CHOOSING A KERNEL FOR CROSS-VALIDATION

A Dissertation

by

OLGA SAVCHUK

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2009

Major Subject: Statistics

CHOOSING A KERNEL FOR CROSS-VALIDATION

A Dissertation

by

OLGA SAVCHUK

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Co-Chairs of Committee,	Jeffrey D. Hart
	Simon J. Sheather
Committee Members,	Qi Li
	Suhasini Subba Rao
Head of Department,	Simon J. Sheather

August 2009

Major Subject: Statistics

ABSTRACT

Choosing A Kernel for Cross-Validation. (August 2009)

Olga Savchuk, B.S., National Technical University of Ukraine;

M.S., National Technical University of Ukraine;

M.S., Texas A&M University

Co–Chairs of Advisory Committee: Dr. Jeffrey D. Hart
Dr. Simon J. Sheather

The statistical properties of cross-validation bandwidths can be improved by choosing an appropriate kernel, which is different from the kernels traditionally used for cross-validation purposes. In the light of this idea, we developed two new methods of bandwidth selection termed: Indirect cross-validation and Robust one-sided cross-validation. The kernels used in the Indirect cross-validation method yield an improvement in the relative bandwidth rate to $n^{-1/4}$, which is substantially better than the $n^{-1/10}$ rate of the least squares cross-validation method. The robust kernels used in the Robust one-sided cross-validation method eliminate the bandwidth bias for the case of regression functions with discontinuous derivatives.

ACKNOWLEDGMENTS

First of all, I would like to express my warm gratitude to my advisors, Dr. Jeffrey D. Hart and Dr. Simon J. Sheather for their help, support, enthusiasm, and interest in my work. I am very fortunate to have so prominent statisticians as my advisors. Let me devote a paragraph to each of them.

Dr. Hart is an outstanding statistician in many fields of statistics, including nonparametric statistics. Moreover, Dr. Hart is an excellent teacher. I am very grateful to Dr. Hart for his willingness to help me. All the five years of my being at Texas A&M I knew that I could turn to Dr. Hart for his advice or help concerning anything which was important to me. I am also very thankful to Dr. Hart for careful reading and correcting the manuscript.

Dr. Sheather is a leader in the nonparametric statistics field. I highly appreciate Dr. Sheather's ability to organize work. Because of Dr. Sheather's talent to achieve the goals, we were able to complete our main research project, Indirect Cross-Validation, fairly fast and with a reasonable amount of effort.

Besides, there are many other people to whom I am grateful for their impact on my life. I would really like to acknowledge my very first academic advisor at National Technical University of Ukraine, Dr. Alexandr Krasilnikov, who first brought me to the world of statistics. I wish to thank Dr. Daren Cline for his excellent courses, especially Stat 614 "Advanced Probability Theory". I appreciate the friendship and help of my classmates, Mandy Hering, Dongling Zhan, and Beverly Gaucher.

Most importantly, I wish to express gratitude to my husband, Dmytro Savchuk, for his enormous help, support and understanding. Without his contribution this dissertation would not have been possible. I also thank my little daughters, Anna and Irina, for bringing so much light and love to my life.

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

TABLE Page

LIST OF FIGURES

FIGURE Page

FIGURE Page

FIGURE Page

CHAPTER I

INTRODUCTION

Any experimenter taking measurements is very likely to face the problem of estimating either a regression function or a probability density function. At best, one may have only a vague idea about qualitative aspects of the function which needs to be estimated. A nonparametric approach suggests that the data should themselves decide which function provides them the best fit without the restrictions imposed by a parametric model.

Kernel methods play an important role in nonparametric function estimation. Kernel density estimation (KDE) was introduced by Rosenblatt (1956) and Parzen (1962) and is one of the most often used probability density estimation methods. Among kernel regression methods the most popular are the local linear estimator (see Fan (1992)), the Gasser-Müller estimator (see Gasser and Müller (1979)), and the Nadaraya-Watson estimator (see Nadaraya (1964) and Watson (1964)). All of the previously mentioned methods require selecting a smoothing parameter, which is usually called the *bandwidth* and is denoted by h.

The choice of h controls the smoothness of the estimator, as the following example illustrates. For this example we generated a sample of size $n = 100$ from a bimodal density and computed kernel density estimates using the three values of h, 0.15, 0.385, and 0.8. The resulting density estimates are shown in Figure 1. The estimate based on $h = 0.15$ is apparently undersmoothed: it is very wiggly and shows several false modes. The estimate based on $h = 0.8$ is oversmoothed: it does not show the bimodal structure of the density. The estimate based on $h = 0.385$ is fairly close to

---

This dissertation follows the style of *Journal of the American Statistical Association.*

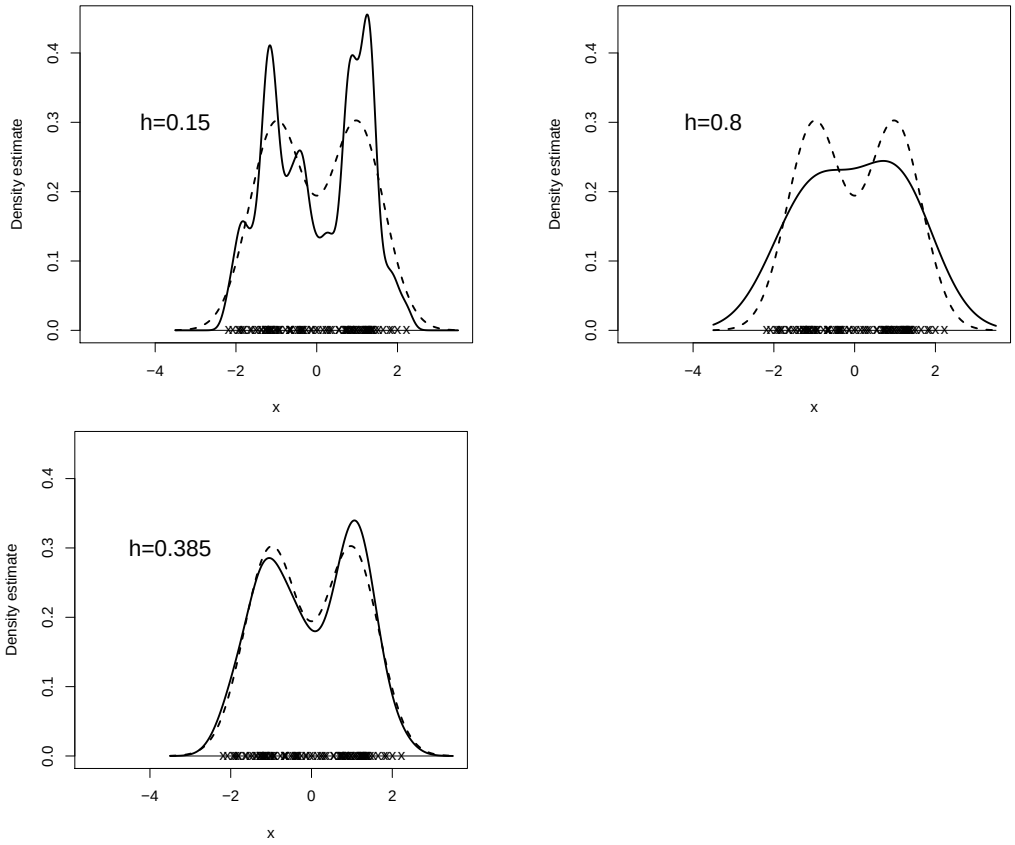


Fig. 1. Density estimates for a bimodal density. Dashed line shows the true density; crosses on the horizontal axes show the data points.

the true density. The bandwidth $h = 0.385$ minimizes the mean integrated squared error (MISE) for the kernel density estimator, which is a popular error criterion in the density and regression estimation problems. The MISE-optimal bandwidth can be computed only if the true function is known, so we will not able to find it for the real data sets. There are many methods that estimate the bandwidth which minimizes MISE or some other optimality criterion. Such bandwidth selection methods are called *data-driven*.

One of the most popular methods of data-driven bandwidth selection is least squares cross-validation (LSCV). The general idea of cross-validation is explained by

the following three steps. First, estimate various models from a subset of the data. Next, predict the remaining observations with each fitted model. Finally, choose the model that minimizes average squared prediction error. In the kernel density estimation and kernel regression contexts different models correspond to different values of h, and each time we predict an observation using all the data points excluding that observation. An estimator constructed from all the data except for one point is called a *leave-one-out estimator*.

There are versions of the LSCV method for the kernel density estimation problem (see Rudemo (1982) and Bowman (1984)) and for the nonparametric regression problem (see Härdle, Hall, and Marron (1988)). The LSCV method is completely automatic, widely used, easy to implement, and is known to perform well on functions which are difficult to estimate (see Loader (1999b), and van Es (1992)). The main drawbacks of the LSCV method are high variability of the selected bandwidths, and a very slow relative convergence rate of the order $n^{-1/10}$.

Many modifications of LSCV have been proposed in an attempt to improve its performance. These include the trimmed cross-validation (TCV) method of Feluch and Koronacki (1992), and the one-sided cross-validation (OSCV) method of Hart and Yi (1998) and Marti'nez Miranda, Nielsen, and Sperlich (2009). The essence of the TCV and OSCV methods is to modify the kernel used to compute the leave-one-out estimator (we will simply call it the *cross-validation kernel*) in order to improve the statistical properties of the cross-validation bandwidths. In particular, the TCV method uses a "trimmed" kernel, obtained from a traditional nonnegative unimodal kernel K by multiplying it by an indicator function, which results in cutting out the center of K. The OSCV method uses a so-called one-sided kernel, which is obtained from K by multiplying it by a line and restricting the kernel's support to nonnegative values. In this research we exploit the idea of transforming the cross-validation kernels

further and propose new modifications of the LSCV method.

In Chapter II we describe a new method of bandwidth selection for kernel density estimation, which is called Indirect Cross-Validation (ICV). Our initial work on the ICV method was inspired by the TCV method, in the sense that we constructed a family of cross-validation kernels by cutting out the middle of the Gaussian kernel. Further research in this direction lead to another family of kernels which are more efficient for cross-validation purposes. The major advance for the ICV method is that it improves the relative bandwidth rate to $n^{-1/4}$. The ICV method outperforms the LSCV method in a simulation study and examples.

Chapter III is independent of Chapter II and is dedicated to extending the OSCV method to the case of nonsmooth regression functions. Subsequently, when we refer to a *nonsmooth* function, we mean a continuous function whose first derivative is bounded and may have a finite number of discontinuities. The points at which the derivative is discontinuous are often called *cusps*. One of the proposed modifications of the ordinary OSCV method, called Robust OSCV, chooses the bandwidth of a so-called robust kernel which eliminates the bias of the OSCV bandwidths in the nonsmooth case, making the method robust to lack of smoothness in the regression function.

Chapter IV contains a summary of our findings and some suggestions for future research.

