2009

Asymptotics and Optimal Bandwidth Selection for Highest Density Region Estimation

R. J. Samworth
University of Cambridge

M. P. Wand
University of Wollongong, mwand@uow.edu.au

Centre for Statistical and Survey Methodology

The University of Wollongong

Working Paper

02-09

Asymptotics and Optimal Bandwidth Selection for Highest Density Region Estimation

Samworth, R.J. and Wand, M.P.

Centre for Statistical & Survey Methodology, University of Wollongong, Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845. Email: anica@uow.edu.au

ASYMPTOTICS AND OPTIMAL BANDWIDTH SELECTION FOR HIGHEST DENSITY REGION ESTIMATION

BY R.J. SAMWORTH AND M.P. WAND

University of Cambridge and University of Wollongong

14th August, 2009

ABSTRACT

We study kernel estimation of highest density regions (HDR). Our main contributions are two-fold. Firstly, we derive a uniform-in-bandwidth asymptotic approximation to a risk that is appropriate for HDR estimation. This approximation is then used to derive a bandwidth selection rule for HDR estimation possessing attractive asymptotic properties. We also present the results of numerical studies that illustrate the benefits of our theory and methodology.

Keywords: Density contour; Density level set; Kernel density estimator; Plug-in bandwidth selection.

1 Introduction

A highest density region (HDR) for a measurement of interest is a region where the underlying density function exceeds some nominal threshold. Given a random sample from that density, HDR estimation typically involves determination of regions where an estimated density is high. Kernel density estimation is the most common approach, but its performance is heavily dependent on the choice of the bandwidth parameter. Automatic selection of a good bandwidth for HDR estimation is the overarching goal of this article.

Figure 1 illustrates the bandwidth selection issue for HDR estimation. The left panel shows five kernel density estimates based on random samples of size 1000 from the normal mixture $\frac{2}{3}N(0,1) + \frac{1}{3}N(0, \frac{1}{100})$ density (Density 4 of Marron and Wand, 1992). In each case the bandwidth is chosen to minimise the integrated squared error (ISE). In the right panel the same random samples are used but, instead, the bandwidths are chosen to minimise an error appropriate for estimation of the 20% HDR (defined formally in Section 2). This region is shown as a thick horizontal line at the base of the plot. It is clear from Figure 1 that optimality for HDR estimation is quite different from ISE-optimality. Low ISE requires that the two curves be close to each other over the whole real line. However, good estimation of the 20% HDR only requires that the 20% HDRs of the kernel density estimates are close to the true region. In particular, the sharp mode of the underlying density has no bearing upon the HDR and there is no need to estimate it well. For this density it is apparent that a bandwidth considerably larger than ISE-optimal bandwidth is best for estimation of the 20% HDR.

In this article we study an asymptotic risk associated with kernel-based HDR estimation and use our theory to develop a plug-in type bandwidth selector. Attractive asymptotic properties of our bandwidth selector are established and good performance is illustrated on simulated data. A self-contained function for use in the `R` environment (R Development Core Team, 2008) is made available on the Internet.

The HDR estimation problem has an established literature. Contributions include Hartigan (1987), Müller & Sawitzki (1991), Polonik (1995), Hyndman (1996), Tsybakov (1997), Baíllo,

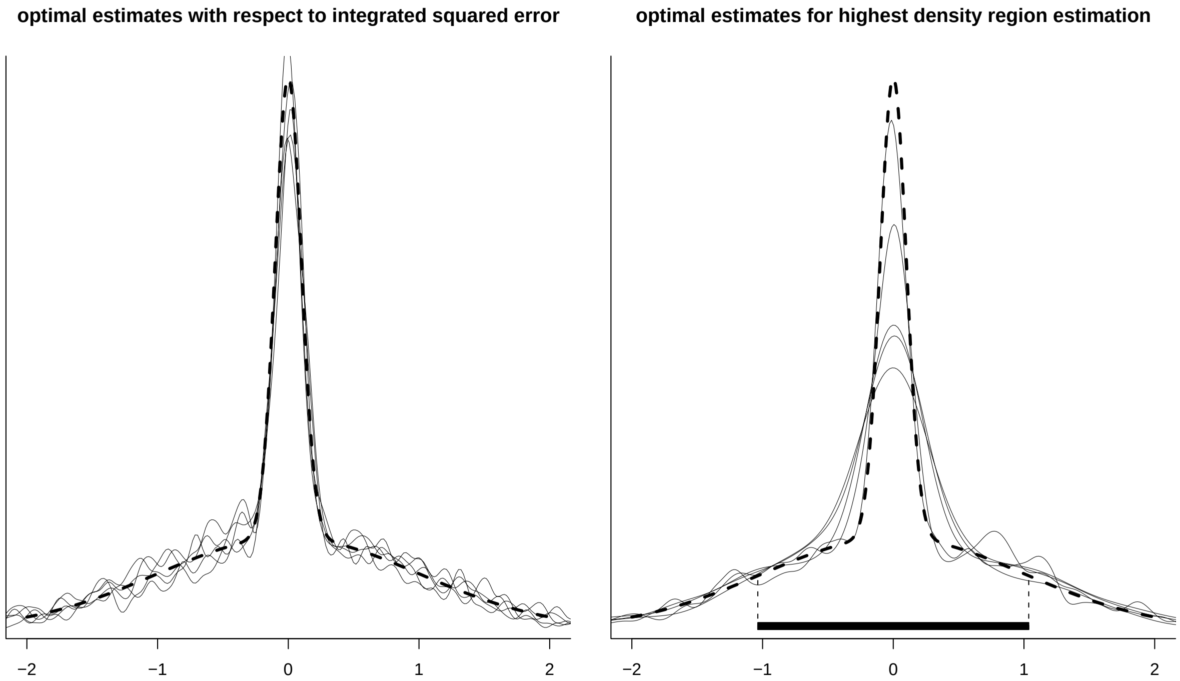

Figure 1: *Left panel: five kernel density estimates based on random samples of size 1000 simulated from the density depicted by the dashed curve. Each estimate is based on the optimal bandwidth with respect to integrated squared error. Right panel: same as the left panel except that the bandwidth is chosen to minimise the error for estimation of the 20% highest density region. This region is shown as a thick horizontal line at the base of the plot and its boundaries are shown as dashed vertical lines.*

Cuesta-Albertos & Cuevas (2001), Baíllo (2003), Cadre (2006), Jang (2006), Rigollet & Vert (2009) and Mason & Polonik (2009). Mason & Polonik (2009) provide a thorough literature review for the problem. Alternative terminology includes estimation of the *density contours*, *density level sets* and *excess mass regions*. This literature is, however, mainly concerned with theoretical results unconnected with the bandwidth selection problem. Jang (2006) is an applied paper on the use of HDR estimation for astronomical sky surveys. However, the bandwidths used there are chosen via classical ISE-based plug-in strategies. The present paper is, to our knowledge, the first to derive theory and bandwidth selection rules that are specifically tailored to the HDR estimation problem.

While our proposed practical bandwidth selector relies on asymptotic approximations, its development comes at a time when sample sizes in applications that benefit from smoothing techniques are becoming very large. The area of application that led to this research, flow cytometry, typically has sample sizes in the hundreds of thousands. The astronomical application in Jang (2006) involves sample sizes in the tens of thousands.

Section 2 presents an approximation to the HDR asymptotic risk. Numerical studies support its use for bandwidth selection. In Section 3 we describe plug-in strategies for bandwidth selection. Asymptotic performance results are established and a simulation study demonstrates practical efficacy. We conclude with an example on daily temperature maxima in Melbourne, Australia. Proofs are deferred to an appendix.

2 Asymptotic Risk Results

Let f be a probability density function on the real line. For $\tau \in (0,1)$, define

$$f_\tau = f_\tau(f) = \inf\biggl\{y \in (0,\infty) : \int_{-\infty}^{\infty} f(x)\mathbb{1}_{\{f(x)\geq y\}}\,dx \leq 1-\tau\biggr\}.$$

We call $R_\tau = \{x \in \mathbb{R} : f(x) \geq f_\tau\}$ the $100(1-\tau)\%$ *highest density region* of f (cf. Hyndman, 1996). If (X_n) is a sequence of independent random variables with density f, the kernel estimator of $f(x)$ based on $X_1,\ldots,X_n$ is

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\Bigl(\frac{x-X_i}{h}\Bigr),$$

where $K : \mathbb{R} \to \mathbb{R}$ satisfies $\int K(x)\,dx = 1$ and is called a *kernel* and $h > 0$ is called the *bandwidth*. Let $\hat{f}_{h,\tau} = f_\tau(\hat{f}_h)$ denote the plug-in estimator of f_τ, so that

$$\hat{f}_{h,\tau} = \inf\biggl\{y \in (0,\infty) : \int_{-\infty}^{\infty} \hat{f}_h(x)\mathbb{1}_{\{\hat{f}_h(x)\geq y\}}\,dx \leq 1-\tau\biggr\}.$$

The corresponding plug-in estimator of R_τ is then $\hat{R}_{h,\tau} = \{x \in \mathbb{R} : \hat{f}_h(x) \geq \hat{f}_{h,\tau}\}$.

Given two Borel subsets A and B of $\mathbb{R}$, we define their proximity through a measure on their symmetric difference $A\triangle B = (A \cap B^c) \cup (A^c \cap B)$. The particular measure μ_f we consider is given by

$$\mu_f(C) = \int_C f(x)\,dx,$$

for all Borel subsets C of $\mathbb{R}$. The error $\mu_f(\hat{R}_{h,\tau}\triangle R_\tau)$ is then then the probability of an observation from f lying in precisely one of $\hat{R}_{h,\tau}$ and R_τ. Compared with Lebesgue measure, μ_f puts more weight on regions where the data will tend to be denser. It also has the advantage of admitting a simple Monte Carlo approximation. This is important in higher dimensional settings where exact computation of $\mu_f(C)$ is difficult.

In Theorem 1, we derive a uniform-in-bandwidth asymptotic expansion for the *risk* $\mathbb{E}\{\mu_f(\hat{R}_{h,\tau}\triangle R_\tau)\}$, which can facilitate a theoretical, optimal choice of bandwidth (cf. Corollary 2). This in turn motivates practical bandwidth selection algorithms whose performance is studied in Theorems 3 and 4. We will make use of the following conditions on the underlying density, bandwidth sequence and kernel:

(A1): f is uniformly continuous on $\mathbb{R}$. There exist finitely many points $x_1 < \ldots < x_{2r}$ such that $f(x_j) = f_\tau$ for $j = 1,\ldots,2r$, and moreover there exists $\delta > 0$ such that f is twice continuously differentiable in $\cup_{j=1}^{r}[x_{2j-1}-\delta, x_{2j}+\delta]$ with $f'(x_{2j-1}) > 0$ and $f'(x_{2j}) < 0$ for $j = 1,\ldots,r$.

(A2): Let $h^- = h_n^-$ and $h^+ = h_n^+$ be non-negative sequences such that $h^- \leq h^+$, such that $n(h^-)^4/\sqrt{\log(1/h^-)} \to \infty$ and such that $h^+ \to 0$ as $n \to \infty$. Then $h = h_n$ is a sequence with $h_n^- \leq h_n \leq h_n^+$ for all n.

(A3): The kernel K is non-negative, continuously differentiable, of bounded variation, and satisfies $\int xK(x)\,dx = 0$ and $\mu_2(K) \equiv \int x^2K(x)\,dx < \infty$. Moreover, K' is of bounded variation, and satisfies $\int K'(x)^2\,dx < \infty$.

