed, similarly as in (37),

$$
E_f\left(\psi_{s_n}^{-p}\left\|\hat{x}_{s_n}-x_f\right\|^p\right) \leq E_f\left(\psi_{s_n}^{-p}\left(2 C^* \sum_{i=1}^d\left|D_i f_{s_n}\left(x_f\right)\right|\right)^p\right)+C_1 \psi_{s_n}^{-p} P_f(\mathcal{A}).
$$

where $\mathcal{A}$ is as in (36) but s^- is replaced with s_n. From Lemma 6 and definitions, for $f \in \mathcal{F}_{s_n}$, $i=1, \ldots d$,

$$
\left|E_f D_i f_{s_n, n}\left(x_f\right)\right| \leq C_2 h^{s_n-1}\left(s_n, \tilde{n}\left(s_n\right)\right)=C_2 n^{-\left(s_n-1\right)/\left(2 s_n+d\right)}=C_2 \psi_{s_n}.
$$

Thus,

$$
\sum_{i=1}^d\left|D_i f_{s_n}\left(x_f\right)\right| \leq \sum_{i=1}^d\left(\left|E_f D_i f_{s_n}\left(x_f\right)\right|+\left|Z_{s_n, i}\right|\right) \leq C_3 \psi_{s_n}+\sum_{i=1}^d\left|Z_{s_n, i}\right|.
$$

We have

$$E_f\left(\psi_{s_n}^{-p}\left\|\hat{x}_{s_n}-x_f\right\|^p\right)\leq C_4\left(1+\psi_{s_n}^{-p}\sum_{i=1}^{d}E_f\left|Z_{s_n,i}\right|^p\right)+C_1\psi_{s_n}^{-p}P_f(\mathcal{A})$$

By Lemma 7,

$$E_f\left|Z_{s_n,i}\right|^p\leq C_5\left(n^{-1/2}h^{-1-d/2}(s_n,\tilde{n}(s_n))\right)^p=C_5\psi_{s_n}^p.$$

From Lemma 4,

$$\limsup_{n\to\infty}\sup_{f\in\mathcal{F}_{s_n}}\psi_{s_n}^{-p}P_f(\mathcal{A})=0,$$

and the lemma follows. □

Proof of Theorem 2. From Lemmas 10 and 11 we have

$$\limsup_{n\to\infty}\sup_{s\in[s_*,s_n)}\mathcal{R}_{n,s}(x_n^*,\psi_s)<\infty.$$

Combining this with Lemma 12 proves the theorem. □

Remark 2. It follows from the Proof of Theorem 2 that we may choose the constant C_η in (14) by

$$C_\eta\geq 4dC^*\left[C_b+K_2^*\sqrt{2pf^*}\right] \tag{43}$$

where

$$C_b=L^*\max_{i=1,\dots,d}\sup_{s_*\leq s'\leq s<\infty}\int\|z\|^s\left|D_iK_{s'}(z)\right|dz \tag{44}$$

and

$$K_2^*=\max_{i=1,\dots,d}\sup_{s\in[s_*,\infty)}\|D_iK_s\|_2.$$

4.3 Proof of Theorem 1

Part 1 of Definition 1 with the normalising factor $\psi_s=\psi_{s,n}$, defined in (8), follows from Theorem 2.

Let us then prove part 2 of Definition 1. Let $S_{s,n}$ satisfy (4) and (5) for some estimator x_n^{**}, with $\Psi_{s,n}=\psi_{s,n}$. Because $n^{-(s'-1)/(2s'+d)}$ is the minimax rate of convergence in the sense of (2), (4) implies that there exists a positive constant $\bar{c}$ such that, for s' as in (5),

$$S_{s',n}\geq\bar{c}n^{-(s'-1)/(2s'+d)} \tag{45}$$

for n large enough. For the current notation, equation (5) reads

$$\lim_{n\to\infty}(S_{s',n}/\psi_{s',n})=0. \tag{46}$$

Note that s' in (5) satisfies $s_* \le s' < s_n$. This is due to the fact that for $s = s_n$, $\psi_{s,n}$ is the minimax rate of convergence, that is, it satisfies (2), and cannot be improved. To make notation similar to Definition 1, denote in the following $s'' = s_n$. Denote $r' = (s'-1)/(2s'+d)$ and let $r, \bar{r}$ be such that