CHAPTER II

INDIRECT CROSS-VALIDATION FOR DENSITY ESTIMATION

1. Introduction

Let $X_1, \ldots, X_n$ be a random sample from an unknown density f. A kernel density estimator of f at the point x is defined as

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\Big(\frac{x - X_i}{h}\Big), \tag{2.1}$$

where $h > 0$ is a smoothing parameter, also known as the bandwidth, and K is the kernel, which is generally chosen to be a unimodal probability density function that is symmetric about zero and has finite variance. A popular choice for K is the Gaussian kernel: $\phi(u) = (2\pi)^{-1/2} \exp(-u^2/2)$. To distinguish between estimators with different kernels, we shall refer to estimator (2.1) with given kernel K as a *K-kernel estimator*.

Practical implementation of the estimator (2.1) requires specification of the smoothing parameter h, also known as the bandwidth. Different bandwidth selection methods are reviewed and compared by Jones, Marron, and Sheather (1996a) and Jones, Marron, and Sheather (1996b). The two most widely used bandwidth selectors are least squares cross-validation, proposed independently by Rudemo (1982) and Bowman (1984), and the Sheather and Jones (1991) plug-in method. Plug-in is often preferred since it produces more stable bandwidths than does LSCV. Nevertheless, the LSCV method is still popular since it requires fewer assumptions than the plug-in method and works well when the density is difficult to estimate; see Loader (1999b), van Es (1992).

The main flaw of LSCV is high variability of the selected bandwidths. Other drawbacks include the tendency of cross-validation curves to exhibit multiple local

minima with the first local minimum being too small (see Hall and Marron (1991)), and the tendency of LSCV to select bandwidths that are much too small when the data exhibit a small amount of autocorrelation (see Hart and Vieu (1990) and Cao, Quintela del Rio, and Vilar Fernandez (1993) for results of numerical studies).

A number of modifications of LSCV have been proposed in an attempt to improve its properties. These include biased cross-validation of Scott and Terrell (1987), a method of Chiu (1991a), the trimmed cross-validation of Feluch and Koronacki (1992), the modified cross-validation of Stute (1992), the one-sided cross-validation of Marti'nez Miranda, Nielsen, and Sperlich (2009), and the method of Ahmad and Ran (2004) based on kernel contrasts. *Indirect cross-validation* is a new modification of the LSCV method.

2. Description of indirect cross-validation

2.1. Notation and definitions

We begin with some notation and definitions that will be used subsequently. For an arbitrary function g, define

$$R(g) = \int g(u)^2 \, du, \quad \mu_{jg} = \int u^j g(u) \, du, \tag{2.2}$$

where here and subsequently integrals are assumed to be over the whole real line. The most popular measures of performance of the kernel estimators (2.1) are integrated squared error (ISE) and mean integrated squared error (MISE). The ISE is defined as

$$ISE(h) = \int \big(\hat{f}_h(x) - f(x)\big)^2 \, dx, \tag{2.3}$$

and MISE is defined as the expectation of ISE. Assume that the underlying density f has second derivative which is continuous and square integrable. Let K be a kernel of

the second order which satisfies the condition $R(K) < \infty$. Then the MISE function has the following asymptotic expansion (see Wand and Jones (1995)):

$$MISE(h) = \frac{R(K)}{nh} + \frac{h^4}{4}\mu_{2K}^2 R(f'') + o\left(\frac{1}{nh} + h^4\right). \tag{2.4}$$

This approximation is valid for $n \to \infty$ so long as $h \to 0$ and $nh \to \infty$. From expression (2.4) it follows that the bandwidth which asymptotically minimizes the MISE of the K-kernel estimator (2.1) has the following form:

$$h_n = \left\{\frac{R(K)}{\mu_{2K}^2 R(f'')}\right\}^{1/5} n^{-1/5}. \tag{2.5}$$

The LSCV criterion is given by

$$LSCV(h) = R(\hat{f}_h) - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{h,-i}(X_i), \tag{2.6}$$

where $\hat{f}_{h,-i}$ denotes the kernel estimator (2.1) constructed from the data without the observation X_i. A well known fact is that $LSCV(h)$ is an unbiased estimator of $MISE(h) - \int f^2(x)\,dx$. For this reason the LSCV method is often called *unbiased cross-validation*.

Let $\hat{h}_{UCV}$ and h_0 denote the bandwidths which minimize the LSCV function (2.6) and the MISE of the K-kernel estimator. Section 2.2 defines the ICV bandwidth, denoted as $\hat{h}_{ICV}$.

2.2. The basic method

We assume that the underlying density f has second derivative which is continuous and square integrable. Our aim is to choose the bandwidth of a *second order* kernel estimator. A second order kernel integrates to 1, has first moment 0, and finite, nonzero second moment. In principle our method can be used to choose the bandwidth

of any second order kernel estimator, but we restrict attention to $K \equiv \phi$, the Gaussian kernel. It is well known that a ϕ-kernel estimator has asymptotic mean integrated squared error (MISE) within 5% of the minimum among all positive, second order kernel estimators.

The essence of the ICV method is to use different kernels at the cross-validation and density estimation stages. The same idea is exploited by the one-sided cross-validation method of Hart and Yi (1998) and Marti'nez Miranda, Nielsen, and Sperlich (2009).

Indirect cross-validation may be described as follows:

1. Select the bandwidth of an L-kernel estimator using least squares cross-validation, and call this bandwidth $\hat{b}_{UCV}$. The kernel L is a second order kernel that is a linear combination of two Gaussian kernels, and will be discussed in detail in Section 2.3.

2. Assuming that the underlying density f has second derivative which is continuous and square integrable, the bandwidths h_n and b_n that asymptotically minimize the $MISE$ of ϕ- and L-kernel estimators, respectively, are related as follows:

$$h_n = \left(\frac{R(\phi)\mu_{2L}^2}{R(L)\mu_{2\phi}^2} \right)^{1/5} b_n \equiv C b_n. \tag{2.7}$$

3. Define the indirect cross-validation bandwidth by $\hat{h}_{ICV} = C\hat{b}_{UCV}$. Expression (2.7) and existing cross-validation theory suggest that $\hat{h}_{ICV}/h_0$ will at least converge to 1 in probability.

It is important that the constant C does not depend on any unknowns. Furthermore, the ICV method does not require any additional computing time compared to the LSCV method.

Theory of Hall and Marron (1987) and Scott and Terrell (1987) shows that the relative error $(\hat{h}_{UCV} - h_0)/h_0$ converges to 0 at the rather disappointing rate of $n^{-1/10}$. In contrast, we will show that $(\hat{h}_{ICV} - h_0)/h_0$ can converge to 0 at the rate $n^{-1/4}$. Kernels L that are sufficient for this result are discussed next.

2.3. Selection kernels

We consider the family of kernels $\mathcal{L} = \{L(\,\cdot\,;\alpha,\sigma) : \alpha > 0, \sigma > 0\}$, where, for all u,

$$L(u;\alpha,\sigma) = (1+\alpha)\phi(u) - \frac{\alpha}{\sigma}\phi\left(\frac{u}{\sigma}\right). \tag{2.8}$$

Note that the Gaussian kernel is a special case of (2.8) when $\sigma = 1$. Each member of $\mathcal{L}$ is symmetric about 0 and has the second moment $\mu_{2L} = 1+\alpha-\alpha\sigma^2$. It follows that kernels in $\mathcal{L}$ are second order, with the exception of those for which $\sigma = \sqrt{(1+\alpha)/\alpha}$.

The family $\mathcal{L}$ can be partitioned into three families: $\mathcal{L}_1$, $\mathcal{L}_2$ and $\mathcal{L}_3$. The first of these is $\mathcal{L}_1 = \left\{L(\cdot;\alpha,\sigma) : \alpha > 0, \sigma < \frac{\alpha}{1+\alpha}\right\}$. Each kernel in $\mathcal{L}_1$ has a negative dip centered at $x = 0$. The kernels in $\mathcal{L}_1$ are ones that "cut-out-the-middle," some examples of which are shown in Figure 2**(a)**.

The second family is $\mathcal{L}_2 = \left\{L(\cdot;\alpha,\sigma) : \alpha > 0, \frac{\alpha}{1+\alpha} \le \sigma \le 1\right\}$. Kernels in $\mathcal{L}_2$ are densities which can be unimodal or bimodal. Note that the Gaussian kernel is a member of this family. The third family is $\mathcal{L}_3 = \left\{L(\cdot;\alpha,\sigma) : \alpha > 0, \sigma > 1\right\}$, each member of which has negative tails. Examples are shown in Figure 2**(b)**.

Kernels in $\mathcal{L}_1$ and $\mathcal{L}_3$ turn out to be highly efficient for cross-validation purposes but very inefficient for estimating f. Indeed, it turns out that an L-kernel estimator based on a sequence of ICV-optimal kernels has $MISE$ that does not converge to 0 faster than $n^{-1/2}$. In contrast, the $MISE$ of the best ϕ-kernel estimator tends to 0 like $n^{-4/5}$. This explains why we do not use L as both a selection *and* an estimation kernel.

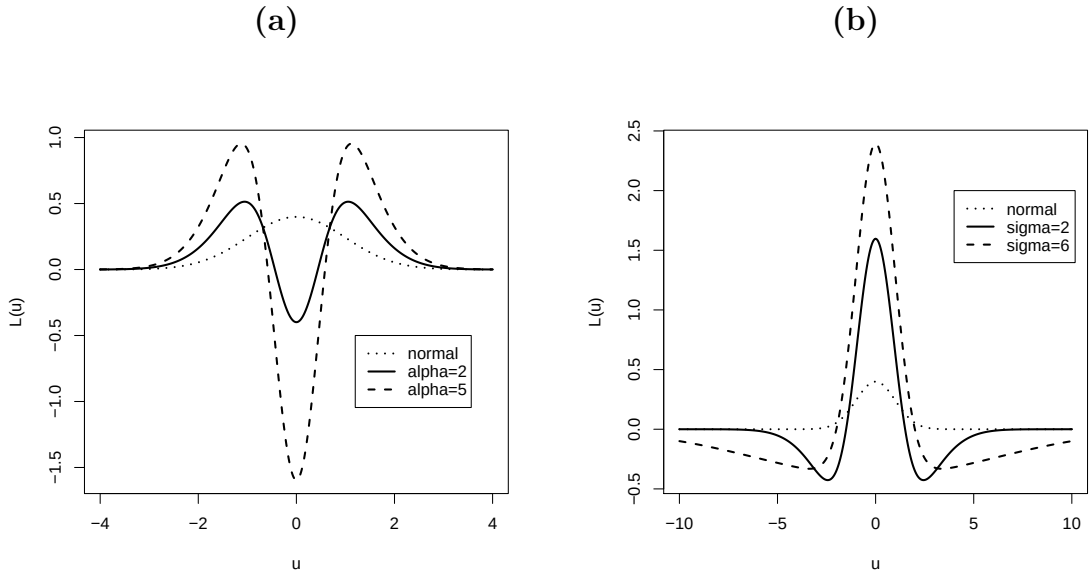


Fig. 2. **(a)** Selection kernels in $\mathcal{L}_1$ which have $\sigma = 0.5$; **(b)** Selection kernels in $\mathcal{L}_3$ with $\alpha = 6$. The dotted curve in both graphs corresponds to the Gaussian kernel.