Assumption (**A1**) in particular implies that f has a γ-exponent with $\gamma = 1$ at level f_τ – in other words, there exists $C > 0$ such that

$$\mu_f(\{x \in \mathbb{R} : |f(x) - f_\tau| \leq \epsilon\}) \leq C\epsilon$$

for sufficiently small $\epsilon > 0$. This type of assumption is common in the literature for this problem – cf. Polonik (1995), Rigollet and Vert (2009). Although there are many parts to condition (**A3**), none is very restrictive. Under (**A1**), f_τ is the unique positive real number satisfying $\int f(x)\mathbb{1}_{\{f(x)\geq f_\tau\}}\,dx = 1-\tau$. In fact, in the course of the proof of Theorem 1 below, we will show that under conditions (**A1**), (**A2**) and (**A3**), $\hat{f}_{h,\tau}$ has an analogous property: that is, with probability one, for all n sufficiently large, $\hat{f}_{h,\tau}$ is the unique positive real number satisfying

$$\int \hat{f}_h(x)\mathbb{1}_{\{\hat{f}_h(x)\geq \hat{f}_{h,\tau}\}}\,dx = 1-\tau.$$

Let Φ and ϕ denote the standard normal distribution function and density function respectively and write $R(K) = \int K^2(x)\,dx$. Define the quantities

$$\begin{aligned} D_1 &= \frac{1}{2}\mu_2(K)\biggl\{\sum_{j=1}^{2r}\frac{1}{|f'(x_j)|}\biggr\}^{-1}\biggl[\sum_{j=1}^{2r}\frac{f''(x_j)}{|f'(x_j)|} + \frac{1}{f_\tau}\sum_{j=1}^{r}\{f'(x_{2j}) - f'(x_{2j-1})\}\biggr], \\ D_2 &= R(K)f_\tau\biggl\{\sum_{j=1}^{2r}\frac{1}{|f'(x_j)|}\biggr\}^{-2}\sum_{j=1}^{2r}\frac{1}{f'(x_j)^2} \\ \text{and } D_{3,j} &= \frac{R(K)f_\tau}{|f'(x_j)|}\biggl\{\sum_{k=1}^{2r}\frac{1}{|f'(x_k)|}\biggr\}^{-1}, \quad j = 1,\ldots,2r. \end{aligned} \tag{2.1}$$

Theorem 1. *Assume (**A1**), (**A2**) and (**A3**). Then*

$$\mathbb{E}\{\mu_f(\hat{R}_{h,\tau}\triangle R_\tau)\} = \sum_{j=1}^{2r}\biggl[\frac{B_{1,j}\phi(B_{2,j}n^{1/2}h^{5/2})}{(nh)^{1/2}} + B_{3,j}h^2\{2\Phi(B_{2,j}n^{1/2}h^{5/2})-1\}\biggr] + o\Bigl(\frac{1}{(nh)^{1/2}}+h^2\Bigr)$$

as $n\to\infty$, uniformly for $h\in[h^-,h^+]$, where

$$B_{1,j} = 2f_\tau\frac{\{R(K)f_\tau - 2D_{3,j} + D_2\}^{1/2}}{|f'(x_j)|}, \quad B_{2,j} = \frac{|\frac{1}{2}\mu_2(K)f''(x_j) - D_1|}{\{R(K)f_\tau - 2D_{3,j} + D_2\}^{1/2}}$$

and

$$B_{3,j} = f_\tau\frac{|\frac{1}{2}\mu_2(K)f''(x_j) - D_1|}{|f'(x_j)|}.$$

The nature of this result is somewhat different from the results in the existing literature, which have tended to focus (sometimes in more general settings) on the order in probability or almost surely of $\mu_f(\hat{R}_{h,\tau}\triangle R_\tau)$ or related measures (e.g. Baíllo, Cuesta-Albertos & Cuevas (2001), Baíllo (2003)). More recent works have derived results on the limiting behaviour of suitably scaled and/or centered versions of $\mu_f(\hat{R}_{h,\tau}\triangle R_\tau)$ (e.g. Cadre (2006), Mason & Polonik (2009)). Rigollet & Vert (2009) provide a finite sample upper bound for the risk, uniformly over certain Hölder classes, with an unspecified constant in the bound. While these theoretical results are certainly of considerable interest, our aim in providing the asymptotic expansion in Theorem 1 is to facilitate practical bandwidth selection algorithms for this problem – see Section 3.

In the course of the proof of Theorem 1, it is shown that

$$R(K)f_\tau - 2D_{3,j} + D_2 = \lim_{n\to\infty}(nh)\mathrm{Var}(\hat{f}_h(x_j) - \hat{f}_{h,\tau}) > 0,$$

so that $B_{1,j}$ is positive. Moreover $B_{2,j}$ and $B_{3,j}$ are non-negative, and will be positive except in pathological circumstances. Assuming $B_{2,j}$ and $B_{3,j}$ are positive, it is easily seen from Theorem 1 that for any sequence of bandwidths satisfying (**A2**), if nh^5 is not bounded away from zero and infinity then $n^{2/5}\mathbb{E}\{\mu_f(\hat{R}_{h,\tau}\triangle R_\tau)\} \to \infty$ along a sub-sequence. On the other hand, if nh^5 is bounded away from zero and infinity, then $n^{2/5}\mathbb{E}\{\mu_f(\hat{R}_{h,\tau}\triangle R_\tau)\}$ is bounded. Notice that all such sequences are permitted by the condition (**A2**). In certain circumstances, there is a rather stronger conclusion:

Corollary 2. *Assume (**A1**) and (**A3**) and that $B_{2,j}$ and $B_{3,j}$ are positive. Assume further that in (**A1**) we have $r = 1$ and the underlying density f is symmetric about some point on the real line. Then there exists a unique $c_{\text{opt}} \in (0, \infty)$, depending on f and K but not n, such that any sequence of bandwidths (h_{opt}) that minimises $\mathbb{E}\{\mu_f(\hat{R}_{h,\tau}\triangle R_\tau)\}$ satisfies*

$$h_{\text{opt}} = c_{\text{opt}} n^{-1/5}\{1 + o(1)\}$$

as $n \to \infty$.

Although the hypotheses on f in Corollary 2 are restrictive, we have not encountered significant problems caused by multiple local minima of the risk function even when these hypotheses are not satisfied.

2.1 Numerical Assessment of Risk Approximation

Theorem 1 yields the asymptotic risk approximation

$$\mathbb{E}\{\mu_f(\hat{R}_{h,\tau}\triangle R_\tau)\} \simeq \sum_{j=1}^{2r}\left[\frac{B_{1,j}\phi(B_{2,j}n^{1/2}h^{5/2})}{(nh)^{1/2}} + B_{3,j}h^2\{2\Phi(B_{2,j}n^{1/2}h^{5/2}) - 1\}\right]. \tag{2.2}$$

In Section 3 we use the right-hand side of (2.2) to develop plug-in bandwidth selection strategies. However, it is prudent to first assess the quality of this approximation to the risk. We now do this through some numerical examples.

For a given f, h and τ, the risk $\mathbb{E}\{\mu_f(\hat{R}_{h,\tau}\triangle R_\tau)\}$ is very difficult to obtain exactly. Instead, we work with a Monte Carlo approximation

$$\frac{1}{M}\sum_{i=1}^{M}\mu_f(\hat{R}_{h,\tau}^{[i]}\triangle R_\tau) \tag{2.3}$$

where $\hat{R}_{h,\tau}^{[1]}, \ldots, \hat{R}_{h,\tau}^{[M]}$ are M simulated realisations of $\hat{R}_{h,\tau}$. For large M (2.3) serves as reasonable proxy for $\mathbb{E}\{\mu_f(\hat{R}_{h,\tau}\triangle R_\tau)\}$ and is henceforth referred to as the 'exact' risk.

Figure 2 compares the asymptotic risk approximation with its 'exact' counterpart for $n = 1000$, f corresponding to Densities 1–10 of Marron and Wand (1992) and $\tau = 0.5$. The kernel K is set to ϕ throughout and the Monte Carlo sample size is $M = 100$. For most of these densities the asymptotic risk approximation is quite good for $n = 1000$. Densities 3 and 4 are the main exceptions; it appears that larger sample sizes are required for the leading terms to be dominant. In particular, for these densities, the difficulty appears to be caused by very large values of $|f'|$ and/or $|f''|$ at the crossing points of $f_{0.5}$ (for Density 4, the level $f_{0.5}$ is very close to the rapid transition from shallow to steep gradient seen in Figure 1).

In Figure 3 we repeated the calculations that produced Figure 2, but with τ set to 0.8 and $n = 100000$. For several densities, including the one depicted in Figure 1, the estimand $R_{0.8}$ corresponds to the fine detail of f. The asymptotic risk approximation is seen to be quite good for these values of τ and n.

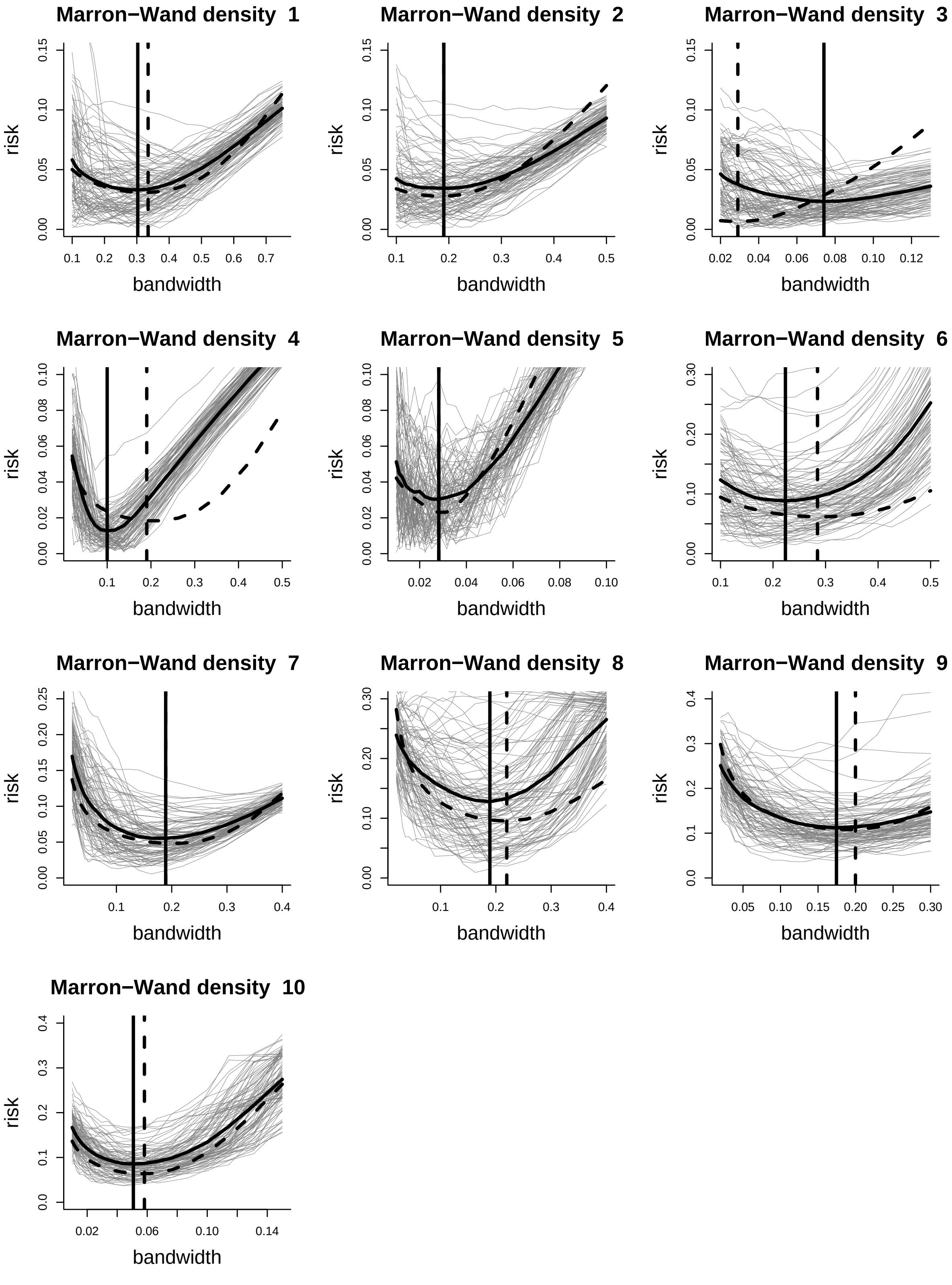

Figure 2: *Comparison of the 'exact' risk $\mathbb{E}\{\mu_f(\hat{R}_{h,0.5}\triangle R_{0.5})\}$ and its asymptotic approximation for the first ten density functions of Marron and Wand (1992) and $n = 1000$. In each panel, the 'exact' risk is obtained by averaging 100 realisations of $\mu_f(\hat{R}_{h,0.5}\triangle R_{0.5})$ (shown as grey curves) and is shown as a solid black curve. The dashed curve is the asymptotic risk approximation corresponding to the right hand side of (2.2). Vertical lines pass through the minima of the 'exact' risk (solid line) and the asymptotic risk (broken line).*