$$0 < r' < \bar{r} < r < \frac{(s''-1)}{2s''+d}$$

and

$$\frac{(1/2)-\bar{r}}{(1/2)-r'} < \frac{1}{2}\,. \tag{47}$$

This choice is possible for sufficiently large n. First, if

$$\liminf_{n\to\infty} \left(S_{s'',n}/n^{-r} \right) = \infty$$

then, for sufficiently large n, using (45) and the fact that $r < (s''-1)/(2s''+d)$,

$$\left(S_{s',n}/\psi_{s',n}\right)\left(S_{s'',n}/\psi_{s'',n}\right) \ge \bar{c} S_{s'',n}/n^{-r} \longrightarrow \infty$$

as $n \to \infty$, that is, (6) holds. Second, if there exists a positive constant C_1 such that

$$\liminf_{n\to\infty} \left(S_{s'',n}/n^{-r} \right) \le C_1$$

then we will come to a contradiction. Indeed, then there exists a subsequence $n_m \to \infty$ such that

$$\limsup_{m\to\infty} \left(S_{s'',n_m}/n_m^{-r} \right) \le C_1. \tag{48}$$

Condition (4) from Definition 1 and (48) imply that

$$C' \ge \tag{49}$$
$$\limsup_{m\to\infty} \min\left\{ a_{n_m}, (n_m^{r-\bar{r}}/C_1)^p \right\} \inf_{x_{n_m}} \max\left\{ \mathcal{R}_{n_m,s'}(x_{n_m}, \psi_{s',n_n}), \mathcal{R}_{n_m,s''}(x_{n_m}, n_m^{-\bar{r}}) \right\}$$

where

$$a_{n_m} = \psi^p_{s',n_m} S^{-p}_{s',n_m} \longrightarrow \infty$$

by (46). We will show that

$$\inf_{x_{n_m}} \max\left\{ \mathcal{R}_{n_m,s'}(x_{n_m}, \psi_{s',n_m}), \mathcal{R}_{n_m,s''}(x_{n_m}, n_m^{-\bar{r}}) \right\} \ge C_2 \tag{50}$$

for some positive constant C_2, which is a contradiction with (49). Our aim is to prove (50) applying the following lemma, which is a special case of Theorem 6 (i) in Tsybakov (1998).

Lemma 13 *Let P_0 and P_1 be probability measures in some measurable space. Let $Q > 0$, $\tau > 0$, $0 < \delta < 1/2$, $0 < l < 1$ be fixed and let $D : \mathbf{R}^d \times \mathbf{R}^d \to [0, \infty)$ be a distance such that,*

$$D(\mathbf{0}, \mathbf{1}) \geq 1 - \delta, \ \ \mathbf{0} = (0, \ldots, 0), \ \ \mathbf{1} = (1, 0, \ldots, 0)$$

and let $P_0 << P_1$ with

$$P_1 \left(\frac{dP_0}{dP_1} \geq \tau \right) \geq 1 - l.$$

Then,

$$\inf_{x_n} \max \left\{ Q^p E_0 D^p(x_n, \mathbf{0}), E_1 D^p(x_n, \mathbf{1}) \right\} \geq \frac{(1-l)(1-2\delta)^p \tau (Q\delta)^p}{(1-2\delta)^p + \tau (Q\delta)^p}$$

where the infimum is taken over all estimators.

Difficult subproblem. We will construct $f_0 \in \mathcal{F}_{s''}$ and $f_1 \in \mathcal{F}_{s'}$ which are sufficiently well separated. We will use a similar type of construction as Tsybakov (1990). Let $f_0(x) = C_0 a^{-d} (1 - \|x/a\|^2)_+$, where $C_0^{-1} = \int_{\mathbf{R}^d} (1 - \|x\|^2)_+ dx$, $(x)_+ = \max\{x, 0\}$, and $a > 0$ will be chosen later to be sufficiently small. Let $\eta_0 : \mathbf{R} \to [0, \infty)$ be an infinitely differentiable function satisfying

$$\eta_0(t) = \begin{cases} 1, & t \in [1/8, 3/8] \\ 0, & t \notin [0, 1/2]. \end{cases}$$