Selection kernels in $\mathcal{L}$ are mixtures of two normal densities, which greatly simplifies computations. This fact has been utilized by Marron and Wand (1992) to derive exact MISE expressions. For kernels L it is possible to derive a closed form expression for the $LSCV$ function. Marron and Wand (1992) also point out that, in addition to their computational advantages, normal mixtures can approximate any density arbitrarily well in various senses. Mixtures of normals are therefore an excellent model for use in simulation studies, a fact which we take advantage of in subsection 8 below.

3. Large sample theory

Large sample theory for the ICV method is developed in this section. The main theoretical result is that the asymptotic MSE of $\hat{h}_{ICV}$ converges to 0 **faster** than the

asymptotic MSE of $\hat{h}_{UCV}$ in two cases:

1. $\sigma \to 0$ (cut-out-the-middle kernels);

2. $\sigma \to \infty$ (negative-tailed kernels).

We consider each of the cases in what follows.

3.1. Asymptotic MSE of the ICV bandwidth when $\sigma \to \infty$

We derive the asymptotic distribution for the ICV bandwidth in the case $\sigma \to \infty$. Before stating our main result, we define some notation:

$$\gamma_L(u) = \int L(w)L(w+u)\,du - 2L(u), \quad \rho_L(u) = u\gamma'(u), \tag{2.9}$$

$$T_n(b) = \sum\sum_{1\le i<j\le n} \left[\gamma_L\left(\frac{X_i - X_j}{b}\right) + \rho_L\left(\frac{X_i - X_j}{b}\right)\right],$$

$$T_n^{(j)}(b) = \frac{\partial^j T_n(b)}{\partial b^j}, \quad j = 1, 2,$$

$$A_\alpha = \frac{3}{\sqrt{2\pi}}(1+\alpha)^2\left[\frac{1}{8}(1+\alpha)^2 - \frac{8}{9\sqrt{3}}(1+\alpha) + \frac{1}{\sqrt{2}}\right],$$

$$C_\alpha = \frac{\sqrt{2A_\alpha}(2\sqrt{\pi})^{9/10}}{5(1+\alpha)^{9/5}\alpha^{1/5}} \qquad \text{and} \qquad D_\alpha = \frac{3}{20}\left(\frac{(1+\alpha)^2}{2\alpha^2\sqrt{\pi}}\right)^{2/5}.$$

Note that to simplify notation, we have suppressed the fact that L, γ and ρ depend on the parameters α and σ. An outline of the proof of the following theorem is given in the Appendix A.

Theorem II.1. *Assume that f and its first five derivatives are continuous and bounded and that $f^{(6)}$ exists and is Lipschitz continuous. Suppose also that*

$$(\hat{b}_{UCV} - b_0)\frac{T_n^{(2)}(\tilde{b})}{T_n^{(1)}(b_0)} = o_p(1) \tag{2.10}$$

for any sequence of random variables $\tilde{b}$ such that $|\tilde{b} - b_0| \le |\hat{b}_{UCV} - b_0|$, a.s. Then, if

$\sigma = o(n)$ and α is fixed,

$$\frac{\hat{h}_{ICV} - h_0}{h_0} = Z_n S_n + B_n + o_p(S_n + B_n),$$

as $n \to \infty$ and $\sigma \to \infty$, where Z_n converges in distribution to a standard normal random variable,

$$S_n = \left(\frac{1}{\sigma^{2/5} n^{1/10}}\right) \frac{R(f)^{1/2}}{R(f'')^{1/10}} C_\alpha, \tag{2.11}$$

and

$$B_n = \left(\frac{\sigma}{n}\right)^{2/5} \frac{R(f''')}{R(f'')^{7/5}} D_\alpha. \tag{2.12}$$

Remarks

(R1) Assumption (2.10) is only slightly stronger than assuming that $\hat{b}_{UCV}/b_0$ converges in probability to 1. The sufficient conditions for (2.10) can be found using techniques as in Hall (1983) and Hall and Marron (1987).

(R2) Theorem 4.1 of Scott and Terrell (1987) on asymptotic normality of LSCV bandwidths is not immediately applicable to our setting for at least three reasons: the kernel L is not positive, it does not have compact support, and, most importantly, it changes with n via the parameter σ.

(R3) The assumption of six derivatives for f is required for a precise quantification of the asymptotic bias of $\hat{h}_{ICV}$. Our proof of asymptotic normality of $\hat{b}_{UCV}$ only requires that f be four times differentiable, which coincides with the conditions of Theorem 4.1 in Scott and Terrell (1987).

(R4) The asymptotic bias B_n is positive, implying that the ICV bandwidth tends to

be larger than the optimal bandwidth. This is consistent with our experience in numerous simulations.

Now let us apply the results of our theorem to determine asymptotically optimal choices for α and σ. The limiting distribution of $(\hat{h}_{ICV} - h_0)/h_0$ has second moment $S_n^2 + B_n^2$, where S_n and B_n are defined by (2.11) and (2.12). Minimizing this expression with respect to σ yields the following asymptotically optimal choice for σ:

$$\sigma_{n,opt} = n^{3/8} \left(\frac{C_\alpha}{D_\alpha} \right)^{5/4} \left[\frac{R(f) R(f'')^{13/5}}{R(f''')^2} \right]^{5/8} =$$
$$\frac{2^{71/16} \pi^{1/2}}{3^{5/4}} \cdot \frac{R(f)^{5/8} R(f'')^{13/8}}{R(f''')^{5/4}} \cdot \frac{\alpha^{3/4}}{(1+\alpha)^2} \left(\frac{3}{8}(1+\alpha)^2 - \frac{8\sqrt{3}}{9}(1+\alpha) + \frac{3}{\sqrt{2}} \right)^{5/8} n^{3/8}. \tag{2.13}$$

Let $MSE(\hat{h}_{ICV}; \alpha, \sigma)$ denote the asymptotic MSE of the ICV bandwidth in the case $\sigma \to \infty$. It follows that

$$MSE(\hat{h}_{ICV}; \alpha, \sigma) = h_0^2 (S_n^2 + B_n^2).$$

Evaluating the MSE of $\hat{h}_{ICV}$ at $\sigma_{n,opt}$ we get the following:

$$MSE(\hat{h}_{ICV}; \alpha, \sigma_{n,opt}) =$$
$$\frac{3}{25 \cdot 2^{13/20} \pi^{1/5}} \cdot \frac{R(f)^{1/2} R(f''')}{R(f'')^{19/10}} \cdot \frac{1}{\alpha} \left(\frac{3}{8}(1+\alpha)^2 - \frac{8\sqrt{3}}{9}(1+\alpha) + \frac{3}{\sqrt{2}} \right)^{1/2} n^{-9/10}$$

From the above expression it follows that α is not confounded with f, meaning that we may determine a single optimal value of α that is independent of f. The function

$$f(\alpha) = \frac{1}{\alpha} \left(\frac{3}{8}(1+\alpha)^2 - \frac{8\sqrt{3}}{9}(1+\alpha) + \frac{3}{\sqrt{2}} \right)^{1/2},$$

normalized by its value at the minimum, is plotted in Figure 3. The minimum of the

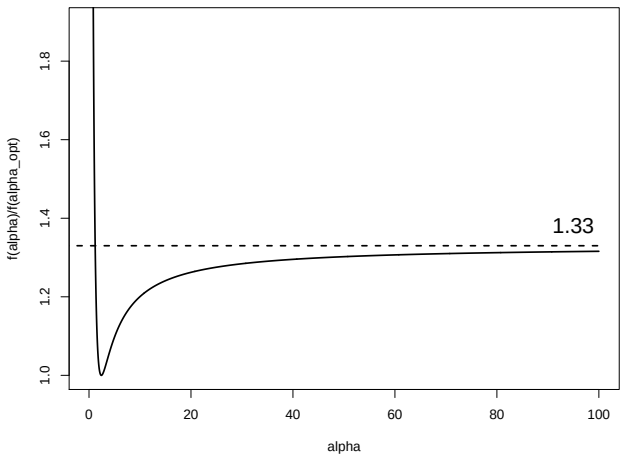


Fig. 3. How the asymptotically optimal MSE of $\hat{h}_{ICV}$ depends on α in the case $\sigma \to \infty$.

function $f(\alpha)$ occurs at the point

$$\alpha_{opt} = \frac{36\sqrt{6} + 9\sqrt{3} - 64}{32 - 9\sqrt{3}} \doteq 2.4233.$$

It turns out that small choices of α lead to an arbitrarily large increase in mean squared error, while the MSE at $\alpha = \infty$ is only about 1.33 times that at the minimum.

The asymptotic MSE of the ICV bandwidth evaluated at α_{opt} and $\sigma_{n,opt}$ has the following form:

$$MSE(\hat{h}_{ICV}; \alpha_{opt}, \sigma_{n,opt}) = C_\infty \frac{R(f)^{1/2} R(f''')}{R(f'')^{19/10}} n^{-9/10}, \tag{2.14}$$

where

$$C_\infty = \frac{\left(45576 - 21087\sqrt{2} + 16384\sqrt{3} - 15552\sqrt{6}\right)^{1/2}}{25\pi^{1/5} \cdot 2^{23/20}(36\sqrt{6} + 9\sqrt{3} - 64)} \doteq 0.0280.$$

It is remarkable that when the asymptotically optimal values of α and σ are used, the asymptotic variance and squared bias make *equal* contributions to the asymptotic

MSE of $\hat{h}_{ICV}$.

From expression (2.14) we can see that the optimal MSE of $\hat{h}_{ICV}$ tends to 0 at the rate $n^{-9/10}$. The corresponding rate for the LSCV method is $n^{-6/10}$. It also follows that

$$\frac{MSE(\hat{h}_{ICV}; \alpha_{opt}, \sigma_{n,opt})}{h_0^2} \sim n^{-1/2},$$

which implies that the relative error of $\hat{h}_{ICV}$ converges to 0 at the rate $n^{-1/4}$. The corresponding rates for LSCV and the Sheather-Jones plug-in rule are $n^{-1/10}$ and $n^{-5/14}$, respectively.

3.2. Asymptotic MSE of the ICV bandwidth when $\sigma \to 0$

Analogous asymptotic theory was developed for the case $\sigma \to 0$, which corresponds to $L \in \mathcal{L}_1$, i.e., kernels that apply negative weights to the smallest spacings in the data. The main theoretical results in this case are outlined below.