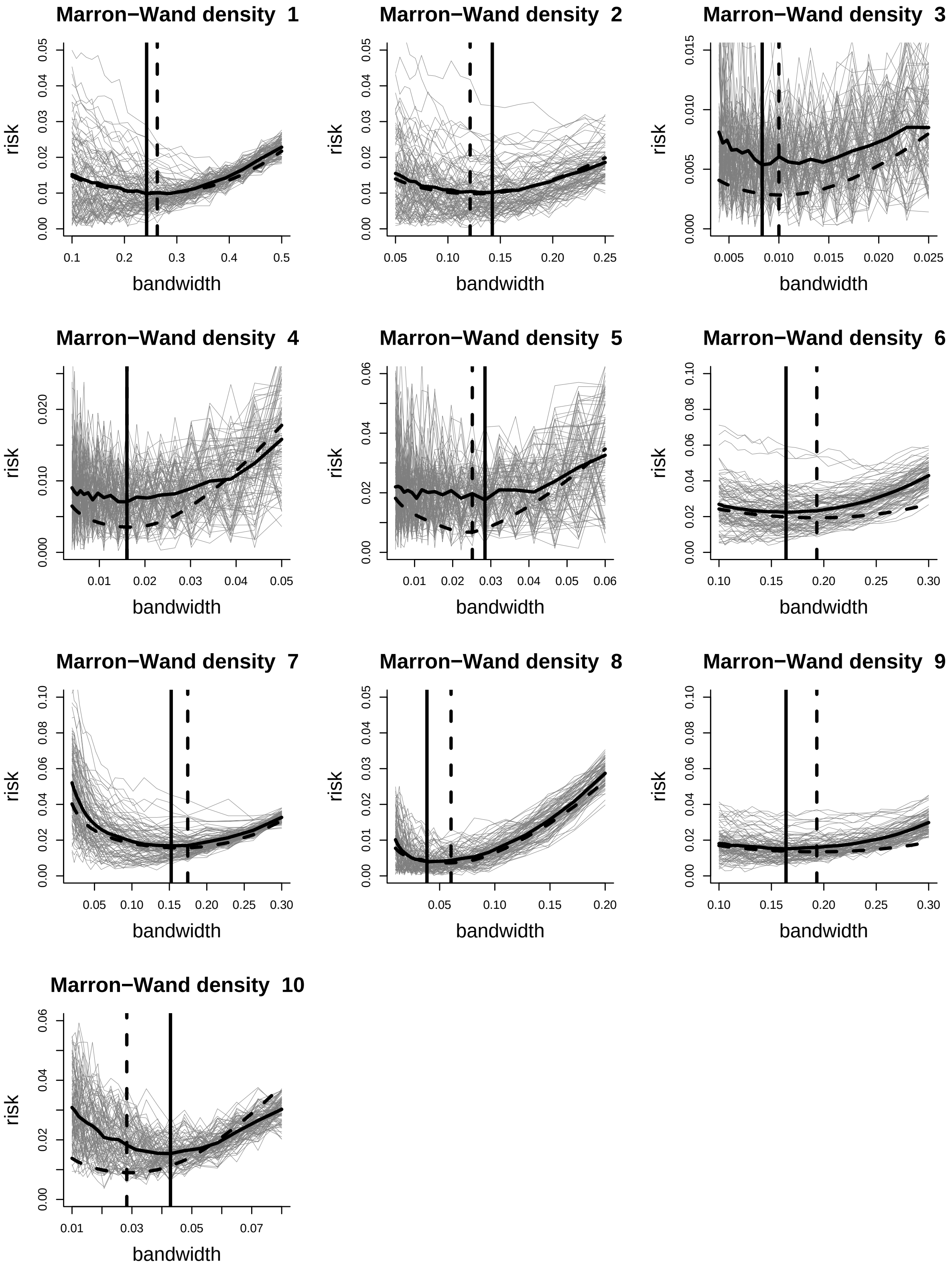

Figure 3: *Comparison of the 'exact' risk $\mathbb{E}\{\mu_f(\hat{R}_{h,0.8}\triangle R_{0.8})\}$ and its asymptotic approximation for the first ten density functions of Marron and Wand (1992) and $n = 100000$. In each panel, the 'exact' risk is obtained by averaging 100 realisations of $\mu_f(\hat{R}_{h,0.8}\triangle R_{0.8})$ (shown as grey curves) and is shown as a solid black curve. The dashed curve is the asymptotic risk approximation corresponding to the right hand side of (2.2). Vertical lines pass through the minima of the 'exact' risk (solid line) and the asymptotic risk (broken line).*

3 Bandwidth Selection

In this section, we assume that, as in Corollary 2, there exists a unique $c_{\text{opt}} \in (0, \infty)$ such that any optimal bandwidth sequence satisfies $h_{\text{opt}} = c_{\text{opt}} n^{-1/5}\{1 + o(1)\}$. In this case, c_{opt} minimises the asymptotic risk, given by

$$AR(c) = \frac{1}{n^{2/5}} \sum_{j=1}^{2r} \biggl[\frac{B_{1,j}}{c^{1/2}} \phi(B_{2,j} c^{5/2}) + B_{3,j} c^2 \{2\Phi(B_{2,j} c^{5/2}) - 1\} \biggr]. \tag{3.1}$$

In order to find a practical bandwidth selector, we seek an estimator $\hat{c}_{\text{opt}}$ of c_{opt}. The natural way to construct such an estimator is by using estimators $\hat{D}_1$, $\hat{D}_2$ and $\hat{D}_{3,j}$ of D_1, D_2 and $D_{3,j}$ respectively to obtain plug-in estimators $\hat{B}_{1,j}$, $\hat{B}_{2,j}$ and $\hat{B}_{3,j}$ of $B_{1,j}$, $B_{2,j}$ and $B_{3,j}$ respectively. These in turn can be used to find $\hat{c}_{\text{opt}} = \text{argmin}_{c \in (0,\infty)} \widehat{AR}_n(c)$, where

$$\widehat{AR}_n(c) = \frac{1}{n^{2/5}} \sum_{j=1}^{2r} \biggl[\frac{\hat{B}_{1,j}}{c^{1/2}} \phi(\hat{B}_{2,j} c^{5/2}) + \hat{B}_{3,j} c^2 \{2\Phi(\hat{B}_{2,j} c^{5/2}) - 1\} \biggr]. \tag{3.2}$$

With probability one, the solution to this minimisation problem will be unique for large n provided that $AR''(c_{\text{opt}}) > 0$ and this solution can easily be found numerically. Our final bandwidth selector is then $\widehat{h}_{\tau\text{HDR}} = \hat{c}_{\text{opt}} n^{-1/5}$.

Note that we have not yet described how to construct the estimators $\hat{D}_1$, $\hat{D}_2$ and $\hat{D}_{3,j}$. Again, we propose plug-in estimators, based on estimates of f_τ as well as $f'(x_j)$ and $f''(x_j)$ for $j = 1, \ldots, 2r$. We assume the kernel K is smooth, and will construct kernel estimators $\hat{f}_{h_0}(\hat{x}_{j,h_0})$, $\hat{f}'_{h_1}(\hat{x}_{j,h_0})$ and $\hat{f}''_{h_2}(\hat{x}_{j,h_0})$ of f_τ, $f'(x_j)$ and $f''(x_j)$ respectively, where $\hat{x}_{j,h_0}$ is an estimator of x_j described below. For the time being, we will use the same kernel K in all cases; this requirement will be dropped later on. Even at this stage it will, however, be important to note that we can use different bandwidths h_0, h_1 and h_2. Recall (e.g. Wand and Jones, 1995, p.49) that, under appropriate conditions, if $h_k \asymp n^{-1/(2k+5)}$ as $n \to \infty$ then $\hat{f}^{(k)}_{h_k}(x_j) - f^{(k)}(x_j) = O_p(n^{-2/(2k+5)})$ and that this order cannot be improved for a non-negative kernel. Here we have used the notation $a_n \asymp b_n$ as $n \to \infty$ to mean $0 < \liminf_{n\to\infty} |a_n/b_n| \leq \limsup_{n\to\infty} |a_n/b_n| < \infty$. Finally, we observe that if h_0 satisfies (**A2**), then with probability one, for all sufficiently large n there exist $\hat{x}_{1,h_0} < \ldots \hat{x}_{2r,h_0}$ such that $\hat{f}_{h_0}(\hat{x}_{j,h_0}) = \hat{f}_{h_0,\tau}$ for each j, and we use $\hat{x}_{j,h_0}$ to estimate x_j. Our theoretical study of the performance of this bandwidth selector requires some additional conditions:

(A4): f has four continuous derivatives in an open set containing each x_j.

(A5): $h_0 \asymp n^{-1/5}$, $h_1 \asymp n^{-1/7}$ and $h_2 \asymp n^{-1/9}$ as $n \to \infty$.

(A6): K has a bounded third derivative, K'' is of bounded variation and $\int |x|^3|K'(x)| + x^4|K''(x)|\, dx < \infty$.

Theorem 3. *Assume (**A1**) and (**A3**)–(**A6**) and that $B_{2,j}$ and $B_{3,j}$ are positive. Assume further that c_{opt} is unique and that $AR''(c_{\text{opt}}) > 0$. Then*

$$\frac{\widehat{h}_{\tau HDR}}{h_{\text{opt}}} = 1 + O_p(n^{-2/9})$$

as $n \to \infty$. Moreover, recalling that $\widehat{h}_{\tau HDR} = \hat{c}_{\text{opt}} n^{-1/5}$, we have

$$\frac{\widehat{AR}_n(\hat{c}_{\text{opt}})}{AR(c_{\text{opt}})} = 1 + O_p(n^{-2/9}).$$

Examining the proof of Theorem 3 reveals that the rate of convergence to zero of the relative error of $\widehat{h}_{\tau\text{HDR}}$ is determined by the rate at which we can estimate $f''(x_j)$ for $j = 1, \ldots, 2r$. This suggests that we might be able to obtain a faster rate of convergence by using a higher order kernel to estimate $f''(x_j)$ (and in fact $f'(x_j)$). Recall that we call K an *Sth order kernel* if

1. $\int K(x)\, dx = 1$
2. $\mu_s(K) \equiv \int x^s K(x)\, dx = 0$ for $s = 1, \ldots, S-1$
3. $\mu_S(K) \equiv \int x^S K(x)\, dx \neq 0$ and $\int |x|^S |K(x)|\, dx < \infty$.

Higher order kernels refer to $S > 2$. The usual objection to the use of higher order kernels, namely that such a kernel cannot be non-negative, is less significant when the aim is to estimate derivatives of a density rather than the density itself. Let the kernels used to estimate $f'(x_j)$ and $f''(x_j)$ be denoted K_1 and K_2 respectively, and continue to denote the respective bandwidths by h_1 and h_2. An improved rate of convergence of the relative error of our bandwidth selector can be obtained by replacing conditions (**A4**), (**A5**) and (**A6**) with the following:

(A7): f has 12 continuous derivatives in an open set containing each x_j.

(A8): $h_0 \asymp n^{-1/5}$, $h_1 \asymp n^{-1/15}$ and $h_2 \asymp n^{-1/25}$ as $n \to \infty$.

(A9): K_1 is a 6th order kernel and has a bounded second derivative with K_1 and K_1' of bounded variation and satisfying $\int x^6 |K_1(x)| + |x|^7 |K_1'(x)|\, dx < \infty$. Moreover, K_2 is a 10th order kernel and has a bounded third derivative with K_2, K_2' and K_2'' of bounded variation and satisfying $\int x^{10} |K_2(x)| + |x|^{11} |K_2'(x)| + x^{12} |K_2''(x)|\, dx < \infty$.

We write $\widehat{\widehat{h}}_{\tau\text{HDR}}$ for the bandwidth selector obtained in a similar way to $\widehat{h}_{\tau\text{HDR}}$, but using the kernels K_1 and K_2 to estimate $f'(x_j)$ and $f''(x_j)$ respectively in the definitions of D_1, D_2, $D_{3,j}$, $B_{1,j}$, $B_{2,j}$ and $B_{3,j}$.

Theorem 4. *Assume (**A1**), (**A3**), (**A7**)–(**A9**) and that $B_{2,j}$ and $B_{3,j}$ are positive. Assume further that c_{opt} is unique and that $AR''(c_{\text{opt}}) > 0$. Then*

$$\frac{\widehat{\widehat{h}}_{\tau HDR}}{h_{\text{opt}}} = 1 + O_p(n^{-2/5})$$

as $n \to \infty$. Moreover, writing $\widehat{\widehat{h}}_{\tau HDR} = \hat{\hat{c}}_{\text{opt}} n^{-1/5}$, we have

$$\frac{\widehat{AR}_n(\hat{\hat{c}}_{\text{opt}})}{AR(c_{\text{opt}})} = 1 + O_p(n^{-2/5}).$$

It is clear that Theorem 3 represents a relatively weak conclusion under relatively weak conditions, while Theorem 4 represents a stronger conclusion under strong conditions. Intermediate results are also possible, but seem to be of little practical interest.

3.1 An Effective Practical Bandwidth Selector

We confine our development of a practical consistent bandwidth selector to the scenario where f satisfies weaker smoothness conditions of Theorem 3. Our end-product is a fast-to-compute bandwidth selector for HDR estimation that possesses the asymptotic properties conveyed by Theorem 3, performs well in simulations, and is readily implemented in R. Indeed, as detailed below, an R function for our procedure is available on the Internet.

The pilot bandwidths h_0, h_1 and h_2 are estimated using direct plug-in strategies with two levels of kernel functional estimation. Chapter 3 of Wand and Jones (1995) provides details on this general approach to bandwidth selection. In the case of h_0 the approach is similar to those proposed by Park and Marron (1990) and Sheather and Jones (1991). Direct plug-in bandwidth selection strategies for density functions and their derivatives involve estimation of functionals of the form

$$\psi_r = \int_{-\infty}^{\infty} f^{(r)}(x) f(x)\,dx.$$

Kernel estimators of ψ_r take the form

$$\widehat{\psi}_r(g) = n^{-2} g^{-r-1} \sum_{i=1}^{n} \sum_{j=1}^{n} L^{(r)}\{(X_i - X_j)/g\}$$

where L is a sufficiently smooth 2nd-order kernel function and $g > 0$ is a bandwidth parameter. Multi-level plug-in strategies use the fact that the asymptotically optimal g, with respect to the mean squared error of $\widehat{\psi}_r(g)$, is $[-2L^{(r)}(0)/\{n\psi_{r+2}\int u^2 L(u)\,du\}]^{1/(r+3)}$. To get the algorithm started we also require *normal scale* estimates of ψ_r, based on the assumption that f is a $N(\mu, \sigma^2)$ density. Normal scale estimates of ψ_r take the form

$$\widehat{\psi}_r^{\text{NS}} = \frac{(-1)^{r/2} r!}{(2\widehat{\sigma})^{r+1}(r/2)!\pi^{1/2}}.$$

Throughout we take $K = L = \phi$, the standard normal kernel. The full algorithm is:

1. The inputs are the random sample $X_1, \ldots, X_n$ and parameter $0 < \tau < 1$ specifying the required HDR.
2. Let $\widehat{\sigma} = \min(\text{sample standard deviation, (sample interquartile range)}/1.349)$ be a robust estimate of scale. (The interquartile range for the standard normal density is approximately 1.349, so this factor ensures approximate unbiased for normally distributed data.)
3. Estimate ψ_8, ψ_{10} and ψ_{12} using normal scale estimates. Explicit expressions for these are $\widehat{\psi}_8^{\text{NS}} = 105/(32\pi^{1/2}\widehat{\sigma}^9)$, $\widehat{\psi}_{10}^{\text{NS}} = -945/(64\pi^{1/2}\widehat{\sigma}^{11})$ and $\widehat{\psi}_{12}^{\text{NS}} = 10395/(128\pi^{1/2}\widehat{\sigma}^{13})$.
4. Estimate ψ_6, ψ_8 and ψ_{10} using kernel estimates $\widehat{\psi}_6(g_{0,1})$, $\widehat{\psi}_8(g_{1,1})$ and $\widehat{\psi}_{10}(g_{1,1})$ where $g_{0,1} = \{30/(\widehat{\psi}_8^{\text{NS}} n)\}^{1/9}$, $g_{1,1} = \{-210/(\widehat{\psi}_{10}^{\text{NS}} n)\}^{1/11}$ and $g_{1,2} = \{1890/(\widehat{\psi}_{12}^{\text{NS}} n)\}^{1/13}$.
5. Estimate ψ_4, ψ_6 and ψ_8 using kernel estimates $\widehat{\psi}_4(g_{0,2})$, $\widehat{\psi}_6(g_{1,2})$ and $\widehat{\psi}_8(g_{2,2})$ where $g_{0,2} = [6/\{\widehat{\psi}_8(g_{0,1})n\}]^{1/7}$, $g_{1,2} = [-30/\{\widehat{\psi}_{10}(g_{1,1})n\}]^{1/9}$ and $g_{2,2} = [210/\{\widehat{\psi}_{12}(g_{1,2})n\}]^{1/11}$.
6. Obtain direct plug-in bandwidths $\widehat{h}^{(r)}$ for estimation of $f^{(r)}$ by replacing ψ_{r+2} in the optimal expression, with respect to asymptotic mean integrated squared error, by $\widehat{\psi}_{r+2}(g_{r,2})$. Explicit expressions are $\widehat{h}_0 = [1/\{2\pi^{1/2}\widehat{\psi}_4(g_{0,2})n\}]^{1/5}$, $\widehat{h}_1 = [-3/\{4\pi^{1/2}\widehat{\psi}_6(g_{1,2})n\}]^{1/7}$ and $\widehat{h}_2 = [15/\{8\pi^{1/2}\widehat{\psi}_8(g_{2,2})n\}]^{1/9}$.
7. Obtain pilot of estimates of f, f' and f'' via Gaussian kernel estimates based on these bandwidths: $\widehat{f}_{\widehat{h}_0}(\cdot)$, $\widehat{f'}_{\widehat{h}_1}(\cdot)$ and $\widehat{f''}_{\widehat{h}_2}(\cdot)$.
8. Use $\widehat{f}_{\widehat{h}_0}(\cdot)$ to obtain pilot estimates of f_τ, r and $x_1, \ldots, x_{2r}$.
9. Substitute the estimates from Steps 6 and 7 into the expressions for $B_{1,j}$, $B_{2,j}$ and $B_{3,j}$ to obtain estimates $\widehat{B}_{1,j}$, $\widehat{B}_{2,j}$ and $\widehat{B}_{3,j}$.

10. The selected bandwidth for Gaussian kernel estimation of the $100(1-\tau)\%$ HDR is $\widehat{h}_{\tau\mathrm{HDR}} = \hat{c}_{\mathrm{opt}} n^{-1/5}$ where $\hat{c}_{\mathrm{opt}} = \mathrm{argmin}_{c\in(0,\infty)}\widehat{AR}_n(c)$, where $\widehat{AR}_n$ was defined in (3.2).

Binned approximations to $\widehat{\psi}_r(g)$ (cf. González-Manteiga, Sanchéz-Sellero, and Wand, 1996) are strongly recommended to allow fast processing of large samples. An `R` function `HDRbw()` that implements the above algorithm has been placed on the second author's 'R Functions and Packages' web-page. At the time of writing the relevant web-site has a link on the address `http://www.uow.edu.au/~mwand`. In addition, the code has been submitted to the authors of the `R` package `hdrcde` (Hyndman and Einbeck, 2009), which supports HDR estimation.

3.2 Simulation Results

We ran a simulation study in which the performance of $\widehat{h}_{\tau\mathrm{HDR}}$ was compared with an established ISE-based selector: least squares cross validation (Bowman, 1982; Bowman, 1984), which we denote by $\widehat{h}_{\mathrm{LSCV}}$. The number of replications in the simulation study was 250. The HDR estimation error $\mu_f(\hat{R}_{h,\tau}\triangle R_\tau)$ was used throughout the study. Figures 4 ($n = 1000$) and 5 ($n = 100000$) summarise the results for the situation where the true f is the normal mixture density from Section 1 and Figure 1. The improvement gained from using the HDR-tailored bandwidth selector is apparent from the graphics, especially for the lower values of τ. Wilcoxon tests applied to the error ratios showed statistically significant improvement of $\widehat{h}_{\tau\mathrm{HDR}}$ at the 5% level for $\tau = 0.2, 0.5$ and 0.8 when $n = 100000$. For $n = 1000$, $\widehat{h}_{\tau\mathrm{HDR}}$ performed better for $\tau = 0.2, 0.5$, while $\widehat{h}_{\mathrm{LSCV}}$ did better for $\tau = 0.8$. This latter result is not a big surprise since good estimation of $R_{0.8}$ requires good estimation of the finger-shaped modal region and this, in turn, requires good ISE performance.

We performed similar simulation comparison for the remaining Densities 1–10 of Marron and Wand (1992). For $n = 1000$ the performance of $\widehat{h}_{\tau\mathrm{HDR}}$ was better than $\widehat{h}_{\mathrm{LSCV}}$ for Densities 1–5; whereas $\widehat{h}_{\mathrm{LSCV}}$ did better for Densities 6–10. This suggests that the asymptotics on which $\widehat{h}_{\tau\mathrm{HDR}}$ relies have not 'kicked in' at $n = 1000$ for these more intricate density functions. We suspect that more sophisticated pilot estimation might improve matters for HDR-based bandwidth selection for lower sample sizes. The $n = 100000$ simulations show superior performance of $\widehat{h}_{\tau\mathrm{HDR}}$, especially $\tau = 0.8$ where it is the 'winner' for 9 out of the 10 density functions. The overarching conclusion is that for common density estimation situations $\widehat{h}_{\tau\mathrm{HDR}}$ is better than $\widehat{h}_{\mathrm{LSCV}}$.

3.3 Application to Daily Temperature Data

We conclude with an application to data on daily maximum temperatures in Melbourne, Australia, for the years 1981–1990. These data were used in Hyndman (1996) to illustrate HDR principles. We revisit them armed with the automatic HDR estimation technology described in Section 3.1 Of interest are the conditional densities

tomorrow's temperature *given* today's temperature is within a fixed interval.

The intervals for the 'today's temperature' values are, in degrees Celsius,

$$[5, 10), [10, 15), \ldots, [40, 45).$$

Figure 6 shows the kernel estimates of the 20%, 50% and 80% HDRs with bandwidths chosen using the rule $\widehat{h}_{\tau\mathrm{HDR}}$ as detailed in Section 3.1. Some interesting bimodality in 'tomorrow's

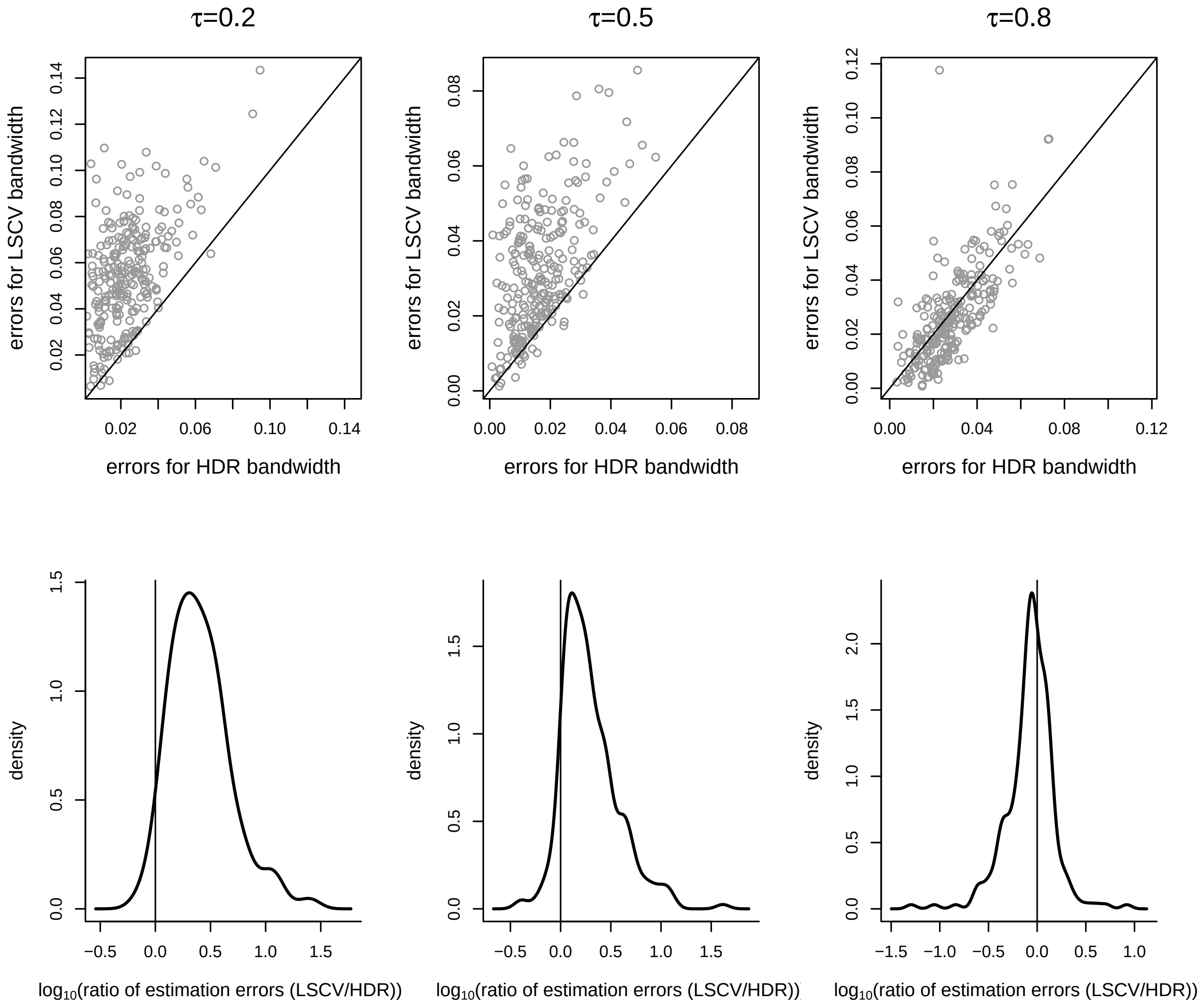

Figure 4: *Summary of simulation comparison between* $\widehat{h}_{\tau HDR}$ *and* $\widehat{h}_{LSCV}$ *for* $\tau = 0.2, 0.5$ *and* 0.8 *and 250 samples of size 1000 generated from Density 4 of Marron and Wand (1992). The upper panels are scatterplots of the errors* $\mu_f(\hat{R}_{h,\tau} \triangle R_\tau)$ *for* $h = \widehat{h}_{LSCV}$ *on the vertical axes and* $h = \widehat{h}_{\tau HDR}$ *on the horizontal axes. The lower panels are kernel density estimates of* $\log_{10}$*((error for* $h = \widehat{h}_{\tau HDR}$*)/(error for* $h = \widehat{h}_{LSCV}$*)).*

temperature' is apparent when conditioned on today's temperature being in the 30–40 degrees Celsius range.