Let $\eta_1(t) = \eta_0(t + 1/2) - \eta_0(t)$ and $\eta(t) = \int_{-\infty}^t \eta_1(u) du$. Define $L : \mathbf{R}^d \to \mathbf{R}$ by $L(x) = \prod_{i=1}^d \eta(x_i)$ and let $f_1 = f_0 + T_n L$ where

$$(T_n L)(x) = (1 - \delta) \tilde{n}^{-1/2} h^{-d/2} \left[L((x - x_1)/h) - L((x - x_2)/h) \right]$$

where $0 < \delta < 1/2$, $h = h(s', \tilde{n}) = \tilde{n}^{-1/(2s'+d)}$, $\tilde{n} = n / \log n$, $x_1 = (3h/8, 0, \ldots, 0)$ and $x_2 = (a/2, 0, \ldots, 0)$. By construction,

$$f_0 \in \mathcal{F}_{s''}, \ \ f_1 \in \mathcal{F}_{s'} \tag{51}$$

where $s'' = s_n$. Furthermore, the mode of f_0 is located at 0 and the mode of f_1 is located at $((1 - \delta) \tilde{n}^{-1/2} h^{-d/2-1}, 0, \ldots, 0)$,

$$x_{f_0} = 0, \ \ x_{f_1} = ((1 - \delta) \tilde{n}^{-1/2} h^{-d/2-1}, 0, \ldots, 0) = ((1 - \delta) \psi_{s',n}, 0, \ldots, 0). \tag{52}$$

Weak convergence. Denote

$$Y_i = \log \left(\frac{f_0(X_i)}{f_1(X_i)} \right) = \log \left(1 - \frac{(T_n L)(X_i)}{f_1(X_i)} \right)$$

and

$$\tau = \exp\{-(1-\delta)p((1/2)-r')\log n\}$$

where $r' = (s'-1)/(2s'+d)$. Now

$$P_{f_1}^{(n)}\left(\frac{dP_{f_0}^{(n)}}{dP_{f_1}^{(n)}} \geq \tau\right) = P_{f_1}^{(n)}\left(\sum_{i=1}^{n} Y_i \geq \log\tau\right) \stackrel{def}{=} 1 - l_n.$$

We will show that

$$\lim_{n\to\infty} l_n = 0. \tag{53}$$

Denote in the following $E_j = E_{f_j}$, $\mathrm{Var}_j = \mathrm{Var}_{f_j}$, $j = 0, 1$. We will show that the weak convergence to the standard Gaussian distribution,

$$B_n^{-1}\sum_{i=1}^{n}(Y_i - E_1 Y_i) \longrightarrow N(0,1), \tag{54}$$

holds under measure P_{f_1}, where $B_n^2 = n\mathrm{Var}_1(Y_1)$. Indeed, there exists $\Delta > 0$ such that for $|t| < \Delta$,

$$-t - C_3 t^2 \leq \log(1-t) \leq -t \tag{55}$$

for some $C_3 > 0$ and then, because $\lim_{n\to\infty}\|(T_nL)/f_1\|_\infty = 0$,

$$\begin{aligned} E_1|Y_1|^q &\leq \int_{\mathbf{R}^d}\left(\frac{|(T_nL)(x)|^q}{f_1^{q-1}(x)}\right)dx \\ &\leq C_4(\tilde{n}h^d)^{-q/2}h^d(1-\delta)^q(f_1(x_1)^{1-q} + f_1(x_2)^{1-q})\int_{\mathbf{R}^d}|L|^q \end{aligned} \tag{56}$$

for $q \geq 1$. We have by (56) with $q = 3$,

$$\begin{aligned} &E_1|Y_1 - E_1Y_1|^3 \leq 8E_1|Y_1|^3 \\ &\leq C_5(\tilde{n}h^d)^{-3/2}h^d(1-\delta)^3(f_1(x_1)^{-2} + f_1(x_2)^{-2})\int_{\mathbf{R}^d}|L|^3. \end{aligned} \tag{57}$$

Also, by (55), by (56), and using Cauchy's inequality, for sufficiently large n,

$$\begin{aligned} E_1Y_1^2 &\geq \int_{\mathbf{R}^d}\left(\frac{(T_nL)(x)}{f_1(x)} + C_3\left(\frac{(T_nL)(x)}{f_1(x)}\right)^2\right)^2 f_1(x)dx \\ &\geq (1-\epsilon)\int_{\mathbf{R}^d}\frac{(T_nL)^2(x)}{f_1(x)}dx \\ &\geq (1-2\epsilon)(\tilde{n}h^d)^{-1}h^d(1-\delta)^2\left[f_1(x_1)^{-1/2} - f_1(x_2)^{-1/2}\right]^2\int_{\mathbf{R}^d}L^2 \end{aligned}$$