In the case when $\sigma \to 0$ the asymptotically MSE-optimal σ has the following form:

$$\sigma^*_{n,opt} = \frac{3^{5/4}}{2^{71/16}\pi^{1/2}} \cdot \frac{R(f''')^{5/4}}{R(f'')^{13/8}R(f)^{5/8}} \cdot \frac{\alpha^2}{(1+\alpha)^{3/4}} \cdot \frac{1}{\left(\frac{3}{8}\alpha^2 + \frac{8\sqrt{3}}{9}\alpha + \frac{3}{\sqrt{2}}\right)^{5/8}} n^{-3/8}$$

The asymptotic MSE evaluated at $\sigma^*_{n,opt}$ has the following form:

$$MSE^*(\hat{h}_{ICV}; \alpha, \sigma^*_{n,opt}) = \frac{3}{25 \cdot 2^{13/20}\pi^{1/5}} \frac{R(f)^{1/2}R(f''')}{R(f'')^{19/10}} \cdot \frac{1}{1+\alpha} \left(\frac{3}{8}\alpha^2 + \frac{8\sqrt{3}}{9}\alpha + \frac{3}{\sqrt{2}}\right)^{1/2} n^{-9/10}.$$

The asymptotically optimal α minimizes the function

$$f^*(\alpha) = \frac{1}{1+\alpha} \left(\frac{3}{8}\alpha^2 + \frac{8\sqrt{3}}{9}\alpha + \frac{3}{\sqrt{2}}\right)^{1/2},$$

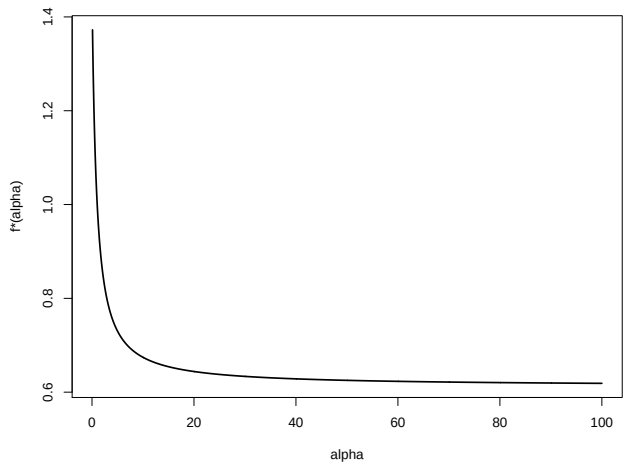


Fig. 4. How the asymptotically optimal MSE of $\hat{h}_{ICV}$ depends on α in the case $\sigma \to 0$.

which is plotted in Figure 4. It is easy to check that its minimum occurs at

$$\alpha^*_{opt} = \infty$$

The minimum asymptotic MSE in the case $\sigma \to 0$ has the following form:

$$MSE^*(\hat{h}_{ICV}; \alpha^*, \sigma^*_{n,opt}) = C_0 \frac{R(f)^{1/2} R(f''')}{R(f'')^{19/20}} n^{-9/10}, \tag{2.15}$$

where

$$C_0 = \frac{3^{3/2}}{25 \cdot 2^{43/2} \pi^{1/5}}.$$

The rates of the asymptotic MSE in the cases $\sigma \to \infty$ and $\sigma \to 0$ are the same. This means that the same optimal rate of $n^{-1/4}$ results from letting $\sigma \to 0$. Moreover, the minimum asymptotic MSE (2.14) and (2.15), corresponding to the cases $\sigma \to \infty$ and $\sigma \to 0$, respectively, have the same form, but different constants of proportionality.

The ratio of the constants is

$$\frac{C_0}{C_\infty} = \frac{3^{3/2}}{2} \cdot \frac{36\sqrt{6} + 9\sqrt{3} - 64}{\left(45576 - 21087\sqrt{2} + 16384\sqrt{3} - 15552\sqrt{6}\right)^{1/2}} \doteq 1.33,$$

implying that the asymptotically optimal negative-tailed kernels are more efficient than the asymptotically optimal cut-out-the-middle kernels. Our simulation studies confirm that using L with large σ does lead to more accurate estimation of the optimal bandwidth.

It is remarkable that when $\alpha \to \infty$ and the asymptotically optimal σ is used in both cases ($\sigma \to \infty$ and $\sigma \to 0$), the asymptotic MSE are equal:

$$MSE(\hat{h}_{ICV}; \infty, \sigma_{n,opt}) = MSE^*(\hat{h}_{ICV}; \infty, \sigma^*_{n,opt}).$$

This means that the negative-tailed kernels with $\sigma_{n,opt}$ and $\alpha \to \infty$ are as efficient as the asymptotically optimal cut-out-the-middle kernels.

Since the asymptotically optimal negative-tailed kernels are superior to the asymptotically optimal cut-out-the-middle kernels, all the subsequent ICV theory was developed for the case $\sigma \to \infty$.

4. Practical choice of α and σ

It is not immediately obvious how to use the asymptotic results developed in Section 3 for practical purposes. The asymptotically optimal α is a known constant, but the asymptotically optimal σ depends on f in a fairly complicated way (2.13). The problem of estimating the constant in (2.13) is potentially as difficult or even more difficult than estimating f itself. A reference estimator for σ based on the standard normal density may be used. In this section we develop a practical purpose model for choosing the parameters α and σ of the selection kernel L defined by (2.8).

4.1. MSE-optimal α and σ

Asymptotic results are not always reliable for practical purposes. In order to have an idea of how good choices of α and σ vary with n and f, we considered the following expression for the asymptotic MSE of the ICV bandwidth:

$$MSE(\alpha,\sigma;f,n) = \left(\frac{1}{4\pi}\right)^{1/5} \frac{R(f''')^2}{R(f'')^{16/5}} n^{-3/5} \left\{ \frac{2}{25} \frac{R(f)R(f'')^{13/5}}{R(f''')^2} \frac{R(\rho_L)}{R(L)^{9/5}(\mu_{2L}^2)^{1/5}} + \right.$$

$$\left. \frac{n^{-3/5}}{400} \left(\frac{R(L)^{2/5}\mu_{2L}\mu_{4L}}{(\mu_{2L}^2)^{7/5}} - \frac{3}{(4\pi)^{1/5}} \right)^2 \right\}, \tag{2.16}$$

where $\rho_L(u)$ is defined by (2.9). Expression (2.16) is valid for either large or small values of σ and uses a slightly enhanced version of the asymptotic bias of $\hat{h}_{ICV}$. The first order bias of $\hat{h}_{ICV}$ is $Cb_0 - h_0$, or $C(b_0 - b_n) + (h_n - h_0)$, where b_n and h_n are the asymptotic MISE minimizers for the L-kernel and ϕ-kernel estimators, respectively. Now, the term $h_n - h_0$ is of smaller order asymptotically than $C(b_0 - b_n)$ and hence was deleted in the theory of Section 3. In expression (2.16) we retain $h_n - h_0$. Notice that the α minimizing expression (2.16) is not free of f.

In order to have an idea of how good choices of α and σ vary with n and f, we determined the minimizers of the asymptotic MSE (2.16) for various sample sizes and densities. It is worth noting that the asymptotically optimal σ (expression (2.13)) is free of location and scale. We may thus choose a single representative of a location-scale family when investigating the effect of f. We considered the following five normal mixtures defined in the article by Marron and Wand (1992):

Gaussian density:	$N(0,1)$
Skewed unimodal density:	$\frac{1}{5}N(0,1) + \frac{1}{5}N\left(\frac{1}{2}, \left(\frac{2}{3}\right)^2\right) + \frac{3}{5}N\left(\frac{13}{12}, \left(\frac{5}{9}\right)^2\right)$
Bimodal density:	$\frac{1}{2}N\left(-1, \left(\frac{2}{3}\right)^2\right) + \frac{1}{2}N\left(1, \left(\frac{2}{3}\right)^2\right)$
Separated bimodal density:	$\frac{1}{2}N\left(-\frac{3}{2}, \left(\frac{1}{2}\right)^2\right) + \frac{1}{2}N\left(\frac{3}{2}, \left(\frac{1}{2}\right)^2\right)$
Skewed bimodal density:	$\frac{3}{4}N(0,1) + \frac{1}{4}N\left(\frac{3}{2}, \left(\frac{1}{3}\right)^2\right)$.

These choices for f provide a fairly representative range of density shapes. In Table I we provide the MSE-optimal choices of α and σ for the above densities at eight sample sizes ranging from $n = 100$ up to $n = 500000$.

Table I. MSE-optimal α and σ.

	Density									
	normal		skewed unimodal		bimodal		separated bimodal		skewed bimodal	
n	α	σ	α	σ	α	σ	α	σ	α	σ
100	3.05	2.79	5.28	1.68	109.68	1.03	16.70	1.19	343.74	1.01
250	2.78	4.04	3.16	2.60	48.46	1.06	4.51	1.84	177.15	1.02
500	2.73	4.97	2.84	3.56	6.21	1.55	3.18	2.58	161.39	1.02
1000	2.69	5.97	2.75	4.49	3.73	2.12	2.84	3.54	123.78	1.03
5000	2.61	8.84	2.66	6.85	2.77	4.26	2.70	5.74	4.71	1.79
20000	2.55	12.40	2.59	9.58	2.68	6.22	2.63	8.08	2.85	3.46
100000	2.50	18.80	2.53	14.27	2.60	9.19	2.56	11.94	2.70	5.65
500000	2.47	29.54	2.49	21.88	2.54	13.65	2.50	18.07	2.62	8.39

The following remarks summarize our findings about α and σ:

1. MSE-optimal α and σ vary greatly from one density to another, which is especially true for "small" sample sizes.

2. For each density, the optimal α decreases monotonically with n and seems to converge to the asymptotically optimal value $\alpha_{opt} = 2.42$ which was derived in Section 3. However, the convergence to the optimal α is very slow, especially for the bimodal densities. Thus, for each unimodal density, the optimal α is within 13.5% of 2.42 at $n = 1000$, and for each bimodal density is within 18% of 2.42 when n is 20,000.

3. The MSE-optimal σ is increasing with sample size for all the densities, which supports the theory of Section 3.

4. For each n, the optimal value of σ (α) is larger (smaller) for the unimodal densities than for the bimodal ones.

5. All of the MSE-optimal α and σ correspond to kernels from $\mathcal{L}_3$, the family of negative-tailed kernels.

4.2. Model for the ICV parameters

We have built a practical purpose model for α and σ using the data outlined in Table I. We used the polynomial regression method. Our independent variable was $\log_{10}(n)$ and the dependent variables were the MSE-optimal values of $\log_{10}(\alpha)$ and $\log_{10}(\sigma)$ found from Table I. The $\log_{10}$ transformations for the MSE-optimal α and σ were needed to stabilize variability. Notice that the five densities defined in Section 4.1 play the role of reference distributions in building our model. Using a sixth degree polynomial for α and a quadratic for σ, we arrived at the following models for α and σ:

Table II. Model choices of α and σ.

n	100	250	500	1000	5000	20000	100000	500000
α_{mod}	25.20	12.77	8.24	5.71	3.23	2.66	2.66	2.62
σ_{mod}	1.39	1.89	2.37	2.95	4.83	7.21	11.22	16.98

$$\begin{aligned} \alpha_{mod} &= 10^{3.390-1.093\log 10(n)+0.025\log 10(n)^3-0.00004\log 10(n)^6}, \\ \sigma_{mod} &= 10^{-0.58+0.386\log 10(n)-0.012\log 10(n)^2}, \quad 100 \le n \le 500000. \end{aligned} \tag{2.17}$$

The MSE-optimal values of $\log_{10}(\alpha)$ and σ together with the model fits are shown in Figure 5. In Table II we give the model choices α_{mod} and σ_{mod} for the same sample sizes as in Table I.