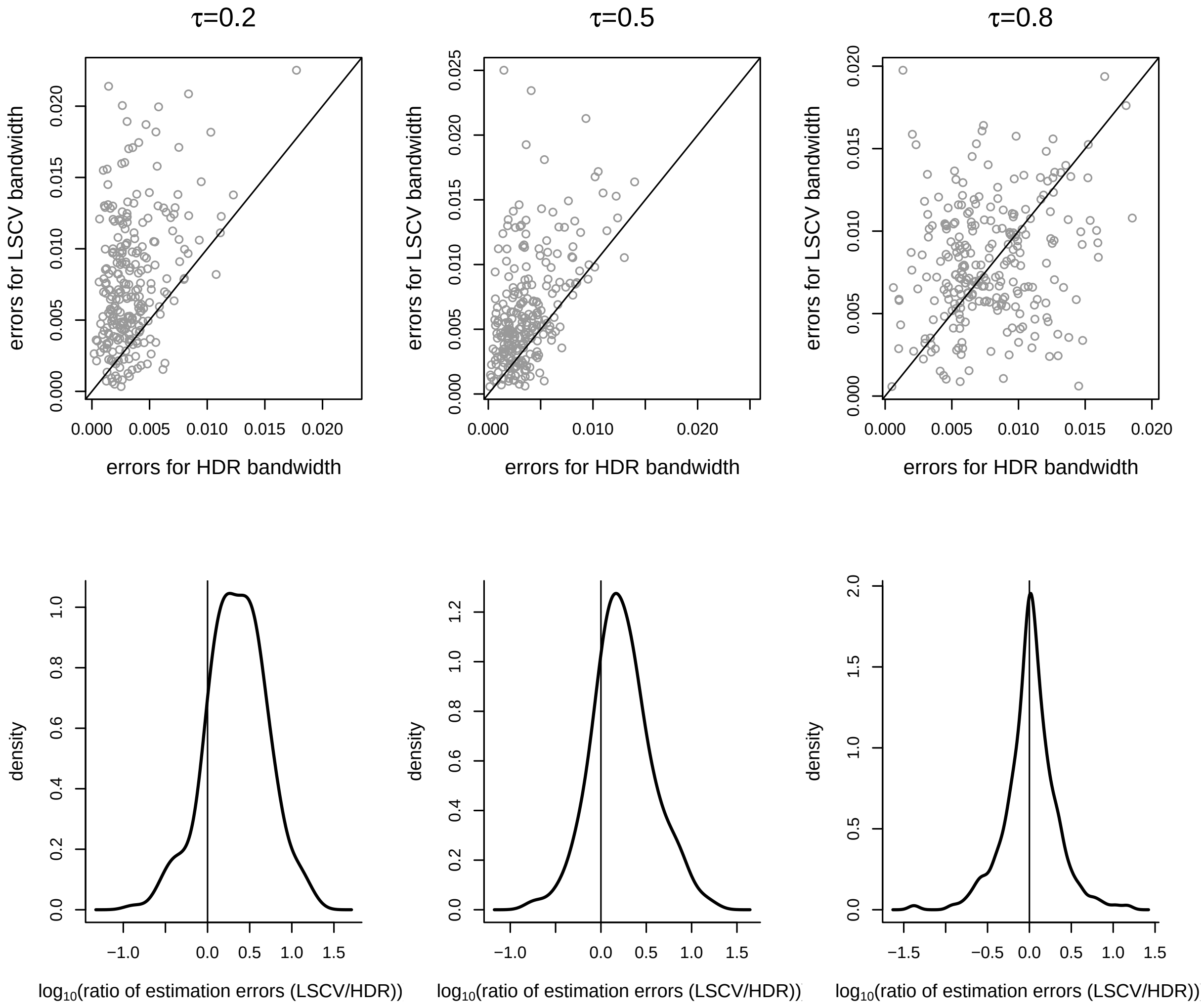

Figure 5: *Summary of simulation comparison between* $\widehat{h}_{\tau HDR}$ *and* $\widehat{h}_{LSCV}$ *for* $\tau = 0.2, 0.5$ *and* 0.8 *and 250 samples of size 100000 generated from Density 4 of Marron and Wand (1992). The upper panels are scatterplots of the errors* $\mu_f(\hat{R}_{h,\tau} \triangle R_\tau)$ *for* $h = \widehat{h}_{LSCV}$ *on the vertical axes and* $h = \widehat{h}_{\tau HDR}$ *on the horizontal axes. The lower panels are kernel density estimates of* $\log_{10}$*((error for* $h = \widehat{h}_{\tau HDR}$*)/(error for* $h = \widehat{h}_{LSCV}$*)).*

Appendix: Proofs

Proof of Theorem 1

Throughout the proof, it is convenient to write $x_0 = -\infty$ and $x_{2r+1} = \infty$ and adopt the convention that $x_0 + a = -\infty$ and $x_{2r+1} + a = \infty$ for all $a \in \mathbb{R}$. Observe that

$$\begin{aligned}\mu_f(\hat{R}_{h,\tau} \triangle R_\tau) &= \int_{-\infty}^{\infty} f(x) \big| \mathbb{1}_{\{\hat{f}_h(x) \geq \hat{f}_{h,\tau}\}} - \mathbb{1}_{\{f(x) \geq f_\tau\}} \big| \, dx \\ &= \sum_{j=0}^{r} \int_{x_{2j}}^{x_{2j+1}} f(x) \mathbb{1}_{\{\hat{f}_h(x) \geq \hat{f}_{h,\tau}\}} \, dx + \sum_{j=1}^{r} \int_{x_{2j-1}}^{x_{2j}} f(x) \mathbb{1}_{\{\hat{f}_h(x) < \hat{f}_{h,\tau}\}} \, dx,\end{aligned}$$

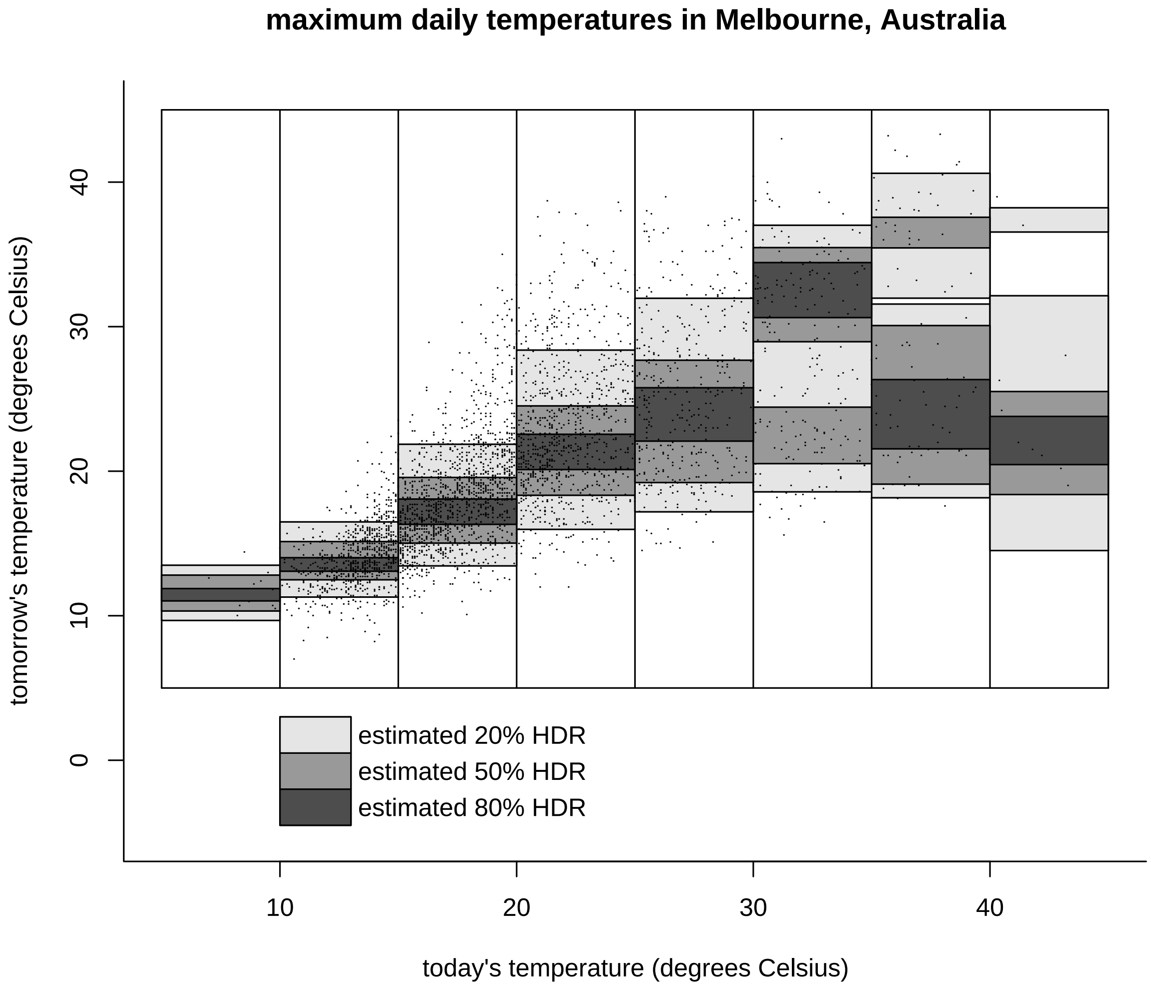

Figure 6: *Estimated kernel HDRs for the conditional densities of tomorrow's temperature given that today's temperature is in a fixed interval. The bandwidth for each HDR estimate is chosen using the selector described in Section 3.1.*

so that

$$\mathbb{E}\{\mu_f(\hat{R}_{h,\tau}\triangle R_\tau)\} = \sum_{j=0}^{r}\int_{x_{2j}}^{x_{2j+1}} f(x)\mathbb{P}\big(\hat{f}_h(x) \geq \hat{f}_{h,\tau}\big)\,dx + \sum_{j=1}^{r}\int_{x_{2j-1}}^{x_{2j}} f(x)\mathbb{P}\big(\hat{f}_h(x) < \hat{f}_{h,\tau}\big)\,dx.$$

The main idea of the proof is that the dominant contribution to $\mathbb{E}\{\mu_f(\hat{R}_{h,\tau}\triangle R_\tau)\}$ comes from a union of $2r$ small intervals, one near each x_j, where $\mathbb{P}\big(\hat{f}_h(x) \geq \hat{f}_{h,\tau}\big)$ is close to $1/2$. In each of these intervals, we can represent $\hat{f}_h(x) - \hat{f}_{h,\tau}$ by a sample mean of independent and identically distributed random variables and a small additional remainder term, and hence apply a normal approximation to deduce the result. For clarity of exposition, we now split the proof into several steps:

Step 1: As a preliminary step, let $\tilde{f} = f + g$ be another uniformly continuous density, and let $\tilde{f}_\tau = f_\tau(\tilde{f})$. Writing $\|\cdot\|_\infty$ for the supremum norm on the real line, we show that there exists $C \geq 1$ such that for all $\epsilon > 0$ sufficiently small, we have $|\tilde{f}_\tau - f_\tau| \leq C\epsilon$ whenever $\|g\|_\infty \equiv \|\tilde{f} - f\|_\infty \leq \epsilon$. To see this, let $L = \sum_{j=1}^{r}(x_{2j} - x_{2j-1})$ and choose $C > 1 + 2L\big\{\frac{1}{4}f_\tau \sum_{j=1}^{2r} \frac{1}{|f'(x_j)|}\big\}^{-1}$. The inverse function theorem (Burkill and Burkill, 2002, Theorem 7.51) gives that for $\epsilon \in \mathbb{R}$ with

$|\epsilon|$ sufficiently small, we can write

$$\{x : f(x) \geq f_\tau + \epsilon\} = \bigcup_{j=1}^{r} [x_{2j-1} + \delta_{\epsilon,2j-1}, x_{2j} - \delta_{\epsilon,2j}],$$

with $\delta_{\epsilon,j} = \frac{\epsilon}{|f'(x_j)|} + O(\epsilon^2)$ as $\epsilon \to 0$. It follows that when $\epsilon > 0$ is sufficiently small, and $\|g\|_\infty \leq \epsilon$, we have

$$\begin{aligned}
\int_{-\infty}^{\infty} \tilde{f}(x) \mathbb{1}_{\{\tilde{f}(x) \geq f_\tau - C\epsilon\}}\, dx &\geq \int_{-\infty}^{\infty} \{f(x) - \epsilon\} \mathbb{1}_{\{f(x) \geq f_\tau - (C-1)\epsilon\}}\, dx \\
&= 1 - \tau + \sum_{j=1}^{r} \int_{x_{2j-1}+\delta_{-\epsilon(C-1),2j-1}}^{x_{2j-1}} f(x)\, dx + \sum_{j=1}^{r} \int_{x_{2j}}^{x_{2j}-\delta_{-\epsilon(C-1),2j}} f(x)\, dx \\
&\qquad - \epsilon \sum_{j=1}^{r} \{x_{2j} - \delta_{-\epsilon(C-1),2j} - (x_{2j-1} + \delta_{-\epsilon(C-1),2j-1})\} \\
&\geq 1 - \tau + \frac{1}{4}(C-1)\epsilon f_\tau \sum_{j=1}^{2r} \frac{1}{|f'(x_j)|} - 2\epsilon L > 1 - \tau.
\end{aligned}$$

Thus $\tilde{f}_\tau \geq f_\tau - C\epsilon$. A very similar argument yields the upper bound $\tilde{f}_\tau \leq f_\tau + C\epsilon$, and this completes **Step 1**.

Remark: Now, for $\delta > 0$ small enough that f has two continuous derivatives in $I_\delta \equiv \cup_{j=1}^{r}[x_{2j-1} - \delta, x_{2j} + \delta]$., let $\|\cdot\|_{I_\delta,\infty}$ denote the supremum norm restricted to I_δ. It will be helpful in **Step 4** to note that a small modification of the above argument may be used to prove that if $\|g\|_\infty$ and $\|g'\|_{I_\delta,\infty}$ are sufficiently small, and if

$$\sum_{j=1}^{r} \int_{x_{2j-1}-\delta}^{x_{2j}+\delta} |g(x)|\, dx = O\Bigl(\sum_{j=1}^{2r} |g(x_j)|\Bigr)$$

as $\sum_{j=1}^{2r} |g(x_j)| \to 0$, then $\tilde{f}_\tau - f_\tau = O\Bigl(\sum_{j=1}^{2r} |g(x_j)|\Bigr)$ as $\sum_{j=1}^{2r} |g(x_j)| \to 0$.