where $0 < \epsilon < 1/3$. Thus, for sufficiently large n,

$$\begin{aligned} &\mathrm{Var}_1(Y_1) = E_1Y_1^2 - (EY_1)^2 \\ &\geq (1-3\epsilon)(\tilde{n}h^d)^{-1}h^d(1-\delta)^2\left[f_1(x_1)^{-1/2} - f_1(x_2)^{-1/2}\right]^2\int_{\mathbf{R}^d}L^2 \end{aligned} \tag{58}$$

where we used also (56) with $q=1$. From (57) and (58) we have

$$\frac{nE_1|Y_1 - E_1Y_1|^3}{(n\mathrm{Var}_1(Y_1))^{3/2}} \le C_6(nh^d)^{-1/2} \longrightarrow 0$$

as $n\to\infty$. The weak convergence in (54) follows from the previous display by the Theorem of Lyapunov, see Theorem 4.9 in Petrov (1995). Then, from (55), from $\int_{\mathbf{R}^d} T_nL = 0$, and from (56) with $q=2$,

$$\begin{aligned} E_1Y_1 &\ge \int_{\mathbf{R}^d}\left(-\frac{(T_nL)(x)}{f_1(x)} - C_3\left(\frac{(T_nL)(x)}{f_1(x)}\right)^2\right) f_1(x)dx \\ &= \int_{\mathbf{R}^d} -C_3\,\frac{(T_nL)^2(x)}{f_1(x)}dx \\ &\ge -C_3(\tilde{n}h^d)^{-1}h^d(1-\delta)^2(f_1(x_1)^{-1} + f_1(x_2)^{-1})\int_{\mathbf{R}^d} L^2. \end{aligned} \tag{59}$$

Then, using (58) and (59),

$$\begin{aligned} &B_n^{-1}(\log\tau - nE_1Y_1) \\ &\le -\sqrt{\log n}\,\frac{(1-\delta)p((1/2)-r') - C_3(1-\delta)^2(f_1(x_1)^{-1}+f_1(x_2)^{-1})\int_{\mathbf{R}^d}L^2}{[(1-3\epsilon)(1-\delta)^2\left[f_1(x_1)^{-1/2} - f_1(x_2)^{-1/2}\right]^2\int_{\mathbf{R}^d}L^2]^{1/2}} \\ &\longrightarrow -\infty \end{aligned} \tag{60}$$

as $n\to\infty$, when $f_1(x_1)$ is sufficient large, which is achieved by choosing a in the definition of f_0 sufficiently small. Now (53) follows from (54) and (60).

Final step. By (52), for any estimator x_n,

$$\|x_n - x_{f_i}\| = \psi_{s',n}\, D\left((1-\delta)^{-1}\psi_{s',n}^{-1}x_n, i\right) \quad i = \mathbf{0},\mathbf{1},$$

where $D(u,v) = (1-\delta)\|u-v\|$, $u,v\in\mathbf{R}^d$. Denoting $Q = \psi_{s',n_m}/n_m^{-\bar{r}}$ we get by (51) and by applying Lemma 13,

$$\begin{aligned} &\inf_{x_{n_m}} \max\left\{\mathcal{R}_{n_m,s'}(x_{n_m},\psi_{s',n_m}), \mathcal{R}_{n_m,s''}(x_{n_m}, n_m^{-\bar{r}})\right\} \\ &\ge \inf_{x_{n_m}} \max\left\{E_1\left(\psi_{s',n_m}^{-p}\,\|x_{n_m} - x_{f_1}\|^p\right), E_0\left(n_m^{p\bar{r}}\,\|x_{n_m} - x_{f_0}\|^p\right)\right\} \\ &= \inf_{x_{n_m}} \max\left\{E_1D^p(x_{n_m},1), Q^pE_0D^p(x_{n_m},0)\right\} \\ &\ge \frac{(1-l_n)(1-2\delta)^p\tau(Q\delta)^p}{(1-2\delta)^p + \tau(Q\delta)^p}. \end{aligned} \tag{61}$$

Here, in view of (8), denoting $r' = (s'-1)/(2s'+d)$,

$$\begin{aligned} \tau Q^p &= \exp\{-(1-\delta)p[(1/2)-r']\log n\}\left(\frac{\psi_{s',n}}{n^{-\bar{r}}}\right)^p \\ &= \exp\{p\log n\,[\bar{r} - (1/2) + \delta((1/2)-r')] + pr'\log\log n\}\,. \end{aligned}$$

Choose $1/2 > \delta > ((1/2) - \bar{r})/((1/2) - r')$. This is possible by (47). Then $\tau Q^p \to \infty$, as $n \to \infty$. Since also $l_n \to 0$, by (53), we conclude that the right hand side of (61) tends to $(1 - 2\delta)^p$, as $n \to \infty$. This proves (50) and thus the theorem. $\square$

Acknowledgements

I wish to thank two referees from helpful comments.