To the extent that unimodal densities are more prevalent than multimodal densities in practice, these model values are biased towards bimodal cases. Our extensive experience shows that the penalty for using good bimodal choices for α and σ when in fact the density is unimodal, is an increase in the upward bias of $\hat{h}_{ICV}$. Our implementation of ICV, however, guards against oversmoothing by using an objective upper bound on the bandwidth, as we explain in detail in Section 7. We thus feel confident in recommending model (2.17) for choosing α and σ in practice, at least until a better method is proposed. Indeed, this model is what we used to choose α and σ in the simulation study reported upon in Section 7.

5. Efficiency of the model-based kernels in bandwidth selection

Define the bandwidth selection efficiency of the selection kernel $L(\cdot;\alpha,\sigma)$ relative to the Gaussian kernel as

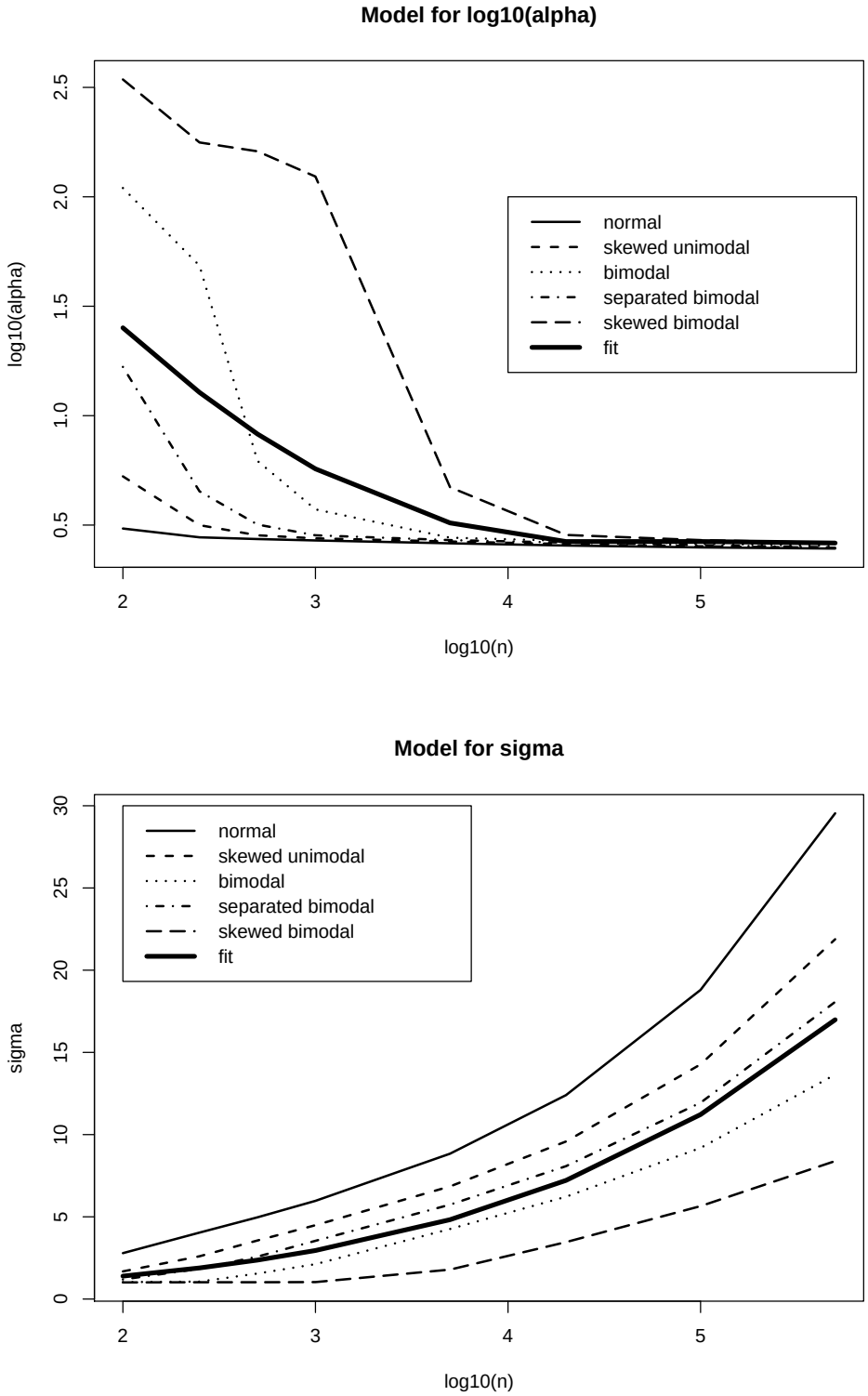


Fig. 5. MSE-optimal $\log_{10}(\alpha)$ and σ and the model fits.

Table III. Efficiencies of $L(\cdot;\alpha_{mod},\sigma_{mod})$ in selecting the bandwidth for five densities at different sample sizes.

n	100	250	500	1000	5000	20000	100000	500000
Gaussian density	0.6273	0.5020	0.3627	0.2479	0.1059	0.0615	0.0373	0.0233
Skewed unimodal	0.6315	0.5072	0.3700	0.2573	0.1180	0.0720	0.0443	0.0277
Bimodal	0.6487	0.5288	0.3999	0.2958	0.1674	0.1150	0.0729	0.0456
Separated bimodal	0.6362	0.5132	0.3783	0.2679	0.1317	0.0839	0.0522	0.0326
Skewed bimodal	0.7087	0.6041	0.5039	0.4300	0.3396	0.2649	0.1724	0.1078

$$E(\alpha,\sigma,f,n)=\frac{\mathrm{MSE}(\alpha,\sigma;f,n)}{\mathrm{MSE}(1,1;f,n)}, \tag{2.18}$$

where $\mathrm{MSE}(\alpha,\sigma,f,n)$ is given by (2.16). The denominator of (2.18) is the asymptotic MSE of the LSCV bandwidth, since the values $\alpha=1$ and $\sigma=1$ correspond to the Gaussian kernel.

Our theory of Section 3 suggests that for the asymptotically optimal kernels the efficiency E tends to 0 at the rate $O\left(n^{-3/10}\right)$ as $n\to\infty$. Even though our practical purpose model (2.17) estimates the asymptotically optimal parameters, it does not use the explicit expressions for $\alpha_{n,opt}$ and $\sigma_{n,opt}$ which guarantee the relative bandwidth rate of $O\left(n^{-1/4}\right)$. What are the efficiencies for the model-based kernels for the sample sizes allowed by the model (2.17)?

Table III gives the efficiencies of the kernels $L(\cdot;\alpha_{mod},\sigma_{mod})$ for eight sample sizes and densities defined in Section 4.1. As we can conclude from Table III, using the model-based kernels $L(\cdot;\alpha_{mod},\sigma_{mod})$ in cross-validation is more appropriate than using the Gaussian kernel for all the considered densities and sample sizes. Moreover, the efficiencies in Table III decrease as n increases, so that using the Gaussian kernel at large sample sizes becomes quite unreasonable. For instance, using the kernel

$L(\cdot;\alpha_{mod},\sigma_{mod})$ leads to more than a fourfold decrease of MSE compared to using the Gaussian kernel K at $n = 1000$ and Gaussian density. Efficiency issues justify the rationale of using the model-based kernel $L(\cdot;\alpha_{mod},\sigma_{mod})$ for the purpose of bandwidth selection.

6. Robustness of ICV to data rounding

The LSCV function (2.6) can be written in the following form:

$$\text{LSCV}(h) = \frac{1}{nh}R(K) + \frac{1}{n^2h}\sum_{i\neq j}\int K(t)K\Big(t+\frac{X_i-X_j}{h}\Big)\,dt - \frac{2}{n(n-1)h}\sum_{i\neq j}K\Big(\frac{X_i-X_j}{h}\Big). \quad (2.19)$$

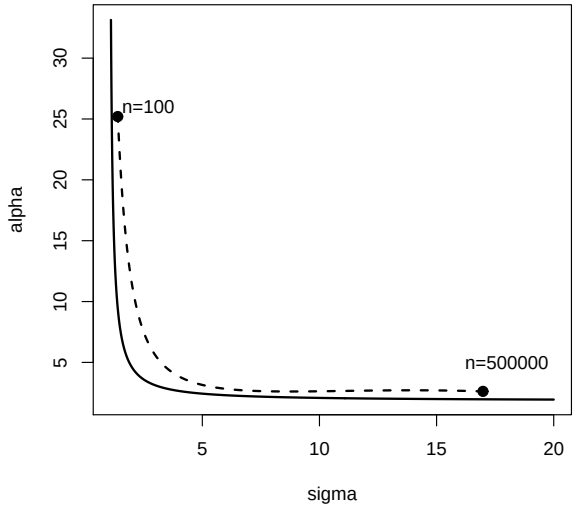
and hence it is clear that LSCV depends on the spacings $X_i - X_j$. Silverman (1986, p.52) showed that if the data are rounded to such an extent that the number of pairs $i < j$ for which $X_i = X_j$ is above a threshold, then $LSCV(h)$ approaches $-\infty$ as h approaches zero. This threshold is $0.27n$ for the Gaussian kernel. Chiu (1991b) showed that for data with ties, the behavior of $LSCV(h)$ as $h \to 0$ is determined by the balance between $R(K)$ and $2K(0)$. In particular, $\lim_{h\to 0} LSCV(h)$ is $-\infty$ and ∞ when $R(K) < 2K(0)$ and $R(K) > 2K(0)$, respectively. The former condition holds necessarily if K is nonnegative and has its maximum at 0. This means that all the traditional kernels have the problem of choosing $h = 0$ when the data are rounded.