Step 2: We show that for each fixed $\delta > 0$,

$$\sum_{j=0}^{r} \int_{x_{2j}+\delta}^{x_{2j+1}-\delta} f(x) \mathbb{P}\bigl(\hat{f}_h(x) \geq \hat{f}_{h,\tau}\bigr)\, dx + \sum_{j=1}^{r} \int_{x_{2j-1}+\delta}^{x_{2j}-\delta} f(x) \mathbb{P}\bigl(\hat{f}_h(x) < \hat{f}_{h,\tau}\bigr)\, dx = o(n^{-1}) \tag{3.3}$$

as $n \to \infty$. In fact, we claim (and it will be straightforward to see) that the error term is of the stated order uniformly for $h \in [h^-, h^+]$. Indeed, we make a similar claim for every error term in each expression below where the bandwidth h appears, but we do not repeat this assertion in future occurrences. As in **Step 1**, observe that under (**A1**), if $\delta > 0$ is sufficiently small, then there exists $\epsilon > 0$ such that $f(x) \leq f_\tau - \epsilon$ for $x \in \cup_{j=0}^{r}[x_{2j} + \delta, x_{2j+1} - \delta]$ and $f(x) \geq f_\tau + \epsilon$ for $x \in \cup_{j=1}^{r}[x_{2j-1} + \delta, x_{2j} - \delta]$. By reducing $\delta > 0$ if necessary, for $x \in \cup_{j=0}^{r}[x_{2j} + \delta, x_{2j+1} - \delta]$,

$$\begin{aligned}
\mathbb{P}\bigl(\hat{f}_h(x) \geq \hat{f}_{h,\tau}\bigr) &\leq \mathbb{P}\bigl(\hat{f}_h(x) - f(x) - (\hat{f}_{h,\tau} - f_\tau) \geq \epsilon\bigr) \\
&\leq \mathbb{P}\bigl(\|\hat{f}_h - f\|_\infty \geq \epsilon/2\bigr) + \mathbb{P}\bigl(|\hat{f}_{h,\tau} - f_\tau| \geq \epsilon/2\bigr) \\
&\leq 2\mathbb{P}\Bigl(\|\hat{f}_h - f\|_\infty \geq \frac{\epsilon}{2C}\Bigr),
\end{aligned} \tag{3.4}$$

where we have used the result of **Step 1** in the last inequality, and C is the constant defined in that step. A very similar argument yields the same upper bound for $\mathbb{P}\bigl(\hat{f}_h(x) < \hat{f}_{h,\tau}\bigr)$ when $x \in \cup_{j=1}^{r}[x_{2j-1} + \delta, x_{2j} - \delta]$. Now, since f is uniformly continuous under (**A1**),

$$\|\mathbb{E}(\hat{f}_h) - f\|_\infty = \sup_{x \in \mathbb{R}} \biggl| \int_{-\infty}^{\infty} K(z)\{f(x - hz) - f(x)\}\, dz \biggr| \to 0 \tag{3.5}$$

as $n \to \infty$. The inequality (3.4), together with the observation (3.5) on the bias of $\hat{f}_h$, yields that for n sufficiently large,

$$\sum_{j=0}^{r} \int_{x_{2j}+\delta}^{x_{2j+1}-\delta} f(x)\mathbb{P}\big(\hat{f}_h(x) \geq \hat{f}_{h,\tau}\big)\,dx + \sum_{j=1}^{r} \int_{x_{2j-1}+\delta}^{x_{2j}-\delta} f(x)\mathbb{P}\big(\hat{f}_h(x) < \hat{f}_{h,\tau}\big)\,dx$$
$$\leq 2\mathbb{P}\Big(\|\hat{f}_h - \mathbb{E}(\hat{f}_h)\|_\infty \geq \frac{\epsilon}{4C}\Big) \leq \exp(-c_1 nh\epsilon^2),$$

for some $c_1 > 0$. Here, the final inequality is an application of Corollary 2.2 of Giné and Guillou (2002) (a consequence of Talagrand's inequality) to the Vapnik–Cervonenkis class of functions $\{K((x-\cdot)/h) : x \in \mathbb{R}, h > 0\}$ (cf. Dudley (1999), Theorems 4.2.1 and 4.2.4). Equation (3.3) follows immediately, and this completes the proof of **Step 2**.

Step 3: We show that (3.3) continues to hold if δ is replaced by a sequence (δ_n) converging to zero, provided that $\delta_n \to 0$ slowly enough that $n^{1/4}\delta_n \to \infty$ and $(h^+)^2 = o(\delta_n)$. In order to complete the proof of **Step 3**, it suffices to show that there exists $\delta > 0$ such that

$$E(\delta,\delta_n) \equiv \sum_{j=1}^{r} \int_{x_{2j-1}-\delta}^{x_{2j-1}-\delta_n} \mathbb{P}\big(\hat{f}_h(x) \geq \hat{f}_{h,\tau}\big)\,dx + \sum_{j=1}^{r} \int_{x_{2j-1}+\delta_n}^{x_{2j-1}+\delta} \mathbb{P}\big(\hat{f}_h(x) < \hat{f}_{h,\tau}\big)\,dx$$
$$\sum_{j=1}^{r} \int_{x_{2j}-\delta}^{x_{2j}-\delta_n} \mathbb{P}\big(\hat{f}_h(x) < \hat{f}_{h,\tau}\big)\,dx + \sum_{j=1}^{r} \int_{x_{2j}+\delta_n}^{x_{2j}+\delta} \mathbb{P}\big(\hat{f}_h(x) \geq \hat{f}_{h,\tau}\big)\,dx = o(n^{-1}).$$

We may assume $\delta > 0$ is small enough that f has two continuous derivatives in I_δ. This enables a straightforward modification to the argument in (3.5) using a Taylor expansion, leading to

$$\|\mathbb{E}(\hat{f}_h) - f\|_{I_\delta,\infty} = O(h^2). \tag{3.6}$$

Now there exists a constant $c_2 > 0$ small enough that if we take $\epsilon_n = c_2\delta_n$, then we have $|f(x) - f_\tau| \geq \epsilon_n$ when $\min_j |x - x_j| \geq \delta_n$. Moreover, $(h^+)^2 = o(\epsilon_n)$, so that for n sufficiently large, the same argument as in **Step 2** yields

$$E(\delta,\delta_n) \leq 2\mathbb{P}\Big(\|\hat{f}_h - \mathbb{E}(\hat{f}_h)\|_\infty \geq \frac{\epsilon_n}{4C}\Big) \leq \exp(-c_1 nh\epsilon_n^2) = o(n^{-1}).$$

This completes the proof of **Step 3**.

Step 4: We seek asymptotic expansions for $\mathbb{E}(\hat{f}_{h,\tau})$ and $\mathrm{Var}(\hat{f}_{h,\tau})$. To this end, for uniformly continuous densities $\tilde{f} = f + g$ that are twice continuously differentiable in I_δ, and for $y \in (0,\infty)$, we define

$$\psi(\tilde{f}, y) = \int_{-\infty}^{\infty} \tilde{f}(x)\mathbb{1}_{\{\tilde{f}(x)\geq y\}}\,dx.$$

The reason for making this definition is that by examining the behaviour of ψ under small changes of its arguments from (f, f_τ), we will be able to study the difference $\hat{f}_{h,\tau} - f_\tau$ in (3.10) below. First, for $\epsilon > 0$ sufficiently small,

$$\begin{aligned}
&\Big|\psi(f, f_\tau + \epsilon) - \psi(f, f_\tau) + \epsilon f_\tau \sum_{j=1}^{2r} \frac{1}{|f'(x_j)|}\Big| \\
&\qquad = \Big|-\int_{-\infty}^{\infty} f(x)\mathbb{1}_{\{f_\tau \leq f(x) < f_\tau+\epsilon\}}\,dx + \epsilon f_\tau \sum_{j=1}^{2r} \frac{1}{|f'(x_j)|}\Big| \\
&\qquad = \Big|-\sum_{j=1}^{r}\Big\{\int_{x_{2j-1}}^{x_{2j-1}+\delta_{\epsilon,2j-1}} f(x)\,dx + \int_{x_{2j}-\delta_{\epsilon,2j}}^{x_{2j}} f(x)\,dx\Big\} + \epsilon f_\tau \sum_{j=1}^{2r} \frac{1}{|f'(x_j)|}\Big| \\
&\qquad = O(\epsilon^2)
\end{aligned} \tag{3.7}$$

as $\epsilon \searrow 0$. A very similar argument shows that the error term is of the same order as $\epsilon \nearrow 0$.

Observe that when $\|g\|_\infty$ and $\|g'\|_{I_\delta,\infty}$ are sufficiently small, $\tilde{f}$ has a non-zero derivative in a neighbourhood of each x_j. It follows that for sufficiently small values of $\|g\|_\infty + \|g'\|_{I_\delta,\infty}$, we can write

$$\{x : \tilde{f}(x) \geq \tilde{f}_\tau\} = \bigcup_{j=1}^{r} [x_{2j-1} + \delta_{\epsilon,2j-1} + \eta_{2j-1} \, , \, x_{2j} - \delta_{\epsilon,2j} - \eta_{2j}],$$

where $\epsilon = \tilde{f}_\tau - f_\tau$. Moreover, provided that $\sum_{j=1}^{r} \int_{x_{2j-1}-\delta}^{x_{2j}+\delta} |g(x)| \, dx = O\bigl(\sum_{j=1}^{2r} |g(x_j)|\bigr)$ and $\sum_{j=1}^{2r} |g(x_j)| = O(\min_j |g(x_j)|)$ as $\sum_{j=1}^{2r} |g(x_j)| + \|g'\|_{I_\delta,\infty} \to 0$, we have that $\eta_j = \frac{-g(x_j)}{|f'(x_j)|} + O(|g(x_j)| \|g'\|_{I_\delta,\infty})$ as $\sum_{j=1}^{2r} |g(x_j)| + \|g'\|_{I_\delta,\infty} \to 0$. Thus we can write

$$\begin{aligned}
&\biggl|\psi(\tilde{f}, \tilde{f}_\tau) - \psi(f, \tilde{f}_\tau) - f_\tau \sum_{j=1}^{2r} \frac{g(x_j)}{|f'(x_j)|} - \sum_{j=1}^{r} \int_{x_{2j-1}}^{x_{2j}} g(x) \, dx\biggr| \\
&\qquad \leq \biggl| \int_{-\infty}^{\infty} f(x)(\mathbb{1}_{\{\tilde{f}(x) \geq \tilde{f}_\tau\}} - \mathbb{1}_{\{f(x) \geq \tilde{f}_\tau\}}) \, dx - f_\tau \sum_{j=1}^{2r} \frac{g(x_j)}{|f'(x_j)|} \biggr| \\
&\qquad\qquad\qquad\qquad + \biggl| \int_{-\infty}^{\infty} g(x)(\mathbb{1}_{\{\tilde{f}(x) \geq \tilde{f}_\tau\}} - \mathbb{1}_{\{f(x) \geq f_\tau\}}) \, dx \biggr| \\
&\qquad = \biggl| \Bigl\{ f_\tau + O\Bigl(\sum_{j=1}^{2r} |g(x_j)|\Bigr) \Bigr\} \sum_{j=1}^{2r} \Bigl\{ \frac{g(x_j)}{|f'(x_j)|} + O(|g(x_j)| \|g'\|_{I_\delta,\infty}) \Bigr\} - f_\tau \sum_{j=1}^{2r} \frac{g(x_j)}{|f'(x_j)|} \biggr| \\
&\qquad\qquad\qquad\qquad + O\Bigl\{ \Bigl(\sum_{j=1}^{2r} |g(x_j)| \Bigr)^2 \Bigr\} \\
&\qquad = O\Bigl\{ \Bigl(\sum_{j=1}^{2r} |g(x_j)| \Bigr)^2 + \|g'\|_{I_\delta,\infty} \sum_{j=1}^{2r} |g(x_j)| \Bigr\} \qquad (3.8)
\end{aligned}$$

as $\sum_{j=1}^{2r} |g(x_j)| + \|g'\|_{I_\delta,\infty} \to 0$. Assuming that $\psi(\tilde{f}, \tilde{f}_\tau) = 1 - \tau$ and that the above conditions on g hold, we have from (3.7) and (3.8) that

$$\begin{aligned}
0 &= \psi(\tilde{f}, \tilde{f}_\tau) - \psi(f, f_\tau) \\
&= \psi(\tilde{f}, \tilde{f}_\tau) - \psi(f, \tilde{f}_\tau) + \psi(f, \tilde{f}_\tau) - \psi(f, f_\tau) \\
&= -\{\tilde{f}_\tau - f_\tau\} f_\tau \sum_{j=1}^{2r} \frac{1}{|f'(x_j)|} + f_\tau \sum_{j=1}^{2r} \frac{g(x_j)}{|f'(x_j)|} + \sum_{j=1}^{r} \int_{x_{2j-1}}^{x_{2j}} g(x) \, dx \\
&\qquad\qquad + O\Bigl\{ \Bigl(\sum_{j=1}^{2r} |g(x_j)| \Bigr)^2 + \|g'\|_{I_\delta,\infty} \sum_{j=1}^{2r} |g(x_j)| \Bigr\} \qquad (3.9)
\end{aligned}$$

as $\sum_{j=1}^{2r} |g(x_j)| + \|g'\|_{I_\delta,\infty} \to 0$.