References

Butucea, C. (2000*a*), 'The adaptive rate of convergence in a problem of pointwise density estimation', *Statist. Probab. Lett.* **47**, 85–90.

Butucea, C. (2000*b*), 'Two adaptive rates of convergence in pointwise density estimation', *Mathematical Methods of Statistics* **9**, 39–64.

Donoho, D. L. and Liu, R. C. (1991), 'Geometrizing rates of convergence III', *Ann. Statist.* **19**, 668–701.

Grenander, U. (1965), 'Some direct estimates of the mode', *Ann. Math. Stat.* **36**, 131–138.

Grund, B. and Hall, P. (1995), 'On the minimisation of L_p error in mode estimation', *Ann. Stat.* **23**, 2264–2284.

Hasminskii, R. (1979), Lower bound for the risks of nonparametric estimates of the mode, *in* J. Jureckova, ed., 'Contributions to Statistics: Jaroslav Hájek Memorial Volume', Academia, Prague, pp. 91–97.

Klemelä, J. and Tsybakov, A. B. (2001), 'Sharp adaptive estimation of linear functionals', *Ann. Statist.* **29**, 1567–1600.

Konakov, V. D. (1973), 'On the asymptotic normality of the mode of multidimensional distributions', *Theory Probab. Appl.* **18**, 794–803.

Lepski, O. V. (1990), 'One problem of adaptive estimation in gaussian white noise', *Theory Probab. Appl.* **35**, 454–466.

Lepski, O. V. (1991), 'Asymptotically minimax adaptive estimation. I: Upper bounds. Optimally adaptive estimates', *Theory Probab. Appl.* **36**, 682–697.

Lepski, O. V. (1992), 'Asymptotically minimax adaptive estimation. II: Schemes without optimal adaptation. Adaptive estimates', *Theory Probab. Appl.* **37**, 433–448.

Mammen, E., Marron, J. S. and Fisher, N. I. (1992), 'Some asymptotics for multimodality tests based on kernel density estimates', *Probab. Theory Relat. Fields* **91**, 115–132.

Müller, D. W. and Sawitzki, G. (1991), 'Excess mass estimates and tests of multimodality', *J. Amer. Statist. Assoc.* **86**, 738–746.

Müller, H.-G. (1989), 'Adaptive nonparametric peak estimation', *Ann. Statist.* **17**(3), 1053–1069.

Park, B. U. and Marron, J. S. (1990), 'Comparison of data-driven bandwidth selectors', *J. Amer. Statist. Assoc.* **85**, 66–72.

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

Petrov, V. V. (1995), *Limit Theorems of Probability Theory*, Clarendon Press, Oxford.

Sheather, S. J. and Jones, M. C. (1991), 'A reliable data-based bandwidth selection method for kernel estimation', *J. Roy. Statist. Soc. Ser. B* **53**, 683–690.

Tsybakov, A. B. (1990), 'Recursive estimation of the mode of a multivariate density', *Probl. Inf. Transm.* **26**, 31–37.

Tsybakov, A. B. (1998), 'Pointwise and sup-norm sharp adaptive estimation of functions on the Sobolev classes', *Ann. Statist.* **26**, 2420–2469.

van de Geer, S. A. (2000), *Empirical Processes in M-Estimation*, Cambridge University Press.

Wand, M. P. and Jones, M. C. (1994), 'Multivariate plug-in bandwidth seletion', *Comput. Statist.* **9**, 97–116.

Ziegler, K. (2000), On adaptive kernel estimation of location and size of the mode of the regression function, Technical report, University of München.

Jussi Klemelä
University of Mannheim
Economics Faculty, Department of Statistics
L 7 3-5 Verfügungsgebäude
68131 Mannheim, Germany
Email: klemela@rumms.uni-mannheim.de