Recall that selection kernels (2.8) are not restricted to be nonnegative. It turns out that there exist α and σ such that $R(L) > 2L(0)$ will hold. We say that selection kernels satisfying this condition are robust to rounding. It can be verified that the negative-tailed selection kernels with $\sigma > 1$ are robust to rounding when

$$\alpha > \frac{-a_\sigma + \sqrt{a_\sigma + (2 - 1/\sqrt{2})b_\sigma}}{b_\sigma}, \quad (2.20)$$

where $a_\sigma = \left(\frac{1}{\sqrt{2}} - \frac{1}{\sqrt{1+\sigma^2}} - 1 + \frac{1}{\sigma}\right)$ and $b_\sigma = \left(\frac{1}{\sqrt{2}} - \frac{2}{\sqrt{1+\sigma^2}} + \frac{1}{\sigma\sqrt{2}}\right)$. It appears that all the selection kernels corresponding to model (2.17) are robust to rounding. Figure 6 shows the region (2.20) and also the curve defined by model (2.17) for $100 \leq n \leq 500000$. Interestingly, the boundary separating robust from nonrobust kernels almost

Fig. 6. Selection kernels robust to rounding have α and σ above the solid curve. The dashed curve corresponds to the model-based selection kernels.



coincides with the (α, σ) pairs defined by that model.

Notice that the fact that $R(L) > 2L(0)$ for the model-based kernels has one more consequence. Consider the behavior of the LSCV(h) function at large values of the bandwidth h. From expression (2.19) it follows that as $h \to \infty$ the asymptotic expression for the LSCV(h) based on the K-kernel estimator has the following form:

$$\text{LSCV}(h) \sim \frac{R(K)}{h} - \frac{2K(0)}{h}.$$

It follows that LSCV(h) $\to 0$ as $h \to \infty$. The sign of LSCV(h) for large h depends on

the sign of the difference $R(K)-2K(0)$. This difference is negative for the traditional kernels and is positive for the model-based kernels L. Then it follows that at large n the ICV criterion function approaches zero from the positive side as $h \to \infty$, implying that when the local minima of the ICV curve are positive, the ICV minimizer will be $h = \infty$. This emphasizes the necessity to restrict the range of h over which we minimize the ICV function. Asymptotically, the problem of a global minimum at $h = \infty$ will go away since an LSCV curve is centered at $-R(f)$ (see Scott and Terrell (1987)).

7. Local ICV

A local version of cross-validation for density estimation was proposed and analyzed independently by Hall and Schucany (1989) and Mielniczuk, Sarda, and Vieu (1989). A local method allows the bandwidth to vary with x, which is desirable when the smoothness of the underlying density varies sufficiently with x. Fan, Hall, Martin, and Patil (1996) proposed a different method of local smoothing that is a hybrid of plug-in and cross-validation methods. Here we propose that ICV be performed locally. The method parallels that of Hall and Schucany (1989) and Mielniczuk, Sarda, and Vieu (1989), with the main difference being that each local bandwidth is chosen by ICV rather than LSCV. We suggest using the *smallest* local minimizer of the ICV curve, since ICV does not have LSCV's tendency to undersmooth.

The local ICV criterion function at the point x is defined as

$$ICV(x,b,w) = \frac{1}{w}\int_{-\infty}^{\infty} \phi\left(\frac{x-u}{w}\right) \hat{f}_b^2(u)\, du - \frac{2}{nw}\sum_{i=1}^{n} \phi\left(\frac{x-X_i}{w}\right) \hat{f}_{b,-i}(X_i),$$

where $\hat{f}_b$ is the kernel density estimate based on a selection kernel L with a smoothing parameter b. The quantity w determines the degree to which the cross-validation is

local, with a very large choice of w corresponding to global ICV. Let $\hat{b}(x)$ be the first local minimizer of $ICV(x, b, w)$ with respect to b for the fixed value of x. Then the bandwidth of a Gaussian kernel estimator at the point x is taken to be $\hat{h}(x) = C\hat{b}(x)$. The constant C is defined by (2.7), and choice of α and σ in the selection kernel L will be discussed in Section 8.

Local LSCV can be criticized on the grounds that, at any x, it promises to be even more unstable than global LSCV since it (effectively) uses only a fraction of the n observations. Because of its much greater stability, ICV seems to be a much more feasible method of local bandwidth selection than does LSCV. We provide evidence of this stability by examples in Section 9.

8. Simulation study

The primary goal of our simulation study is to compare ICV with ordinary LSCV. However, we will also include the Sheather-Jones plug-in method in the study. We considered the four sample sizes $n = 100$, 250, 500 and 5000, and sampled from each of the five densities listed in Section 4.1. For each combination of density and sample size, 1000 replications were performed.

Let $\hat{h}_0$ denote the minimizer of $ISE(h)$ for a Gaussian kernel estimator. For each replication, we computed $\hat{h}_0$, $\hat{h}^*_{ICV}$, $\hat{h}_{UCV}$ and $\hat{h}_{SJPI}$. The definition of $\hat{h}^*_{ICV}$ is as follows:

$$\hat{h}^*_{ICV} = \min(\hat{h}_{ICV}, \hat{h}_{OS}), \tag{2.21}$$

where

$$\hat{h}_{OS} = \left(\frac{243}{35}\right)^{1/5} \left(\frac{R(\phi)}{\mu_{2\phi}^2}\right)^{1/5} s \cdot n^{-1/5} = \left(\frac{243}{35} \cdot \frac{1}{2\sqrt{\pi}}\right)^{1/5} s \cdot n^{-1/5}$$

is the oversmoothed bandwidth of Terrell (1990); s is the sample standard deviation

computed for the data $x_1, \ldots, x_n$. It is arguable that *no* data-driven bandwidth should be larger than $\hat{h}_{OS}$ since this statistic estimates an upper limit for *all* MISE-optimal bandwidths (under standard smoothness conditions). Since $\hat{h}_{ICV}$ tends to be biased upwards, using the bandwidth $\hat{h}_{OS}$ as an upper bound for the bandwidth search interval is a convenient means of limiting the bias. In Table XI of Appendix B we give the percentage of times when the upper bound of $\hat{h}_{OS}$ is used in the bandwidth selection rule (2.21). In all cases the parameters α and σ in the selection kernel L were chosen according to model (2.17).

For any random variable Y defined in each replication of our simulation, we denote the average, standard deviation and median of Y over all replications (with n and f fixed) by $\widehat{\mathrm{E}}(Y)$, $\widehat{\mathrm{SD}}(Y)$ and $\widehat{\mathrm{Median}}(Y)$. To evaluate the bandwidth selectors we computed $\widehat{\mathrm{E}}\{ISE(\hat{h})/ISE(\hat{h}_0)\}$ and $\widehat{\mathrm{Median}}\{ISE(\hat{h})/ISE(\hat{h}_0)\}$ for $\hat{h}$ equal to each of $\hat{h}^*_{ICV}$, $\hat{h}_{UCV}$ and $\hat{h}_{SJPI}$. We also computed the performance measure $\widehat{E}\left(\hat{h} - \hat{E}(\hat{h}_0)\right)^2$, which estimates the MSE of the bandwidth $\hat{h}$.

Our simulation results for the "normal" and "bimodal" densities, as defined in Section 4.1, are given in Tables IV and V and Figures 7 and 8. Results for the "skewed unimodal", "separated bimodal" and "skewed bimodal" densities are given in the Appendix B. Our main observations and conclusions are summarized as follows.

1. The reduced variability of the ICV bandwidth is evident in our study. The ratio $\widehat{\mathrm{SD}}(\hat{h}^*_{ICV})/\widehat{\mathrm{SD}}(\hat{h}_{UCV})$ ranged between 0.21 and 0.97 in the twenty settings considered. However, the variances of the ICV bandwidths were always higher compared to the Sheather-Jones plug-in bandwidths. It is also worth noting that the ratio of sample standard deviations of the ICV and LSCV bandwidths decreases as the sample size n increases.

2. The ratio $\widehat{E}\left(\hat{h}^*_{ICV} - \widehat{E}\hat{h}_0\right)^2 / \widehat{E}\left(\hat{h}_{UCV} - \widehat{E}\hat{h}_0\right)^2$ ranged between 0.04 and 0.70

in the sixteen settings excluding the skewed bimodal density. For the skewed bimodal density, the ratio was 0.84, 1.27, 1.09, and 0.40 at the respective sample sizes 100, 250, 500 and 5000. The fact that this ratio was larger than 1 in two cases was a result of ICV's bias, since the sample standard deviation of the ICV bandwidth was smaller than that for the LSCV bandwidth in all twenty settings. Notice that plug-in always had a smaller value of $\widehat{E}\left(\hat{h} - \widehat{E}\hat{h}_0\right)^2$ than did ICV.

3. The most important observation is that the values of $\widehat{\mathrm{E}}\big(ISE(\hat{h})/ISE(\hat{h}_0)\big)$ were smaller for ICV than for LSCV for all combinations of densities and sample sizes. The values of $\widehat{\mathrm{Median}}\big(ISE(\hat{h})/ISE(\hat{h}_0)\big)$ were smaller for ICV than for LSCV in all but one case, which corresponds to the large bias case when the density is skewed bimodal and $n = 250$. In this case $\widehat{\mathrm{Median}}\big(ISE(\hat{h})/ISE(\hat{h}_0)\big)$ was 1.0013 times greater for ICV than for LSCV. Being close to LSCV in bimodal case is not bad since in that case LSCV performs well.

4. Despite the fact that the LSCV bandwidth is asymptotically normally distributed (see Hall and Marron (1987)), its distribution in finite samples tends to be skewed to the left. In our simulations we have noticed that the distribution of the ICV bandwidth is less skewed than that of the LSCV bandwidth. A typical case is illustrated in Figure 9, where kernel density estimates for the two data-driven bandwidths are plotted from the simulation with the skewed unimodal density at $n = 250$. Also plotted is a density estimate for the ISE-optimal bandwidths. Note that the ICV density is more concentrated near the middle of the ISE-optimal distribution than the density estimate for LSCV.

5. Usually the ICV bandwidths cluster more tightly about the MISE minimizer h_0 as opposed to the LSCV bandwidths. A typical example is given in Figure 10

which provides scatterplots of the bandwidths $\hat{h}_{UCV}$ and $\hat{h}_{ICV}$ versus the ISE-optimal bandwidths $\hat{h}_0$ in the case of the Gaussian density and $n = 500$. In this case the MISE minimizer is $h_0 = 0.315$, and the ICV bandwidths are better concentrated about it compared to the LSCV bandwidths. Notice that the sample correlation coefficients were -0.52 and -0.60 for LSCV and ICV, respectively. The fact that these correlations are negative is a well-established phenomenon; see, for example Hall and Johnstone (1992).