We want to apply (3.9) with $\tilde{f} = \hat{f}_h$, so that $g = \hat{f}_h - f$. In order to do this, we must recall observation (3.5) on the bias of $\hat{f}_h$, and the fact that $\|\hat{f}_h - \mathbb{E}(\hat{f}_h)\|_\infty = O_{a.s.}\bigl(\frac{\sqrt{\log 1/h}}{(nh)^{1/2}}\bigr)$ from an application of Corollary 2.2 of Giné and Guillou (2002). It follows that $\|\hat{f}_h - f\|_\infty \overset{a.s.}{\to} 0$. Similarly, $\|\mathbb{E}(\hat{f}'_h) - f'\|_{I_\delta,\infty} = O(h^2)$, and a further application of Corollary 2.2 of Giné and Guillou (2002) gives $\|\hat{f}'_h - \mathbb{E}(\hat{f}'_h)\|_{I_\delta,\infty} = O_{a.s.}\bigl(\frac{\sqrt{\log 1/h}}{(nh^3)^{1/2}}\bigr)$. Thus $\|\hat{f}'_h - f'\|_{I_\delta,\infty} \overset{a.s.}{\to} 0$. This in turn implies that with probability one, for n sufficiently large, $\hat{f}_{h,\tau}$ is the unique solution to $\psi(\hat{f}_h, \hat{f}_{h,\tau}) = 1 - \tau$, or equivalently $\int \hat{f}_h(x)\mathbb{1}_{\{\hat{f}_h(x)\geq \hat{f}_{h,\tau}\}}\,dx = 1-\tau$, as claimed in Section 2. It remains to note that

$$\frac{\sum_{j=1}^{r}\int_{x_{2j-1}-\delta}^{x_{2j}+\delta}|\hat{f}_h(x)-f(x)|\,dx}{\sum_{j=1}^{2r}|\hat{f}_h(x_j)-f(x_j)|} = O_p(1) \quad \text{and} \quad \frac{\sum_{j=1}^{2r}|\hat{f}_h(x_j)-f(x_j)|}{\min_j|\hat{f}_h(x_j)-f(x_j)|} = O_p(1).$$

It follows that we can now substitute $g = \hat{f}_h - f$ in (3.9) to deduce that

$$\hat{f}_{h,\tau} - f_\tau = \biggl\{\sum_{j=1}^{2r}\frac{1}{|f'(x_j)|}\biggr\}^{-1}\biggl\{\sum_{j=1}^{2r}\frac{\hat{f}_h(x_j)-f(x_j)}{|f'(x_j)|} + \frac{1}{f_\tau}\sum_{j=1}^{r}\int_{x_{2j-1}}^{x_{2j}}\hat{f}_h(x)-f(x)\,dx\biggr\} + O_p\Bigl(\frac{\sqrt{\log(1/h)}}{nh^2} + \frac{h^{1/2}\sqrt{\log(1/h)}}{n^{1/2}} + h^4\Bigr). \quad (3.10)$$

Equation (3.10) shows that we can write the difference $\hat{f}_{h,\tau} - f_\tau$ as a sample mean of independent and identically distributed random variables and a small additional remainder term. Notice from the bandwidth condition on h^- in **(A2)** that $\frac{\sqrt{\log(1/h)}}{nh^2} = o(h^2)$. Next, observe that

$$\sum_{j=1}^{2r}\frac{\mathbb{E}\{\hat{f}_h(x_j)\}-f(x_j)}{|f'(x_j)|} + \frac{1}{f_\tau}\sum_{j=1}^{r}\int_{x_{2j-1}}^{x_{2j}}\mathbb{E}\{\hat{f}_h(x)\}-f(x)\,dx = D_1\sum_{j=1}^{2r}\frac{1}{|f'(x_j)|}h^2 + o(h^2),$$

where D_1 is given in (2.1). Thus, in order to prove that

$$\mathbb{E}(\hat{f}_{h,\tau}) = f_\tau + D_1h^2 + o(h^2), \quad (3.11)$$

it suffices by (3.9) and **Step 1** to show that for any $\eta > 0$,

$$\mathbb{E}(|\hat{f}_{h,\tau} - f_\tau - D_1h^2|\mathbb{1}_{\{\sum_{j=1}^{2r}|\hat{f}_h(x_j)-f(x_j)|+\|\hat{f}_h'-f'\|_{I_\delta,\infty}>\eta\}}) = o(h^2).$$

But this follows by Cauchy–Schwarz, because **Step 1** may be used to show that $\mathbb{E}(\hat{f}_{h,\tau}^2) = O(1)$, and also

$$\mathbb{P}\Bigl(\sum_{j=1}^{2r}|\hat{f}_h(x_j)-f(x_j)| > \eta/2\Bigr) + \mathbb{P}\bigl(\|\hat{f}_h' - f'\|_{I_\delta,\infty} > \eta/2\bigr) = o(n^{-1}).$$

We therefore deduce (3.11).

In a very similar way, we can also use (3.9) and the fact that

$$\sum_{j=1}^{2r}\frac{\mathrm{Var}\{\hat{f}_h(x_j)\}}{f'(x_j)^2} = \frac{D_2}{nh}\biggl\{\sum_{j=1}^{2r}\frac{1}{|f'(x_j)|}\biggr\}^2 + o\Bigl(\frac{1}{nh}\Bigr),$$

where D_2 is given in (2.1), to deduce that

$$\mathrm{Var}(\hat{f}_{h,\tau}) = \frac{D_2}{nh} + o\Bigl(\frac{1}{nh}\Bigr). \quad (3.12)$$

Step 5: We can use the results of **Step 4** to shrink the region of interest still further. From the result of **Step 3** we can write

$$\begin{aligned}\mathbb{E}\{\mu_f(\hat{R}_{h,\tau}\triangle R_\tau)\} &= \sum_{j=1}^{r}\int_{x_{2j-1}-\delta_n}^{x_{2j-1}+\delta_n} f(x)\bigl|\mathbb{P}\bigl(\hat{f}_h(x) < \hat{f}_{h,\tau}\bigr) - \mathbb{1}_{\{x<x_{2j-1}\}}\bigr|\,dx \\ &\quad + \sum_{j=1}^{r}\int_{x_{2j}-\delta_n}^{x_{2j}+\delta_n} f(x)\bigl|\mathbb{P}\bigl(\hat{f}_h(x) < \hat{f}_{h,\tau}\bigr) - \mathbb{1}_{\{x\geq x_{2j}\}}\bigr|\,dx + o(n^{-1}) \\ &= \frac{f_\tau}{(nh)^{1/2}}\sum_{j=1}^{r}\int_{-(nh)^{1/2}\delta_n}^{(nh)^{1/2}\delta_n}|\mathbb{P}\bigl(\hat{f}_h(x_{2j-1}+(nh)^{-1/2}t) < \hat{f}_{h,\tau}\bigr) - \mathbb{1}_{\{t<0\}}| \\ &\quad + \bigl|\mathbb{P}\bigl(\hat{f}_h(x_{2j}+(nh)^{-1/2}t) < \hat{f}_{h,\tau}\bigr) - \mathbb{1}_{\{t\geq 0\}}\bigr|\,dt + o(n^{-1}). \quad (3.13)\end{aligned}$$

For brevity, we write $x_j^t = x_j + (nh)^{-1/2}t$. Now, for each $j = 1, \ldots, 2r$, we see that for n sufficiently large, $\mathbb{E}\{\hat{f}_h(x_j^t) - \hat{f}_{h,\tau}\}$ is a strictly monotone function of $t \in [-(nh)^{1/2}\delta_n, (nh)^{1/2}\delta_n]$, with a unique zero t_j^*, say. Moreover,

$$t_j^* = \Big\{D_1 - \frac{1}{2}\mu_2(K)f''(x_j)\Big\}\{f'(x_j)\}^{-1}n^{1/2}h^{5/2}\{1 + o(1)\}.$$

Fix a sequence (t_n) diverging to infinity and let $I_j^n = [-(nh)^{1/2}\delta_n, (nh)^{1/2}\delta_n] \setminus [t_j^* - t_n, t_j^* + t_n]$. We claim that

$$\sum_{j=1}^{r}\bigg\{\int_{I_{2j-1}^n}|\mathbb{P}\big(\hat{f}_h(x_{2j-1}^t) < \hat{f}_{h,\tau}\big) - \mathbb{1}_{\{t<0\}}|\,dt + \int_{I_{2j}^n}\big|\mathbb{P}\big(\hat{f}_h(x_{2j}^t) < \hat{f}_{h,\tau}\big) - \mathbb{1}_{\{t\geq 0\}}\big|\,dt\bigg\} \to 0 \quad (3.14)$$

as $n \to \infty$. Now there exists $c_3 > 0$ such that for all $t \in \cup_{j=1}^{2r} I_j^n$ and n sufficiently large, we have $|\mathbb{E}\{\hat{f}_h(x_j^t) - \hat{f}_{h,\tau}\}| \geq c_3(nh)^{-1/2}t_n$. Thus there exists $c_4 > 0$ such that for all n sufficiently large,

$$\begin{aligned}|\mathbb{P}\big(&\hat{f}_h(x_{2j-1}^t) < \hat{f}_{h,\tau}\big) - \mathbb{1}_{\{t<0\}}| \\ &\leq \mathbb{P}\bigg(\bigg|\frac{\hat{f}_h(x_{2j-1}^t) - \mathbb{E}\{\hat{f}_h(x_{2j-1}^t)\}}{\mathrm{Var}^{1/2}\{\hat{f}_h(x_{2j-1}^t)\}}\bigg| \geq c_4 t_n\bigg) + \mathbb{P}\bigg(\bigg|\frac{\hat{f}_{h,\tau} - \mathbb{E}(\hat{f}_{h,\tau})}{\mathrm{Var}^{1/2}(\hat{f}_{h,\tau})}\bigg| \geq c_4 t_n\bigg) \to 0,\end{aligned}$$

uniformly for $t \in \cup_{j=1}^r I_{2j-1}^n$. Since also $\big|\mathbb{P}\big(\hat{f}_h(x_{2j}^t) < \hat{f}_{h,\tau}\big) - \mathbb{1}_{\{t\geq 0\}}\big| \to 0$ uniformly for $t \in \cup_{j=1}^r I_{2j}^n$, we deduce (3.14).

Step 6: We also require an asymptotic expansion for $\mathrm{Cov}(\hat{f}_h(x_j^t), \hat{f}_{h,\tau})$, for $t \in [t_j^* - t_n, t_j^* + t_n]$. In fact, provided (t_n) diverges sufficiently slowly, we have

$$\mathrm{Cov}(\hat{f}_h(x_j^t), \hat{f}_{h,\tau}) = \frac{D_{3,j}}{nh} + o\Big(\frac{1}{nh}\Big),$$

uniformly for $t \in [t_j^* - t_n, t_j^* + t_n]$, where $D_{3,j}$ is given at (2.1). This follows from the expansion (3.9) and the fact that provided (t_n) diverges sufficiently slowly,

$$\begin{aligned}\mathbb{E}\Big\{\frac{1}{h^2}K\Big(\frac{x_j - X_1}{h}\Big)K\Big(\frac{x_j^t - X_1}{h}\Big)\Big\} &= \frac{1}{h}\int_{-\infty}^{\infty} K(z)K\Big(\frac{(nh)^{-1/2}t + hz}{h}\Big)f(x_j - hz)\,dz \\ &= \frac{1}{h}f_\tau R(K) + o(h^{-1}),\end{aligned}$$

uniformly for $t \in [t_j^* - t_n, t_j^* + t_n]$.

Step 7: To complete the proof of Theorem 1, it suffices by (3.13) and (3.14) to show that there exists a sequence (t_n) diverging to infinity such that

$$\begin{aligned}\frac{f_\tau}{(nh)^{1/2}}\sum_{j=1}^{r}\bigg\{\int_{t_{2j-1}^*-t_n}^{t_{2j-1}^*+t_n}&|\mathbb{P}\big(\hat{f}_h(x_{2j-1}^t) < \hat{f}_{h,\tau}\big) - \mathbb{1}_{\{t<0\}}|\,dt + \int_{t_{2j}^*-t_n}^{t_{2j}^*+t_n}\big|\mathbb{P}\big(\hat{f}_h(x_{2j}^t) < \hat{f}_{h,\tau}\big) - \mathbb{1}_{\{t\geq 0\}}\big|\,dt\bigg\} \\ &= \sum_{j=1}^{2r}\bigg[\frac{B_{1,j}\phi(B_{2,j}n^{1/2}h^{5/2})}{(nh)^{1/2}} + B_{3,j}h^2\{2\Phi(B_{2,j}n^{1/2}h^{5/2}) - 1\}\bigg] + o\Big(\frac{1}{(nh)^{1/2}} + h^2\Big).\end{aligned}$$

For $i = 1, \ldots, n$, let $Z_{ni}(x) = h^{-1}K\big(\frac{x - X_i}{h}\big)$ and let $\bar{Y}_n = n^{-1}\sum_{i=1}^n Y_{ni}$, where

$$Y_{ni} = Z_{ni}(x_j^t) - f_\tau - \bigg\{\sum_{k=1}^{2r}\frac{1}{|f'(x_k)|}\bigg\}^{-1}\bigg[\sum_{k=1}^{2r}\frac{Z_{ni}(x_k) - f(x_k)}{|f'(x_k)|} + \frac{1}{f_\tau}\sum_{k=1}^{r}\int_{x_{2k-1}}^{x_{2k}} Z_{ni}(x) - f(x)\,dx\bigg].$$

By (3.9) and (3.10), we can write $\hat{f}_h(x_j^t) - \hat{f}_{h,\tau} = \bar{Y}_n + R_n$, where $R_n - \mathbb{E}(R_n) = o_p\{(nh)^{-1/2}\}$. Since $\mathrm{Var}(\bar{Y}_n) = O\{(nh)^{-1}\}$ uniformly for $t \in [t_j^* - t_n, t_j^* + t_n]$, we choose (t_n) to diverge to infinity so slowly that

- $\mathbb{P}\Bigl(\frac{|R_n-\mathbb{E}(R_n)|}{\mathrm{Var}^{1/2}(\bar{Y}_n)} > \frac{1}{t_n^2}\Bigr) \leq \frac{1}{t_n^2}$, uniformly for $t \in [t_j^* - t_n, t_j^* + t_n]$
- $(nh)\mathrm{Var}(\bar{Y}_n) = R(K)f_\tau - 2D_{3,j} + D_2 + o(t_n^{-1})$, uniformly for $t \in [t_j^* - t_n, t_j^* + t_n]$
- $\mathbb{E}(\bar{Y}_n + R_n) = \{(nh)^{-1/2}tf'(x_j) + D_4h^2\}\{1 + o(t_n^{-1})\}$, uniformly for $t \in [t_j^* - t_n, t_j^* + t_n]$, where $D_4 = \frac{1}{2}\mu_2(K)f''(x_j) - D_1$
- $t_n = o(n^{1/6})$.