Table IV. Simulation results for the Gaussian density.

n	**LSCV**	**SJPI**	**ICV**	**ISE**
$\widehat{\mathrm{E}}(\hat{h})$				
100	0.4452	0.3934	0.4153	0.4316
250	0.3640	0.3388	0.3494	0.3549
500	0.3109	0.2980	0.3086	0.3081
5000	0.1836	0.1899	0.1977	0.1953
$\widehat{\mathrm{SD}}(\hat{h}) \cdot 10^2$				
100	12.3217	6.4324	6.5230	7.5201
250	8.3577	3.7174	4.4478	6.2730
500	7.1117	2.6030	3.0802	5.6350
5000	3.9008	0.6190	0.8204	3.0928
$\widehat{\mathrm{E}}(\hat{h} - \widehat{\mathrm{E}}(\hat{h}_0))^2 \cdot 10^4$				
100	153.5291	55.9547	45.1705	
250	70.6115	16.3766	20.0568	
500	50.6085	7.7748	9.4813	
5000	16.5621	0.6679	0.7311	
$\widehat{\mathrm{E}}\big(ISE(\hat{h})/ISE(\hat{h}_0)\big)$				
100	2.4700	1.9080	1.7218	
250	1.9159	1.5056	1.4757	
500	1.7581	1.3773	1.3610	
5000	1.4132	1.1146	1.1031	
$\widehat{\mathrm{Median}}\big(ISE(\hat{h})/ISE(\hat{h}_0)\big)$				
100	1.3111	1.1570	1.1123	
250	1.2172	1.1041	1.0937	
500	1.2140	1.1031	1.0961	
5000	1.1091	1.0447	1.0518	

Table V. Simulation results for the Bimodal density.

n	**LSCV**	**SJPI**	**ICV**	**ISE**
$\widehat{\mathrm{E}}(\hat{h})$				
100	0.4291	0.3945	0.4196	0.3824
250	0.3136	0.3116	0.3285	0.2972
500	0.2593	0.2624	0.2745	0.2532
5000	0.1526	0.1571	0.1626	0.1548
$\widehat{\mathrm{SD}}(\hat{h}) \cdot 10^2$				
100	13.5653	7.4443	9.5668	7.6090
250	8.4673	4.1878	6.5092	4.2943
500	5.7059	2.4444	4.2008	3.5598
5000	2.4629	0.4795	0.8146	1.9650
$\widehat{\mathrm{E}}(\hat{h} - \widehat{\mathrm{E}}(\hat{h}_0))^2 \cdot 10^4$				
100	205.6555	56.8404	105.2554	
250	74.3324	19.6074	52.1298	
500	32.8927	6.8119	22.1647	
5000	6.1066	0.2820	1.2669	
$\widehat{\mathrm{E}}\big(ISE(\hat{h})/ISE(\hat{h}_0)\big)$				
100	1.6995	1.3273	1.3614	
250	1.5160	1.2091	1.2874	
500	1.4167	1.1507	1.1917	
5000	2.0643	1.0684	1.0768	
$\widehat{\mathrm{Median}}\big(ISE(\hat{h})/ISE(\hat{h}_0)\big)$				
100	1.2095	1.0874	1.1336	
250	1.1609	1.0834	1.1270	
500	1.1224	1.0607	1.0942	
5000	1.0583	1.0307	1.0365	

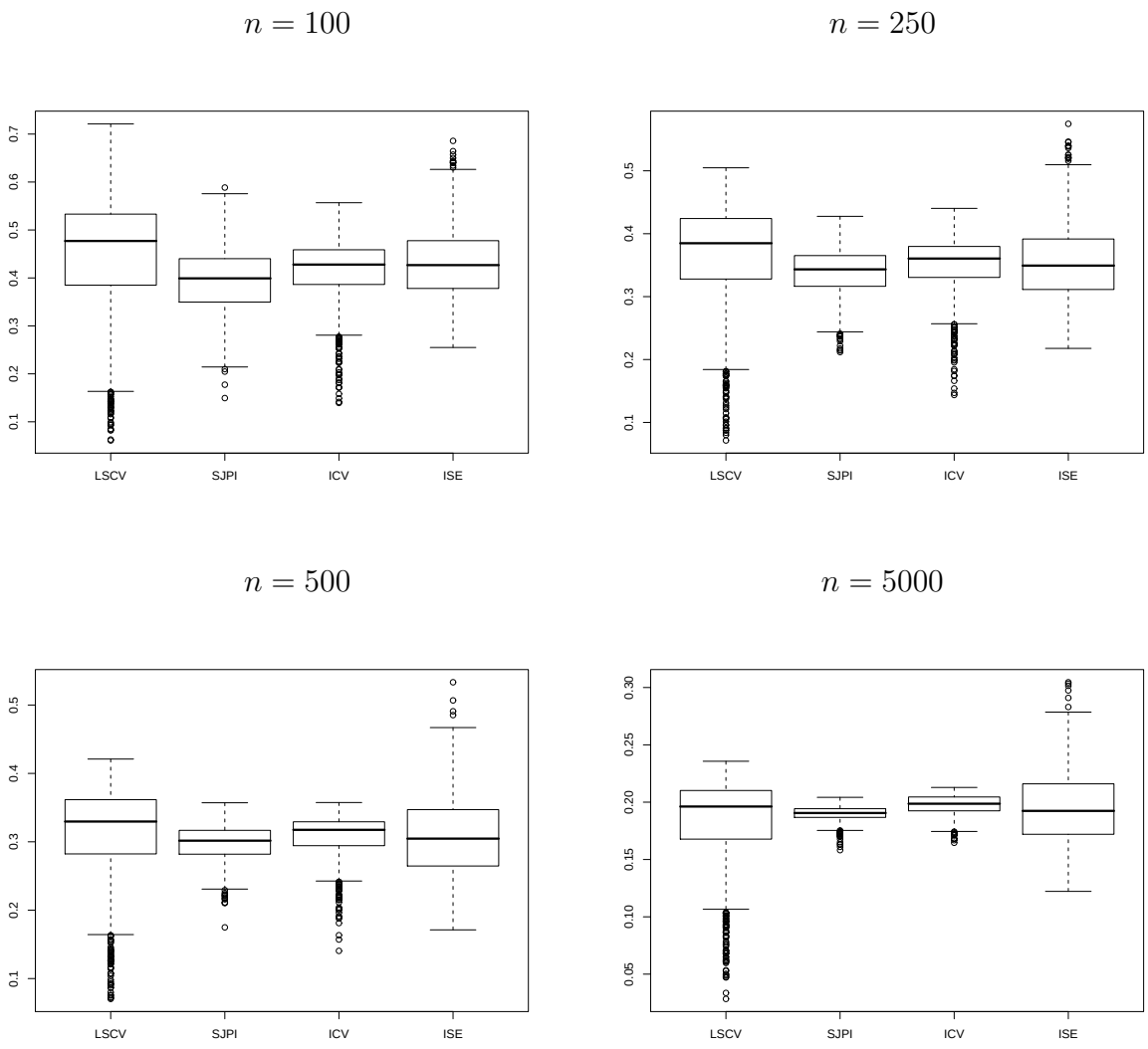


Fig. 7. Boxplots for the data-driven bandwidths in the case of the Gaussian density.

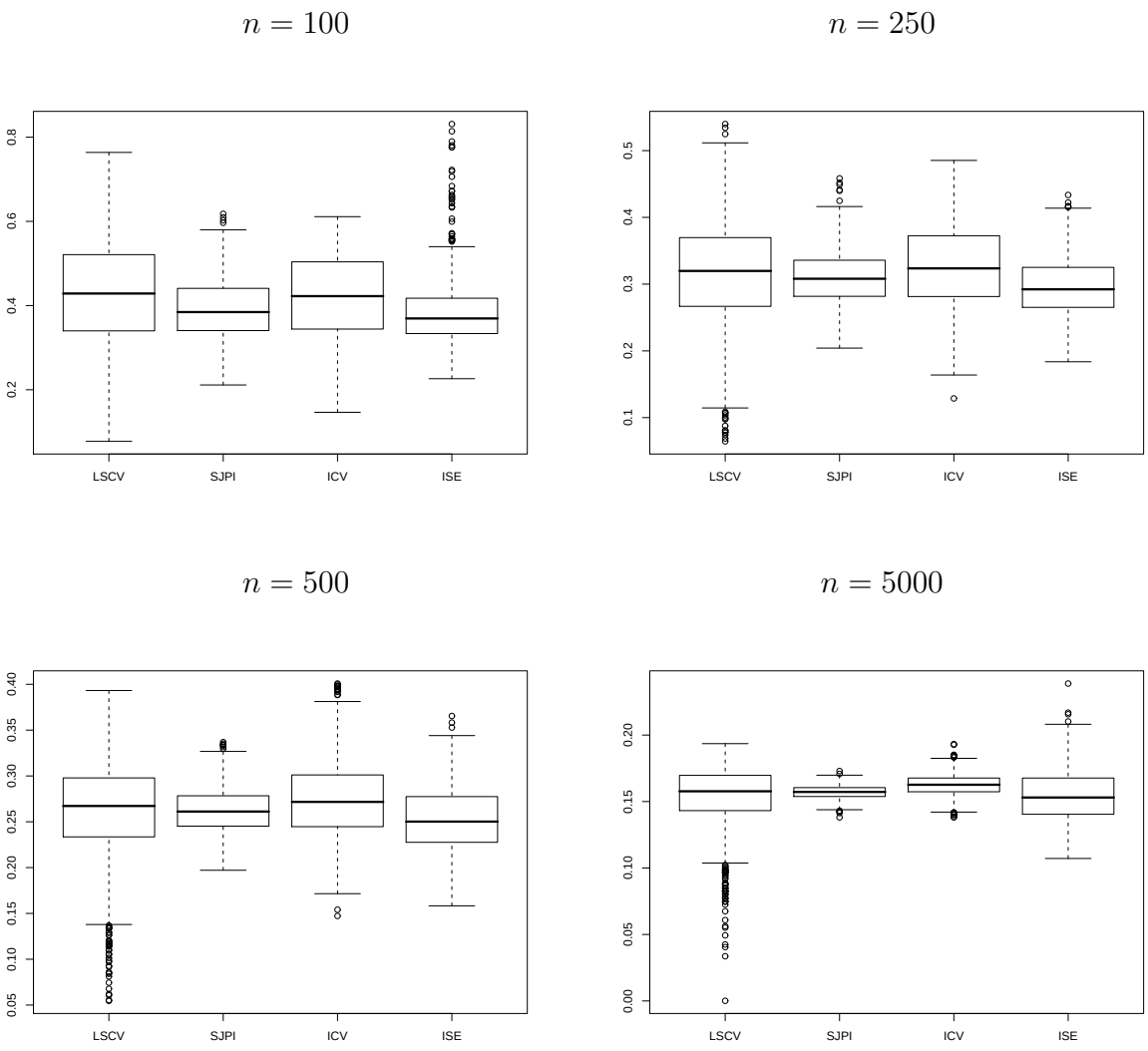


Fig. 8. Boxplots for the data-driven bandwidths in the case of the Bimodal density.