Then

$$\begin{aligned}
&\mathbb{P}\bigl(\hat{f}_h(x_j^t) < \hat{f}_{h,\tau}\bigr) - \Phi\biggl(\frac{-tf'(x_j) - D_4n^{1/2}h^{5/2}}{\{R(K)f_\tau - 2D_{3,j} + D_2\}^{1/2}}\biggr)\\
&\quad \leq \mathbb{P}\biggl(\frac{|R_n - \mathbb{E}(R_n)|}{\mathrm{Var}^{1/2}(\bar{Y}_n)} > \frac{1}{t_n^2}\biggr) + \mathbb{P}\biggl(\frac{\bar{Y}_n - \mathbb{E}(\bar{Y}_n)}{\mathrm{Var}^{1/2}(\bar{Y}_n)} \leq \frac{-\mathbb{E}(\bar{Y}_n + R_n)}{\mathrm{Var}^{1/2}(\bar{Y}_n)} + \frac{1}{t_n^2}\biggr)\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad - \Phi\biggl(\frac{-tf'(x_j) - D_4n^{1/2}h^{5/2}}{\{R(K)f_\tau - 2D_{3,j} + D_2\}^{1/2}}\biggr)\\
&\quad = O\Bigl(\frac{1}{t_n^2} + \frac{1}{(nh)^{1/2}}\Bigr) + \Phi\biggl(\frac{-\mathbb{E}(\bar{Y}_n + R_n)}{\mathrm{Var}^{1/2}(\bar{Y}_n)}\biggr) - \Phi\biggl(\frac{-tf'(x_j) - D_4n^{1/2}h^{5/2}}{\{R(K)f_\tau - 2D_{3,j} + D_2\}^{1/2}}\biggr)\\
&\quad = o(t_n^{-1}),
\end{aligned}$$

uniformly for $t \in [t_j^* - t_n, t_j^* + t_n]$. Here we have used the Berry–Esseen inequality to reach the penultimate line. A very similar argument yields a lower bound of the same order. The proof of **Step 7**, and hence the proof of Theorem 1, is now completed by the observation that

$$\begin{aligned}
\frac{f_\tau}{(nh)^{1/2}}\sum_{j=1}^{r}\biggl\{\int_{-\infty}^{\infty}\biggl|\Phi\biggl(\frac{-tf'(x_{2j-1}) - D_4n^{1/2}h^{5/2}}{\{R(K)f_\tau - 2D_{3,j} + D_2\}^{1/2}}\biggr) - \mathbb{1}_{\{t<0\}}\biggr|&\\
+ \biggl|\Phi\biggl(\frac{-tf'(x_{2j}) - D_4n^{1/2}h^{5/2}}{\{R(K)f_\tau - 2D_{3,j} + D_2\}^{1/2}}\biggr) - \mathbb{1}_{\{t\geq0\}}\biggr|\,dt\biggr\}&\\
= \sum_{j=1}^{2r}\biggl[\frac{B_{1,j}\phi(B_{2,j}n^{1/2}h^{5/2})}{(nh)^{1/2}} + B_{3,j}h^2\{2\Phi(B_{2,j}n^{1/2}h^{5/2}) - 1\}\biggr]&.
\end{aligned}$$

□

Proof of Corollary 2

We know that for any sequence of bandwidths minimising the risk, nh^5 is bounded away from zero and infinity, so we may restrict attention to this case. The important point to note is that under the hypotheses of the corollary, $B_{1,j}$, $B_{2,j}$ and $B_{3,j}$ do not depend on j, so we write them as B_1, B_2 and B_3 respectively.

By making the substitution $x = B_2n^{1/2}h^{5/2}$, there exist positive constants $a = 2B_1B_2^{1/5}$ and $b = B_3/(B_1B_2)$ such that $\lim_{n\to\infty} n^{2/5}\mathbb{E}\{\mu_f(\hat{R}_{h,\tau}\triangle R_\tau)\} = au(x)$, where $u(x) = x^{-1/5}\phi(x) + bx^{4/5}\{2\Phi(x) - 1\}$. Since u is continuous with $u(x) \to \infty$ as $x \searrow 0$ and $x \to \infty$, it attains its minimum in $(0, \infty)$. To show this minimum is unique, it suffices to show that $v(x)$ has a unique zero in $(0, \infty)$, where

$$v(x) = \frac{5x^{6/5}}{\phi(x)}u'(x) = -1 + \frac{4bx\{2\Phi(x) - 1\}}{\phi(x)} + 5(2b - 1)x^2.$$

Now we have

$$v'(x) = 2(14b-5)x + \frac{4b(1+x^2)\{2\Phi(x)-1\}}{\phi(x)}$$
$$v''(x) = 2(18b-5) + 8bx^2 + \frac{4b(3x+x^3)\{2\Phi(x)-1\}}{\phi(x)}.$$

There are therefore two cases to consider: if $b \geq 5/18$, then v is strictly convex, so since $v(0+) = -1$ and $v(x) \to \infty$ as $x \to \infty$, we see that v has a unique zero in $(0, \infty)$. On the other hand, if $b < 5/18$, then there exists $x^* \in (0, \infty)$ such that $v''(x) < 0$ for $x \in (0, x^*)$ and $v''(x) > 0$ for $x \in (x^*, \infty)$. But if $b < 5/18$ then $v'(x) < 0$, for sufficiently small $x > 0$, so from $v(0+) = -1$, it again follows that v has a unique zero.

Write $x_{\min}$ for the unique minimum of u in $(0, \infty)$, and let $c_{\mathrm{opt}} = (x_{\min}/B_2)^{2/5}$. We conclude that any optimal bandwidth sequence (h_{opt}), in the sense of minimising $\mathbb{E}\{\mu_f(\hat{R}_{h,\tau} \triangle R_\tau)\}$, must satisfy $h_{\mathrm{opt}} = c_{\mathrm{opt}} n^{-1/5}\{1 + o(1)\}$ as $n \to \infty$. □

Proof of Theorem 3

We require a bound on $|\hat{x}_{j,h_0} - x_j|$ for $j = 1, \ldots, 2r$. To this end, let $\tilde{f} = f + g$ be another density satisfying the same conditions as f. From **Step 4** of the proof of Theorem 1, we see that for sufficiently small values of $\|g\|_\infty + \|g'\|_{I_\delta,\infty}$, there exist precisely $2r$ values $\tilde{x}_1 < \ldots < \tilde{x}_{2r}$ such that $\tilde{f}(\tilde{x}_j) = \tilde{f}_\tau$. Moreover, provided $\sum_{j=1}^r \int_{x_{2j-1}-\delta}^{x_{2j}+\delta} |g(x)|\,dx = O\bigl(\sum_{j=1}^{2r} |g(x_j)|\bigr)$ as $\sum_{j=1}^{2r} |g(x_j)| + \|g'\|_{I_\delta,\infty} \to 0$, we have $\tilde{x}_j - x_j = O(|g(x_j)|)$ as $\sum_{j=1}^{2r} |g(x_j)| + \|g'\|_{I_\delta,\infty} \to 0$. Substituting $\tilde{f} = \hat{f}_{h_0}$, so that $g = \hat{f}_{h_0} - f$ and $\tilde{x}_j = \hat{x}_{j,h_0}$, we have $|\hat{x}_{j,h_0} - x_j| = O_p(n^{-2/5})$.

It follows that $\hat{D}_1 = D_1 + O_p(n^{-2/9})$, the crucial fact being that $\hat{f}''_{h_2}(\hat{x}_{j,h_0}) - f''(x_j) = O_p(n^{-2/9})$. Similarly, $\hat{D}_2 = D_2 + O_p(n^{-2/7})$ and $\hat{D}_{3,j} = D_{3,j} + O_p(n^{-2/7})$ for $j = 1, \ldots, 2r$. Thus $\hat{B}_{1,j} = B_{1,j} + O_p(n^{-2/7})$, $\hat{B}_{2,j} = B_{2,j} + O_p(n^{-2/9})$ and $\hat{B}_{3,j} = B_{3,j} + O_p(n^{-2/9})$. We deduce that for any $0 < c_1 < c_2 < \infty$, we have $\widehat{AR}_n(c) = AR(c)\bigl\{1 + O_p(n^{-2/9})\bigr\}$, uniformly for $c \in [c_1, c_2]$, and a standard Taylor expansion argument then gives that $\hat{c}_{\mathrm{opt}} = c_{\mathrm{opt}}\bigl\{1 + O_p(n^{-2/9})\bigr\}$. Both conclusions of the theorem follow immediately. □

Proof of Theorem 4

Let $z_n = \delta/h_2$, where δ is small enough that f has 12 continuous derivatives in $\cup_{j=1}^{2r}[x_j - \delta, x_j + \delta]$. Under the conditions of the theorem, we may integrate by parts twice and apply a Taylor expansion to obtain

$$|\mathbb{E}\{\hat{f}''_{h_2}(x_j)\} - f''(x_j)| = \left| \int_{-z_n}^{z_n} K_2(z)\{f''(x_j - h_2 z) - f''(x_j)\}\,dz \right| + o(h_2^{10}) = O(h_2^{10}).$$

This expression for the bias can be combined with the standard fact that $\mathrm{Var}\hat{f}''_{h_2}(x_j) = O\{(nh_2^5)^{-1}\}$ and the bound on $|\hat{x}_{j,h_0} - x_j|$ from the proof of Theorem 3 to yield $\hat{f}''_{h_2}(\hat{x}_{j,h_0}) - f''(x_j) = O_p(n^{-2/5})$. Similar computations give $\hat{f}'_{h_1}(\hat{x}_{j,h_0}) - f'(x_j) = O_p(n^{-2/5})$. The rest of the proof mirrors the proof of Theorem 3. □

Acknowledgments

This research was partially supported by Australian Research Council grant DP055651. The authors are grateful to Tarn Duong, Inge Koch, Steve Marron and Richard Nickl for their comments on aspects of this research, and to the organisers of a workshop on statistical research held at the Keystone Resort, Colorado, USA, on 4th-8th June, 2007.

References

Baíllo, A. (2003). Total error in a plug-in estimator of level sets. *Statistics and Probability Letters*, **65**, 411–417.

Baíllo, A., Cuesta-Albertos, J. and Cuevas, A. (2001). Convergence rates in nonparametric estimation of level sets. *Statistics and Probability Letters*, **53**, 27–35.

Bowman, A.W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, **71**, 353–60.

Burkill, J.C. and Burkill, H. (2002). *A Second Course in Mathematical Analysis.* Cambridge: Cambridge University Press.

Cadre, B. (2006). Kernel estimation of density level sets. *Journal of Multivariate Analysis*, **97**, 999–1023.

Dudley, R. (1999). *Uniform Central Limit Theorems*, Cambridge: Cambridge University Press.

Giné, E. and Guillou, A. (2002). Rates of strong uniform consistency for multivariate kernel density estimators. *Annals of the Institute of Henri Poincaré– Probability and Statistics*, **6**, 907–921.

González-Manteiga, W., Sanchéz-Sellero, C. and Wand, M. P. (1996) Accuracy of binned kernel functional approximations. *Computational Statistics and Data Analysis*, **22**, 1–16.

Hartigan, J.A. (1987). Estimation of a convex density contour in two dimensions. *Journal of the American Statistical Association*, **82**, 267–270.

Hyndman, R.J. (1996). Computing and graphing highest density regions. *The American Statistician*, **50**, 120–126.

Hyndman, R.J. and Einbeck, J. (2007). `hdrcde`. Highest density regions and conditional density estimation. R package. `http://cran.r-project.org`.

Jang, W. (2006). Nonparametric density estimation and clustering in astronomical sky surveys. *Computational Statistics and Data Analysis*, **50**, 760–774.

Marron, J. S. and Wand, M. P. (1992). Exact mean integrated squared error. *The Annals of*

Statistics, **20**, 712–736 .

Mason, D. M. and Polonik, W. (2009). Asymptotic normality of plug-in level set estimates. *Annals of Applied Probability*, **19**, 1108–1142.

Müller, D.W. and Sawitzki, G. (1991). Excess mass estimates and tests for multimodality. *Journal of the American Statistical Association*, **86**, 738–746.

Park, B.U. and Marron, J.S. (1990) Comparison of data-driven bandwidth selectors. *Journal of the American Statistical Association*, **85**, 66–72.

Polonik, W. (1995). Measuring mass concentrations and estimating density contour clusters – an excess mass approach. *The Annals of Statistics*, **23**, 855–881.

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL `http://www.R-project.org`.

Rigollet, P. and Vert, R. (2009). Optimal rates for plug-in estimators of density level sets. *Bernoulli*, to appear.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, **9**, 65–78.

Sheather, S.J. and Jones, M.C. (1991) A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, Series B*, **53**, 683–690.

Tsybakov, A. B. (1997). On nonparametric estimation of density level sets. *The Annals of Statistics*, **25**, 948–969.

Wand, M.P. and Jones, M.C. (1995). *Kernel Smoothing*, Boca Raton, Florida: Chapman and Hall, CRC Press.