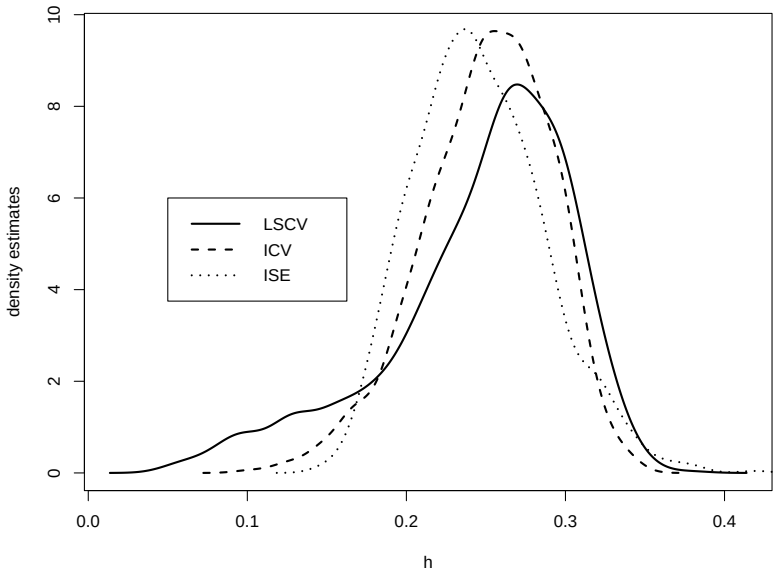


Fig. 9. Kernel density estimates for random bandwidths from the simulation with the Skewed unimodal density and $n = 250$.

A problem we have noticed with the ICV method is that its criterion function can have two local minima when the sample size is moderate and the density has two modes. The following example illustrates the problem. Let $ICV(h)$ denote the ICV criterion function which is computed using kernel L in place of K in the cross-validation function (2.6). In Figure 11**(a)** we have plotted three functions $ICV\left(\frac{h}{C}\right)$ for the case of the separated bimodal density and $n = 100$. The minimizers of the solid, dashed and dotted lines occur at the h-values 0.2991, 2.0467 and 0.2204, respectively. For comparison, the corresponding bandwidths chosen by the Sheather-Jones plug-in method are 0.3240, 0.2508 and 0.2467. The value of $h = 2.0467$ which minimizes the dashed $ICV\left(\frac{h}{C}\right)$ curve is obviously too large. The local minimum at 0.1295 would yield a much more reasonable estimate. The problem of choosing too large

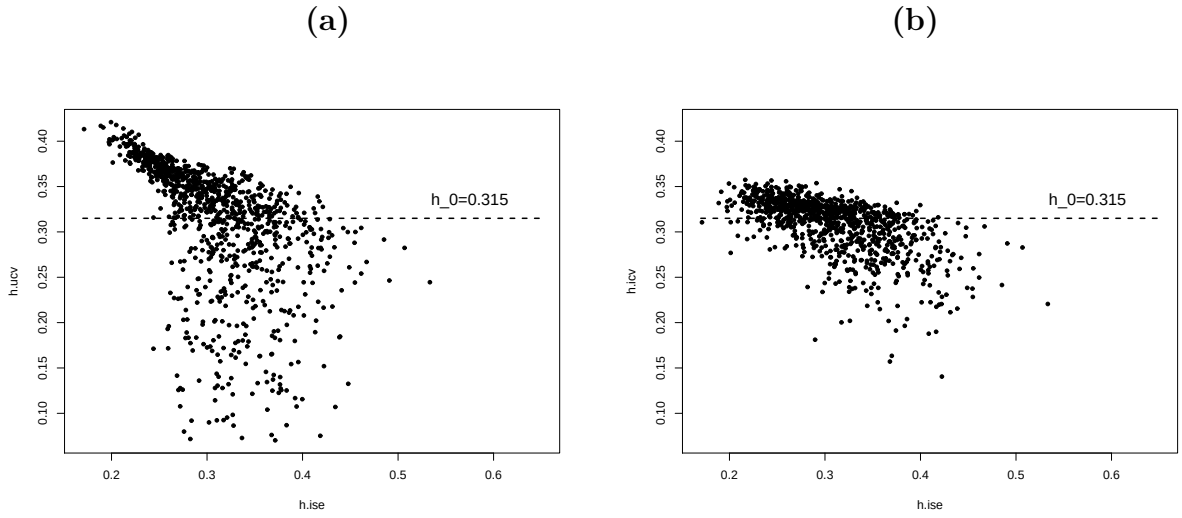


Fig. 10. Scatterplots of $\hat{h}$ vs. $\hat{h}_0$ for the case of the Gaussian density and $n = 500$, with $\hat{h}$ corresponding to the **(a)** LSCV and **(b)** ICV bandwidths.

a bandwidth from the second local minimum is mitigated by using the rule (2.21). Indeed, the oversmoothed bandwidths for the three samples are shown by the vertical lines in Figure 11 and were 0.7404, 0.7580 and 0.7341. Note that the problem with the ICV curve having two local minima of approximately the same value quickly goes away as the sample size increases. This is illustrated in Figure 11**(b)**, where we have plotted three $ICV\left(\frac{h}{C}\right)$ curves for the separated bimodal case with $n = 500$. Thus, the selection rule $\hat{h}^*_{ICV}$ given by (2.21) rather than just $\hat{h}_{ICV}$ appears to be useful mostly for small and moderate sample sizes.

9. Examples

In this Section we illustrate the use of ICV with five examples. The purpose of the first two examples is to compare the performance of ICV, LSCV, and Sheather-Jones plug-in methods for choosing a global bandwidth. The third example illustrates the

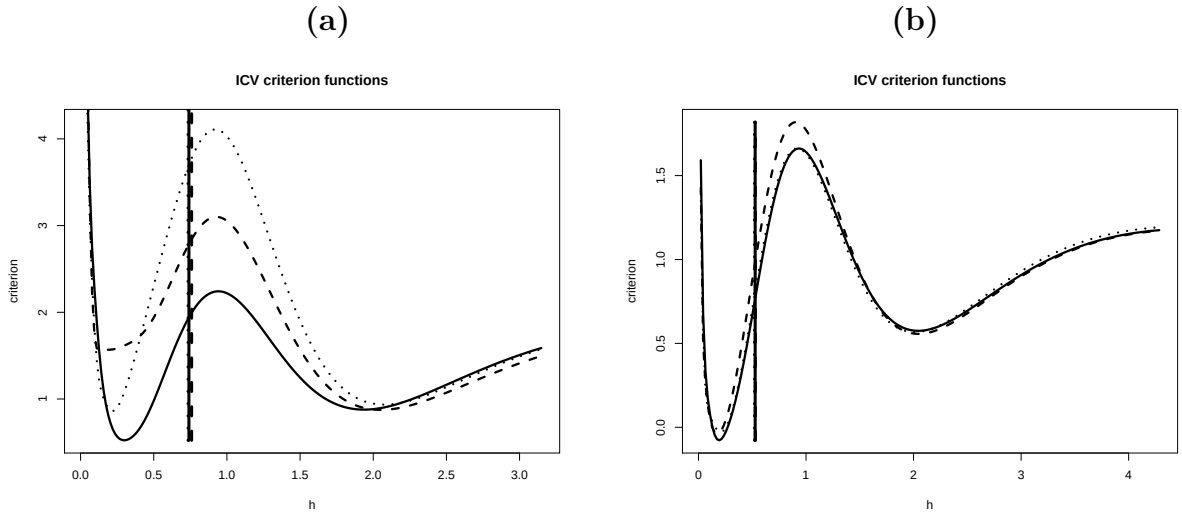


Fig. 11. Three $ICV\left(\frac{h}{C}\right)$ functions in the case of the separated bimodal density at **(a)** $n = 100$ and **(b)** $n = 500$. Vertical lines show the location of $\hat{h}_{OS}$.

benefit of using ICV for rounded data. The last two examples show an advantage of applying the ICV method locally.

9.1. Mortgage defaulters

In this example we analyze the credit scores of Fannie Mae clients who defaulted on their loans. The mortgages considered were purchased in "bulk" lots by Fannie Mae from primary banking institutions. The data set of size $n = 402$ was taken from the website http://www.dataminingbook.com associated with the book of Shmueli, Patel, and Bruce (2006).

The $LSCV(h)$ and $ICV\left(\frac{h}{C}\right)$ curves for the mortgage defaulters data are given in Figure 12. It turns out that the LSCV curve tends to $-\infty$ when $h \to 0$, but has a local minimum at about 2.84. In Figure 13 we have plotted an unsmoothed frequency histogram and the LSCV, ICV and Sheather-Jones plug-in density estimates for the

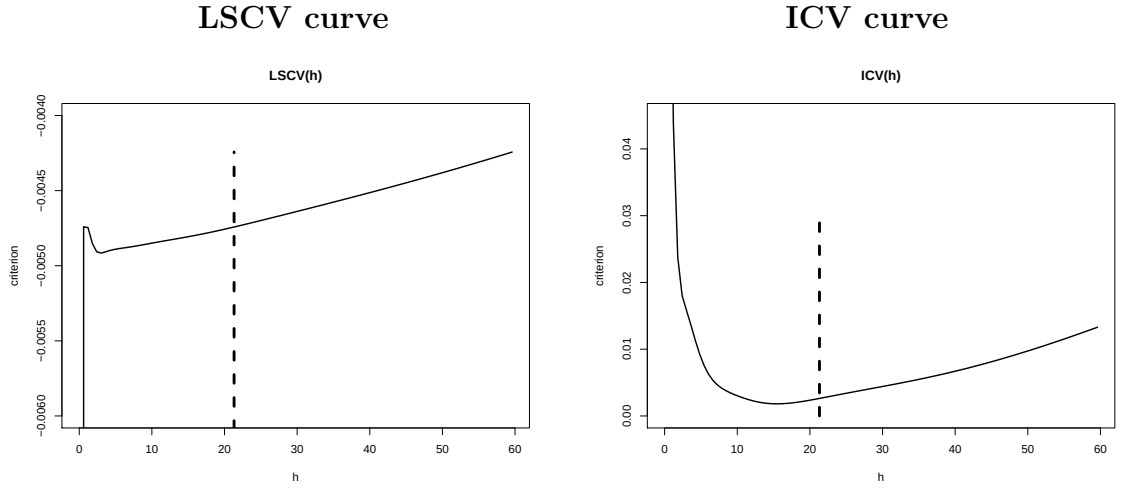


Fig. 12. $LSCV(h)$ and $ICV\left(\frac{h}{C}\right)$ curves for the data on credit scores for the defaulters. Vertical dashed lines show the location of the oversmoothed bandwidth $\hat{h}_{OS}$.

credit scores. The class interval size in the unsmoothed histogram was chosen to be 1, which is equal to the accuracy to which the data have been reported. We used the largest local minimizer of the LSCV curve, $\hat{h}_{UCV} = 2.84$, as suggested by Park and Marron (1990). The resulting LSCV estimate is severely undersmoothed. Both the Sheather-Jones plug-in and ICV density estimates show a single mode around 675 and look similar, with the ICV estimate being somewhat smoother.

Interestingly, a high percentage of the defaulters have credit scores less than 620, which many lenders consider the minimum score that qualifies for a loan; see Desmond (2008).

9.2. PGA data

In this example the data are the average numbers of putts per round played, for the top 175 players on the 1980 and 2001 PGA golf tours. The question of interest is whether there has been any improvement from 1980 to 2001. This data set has already

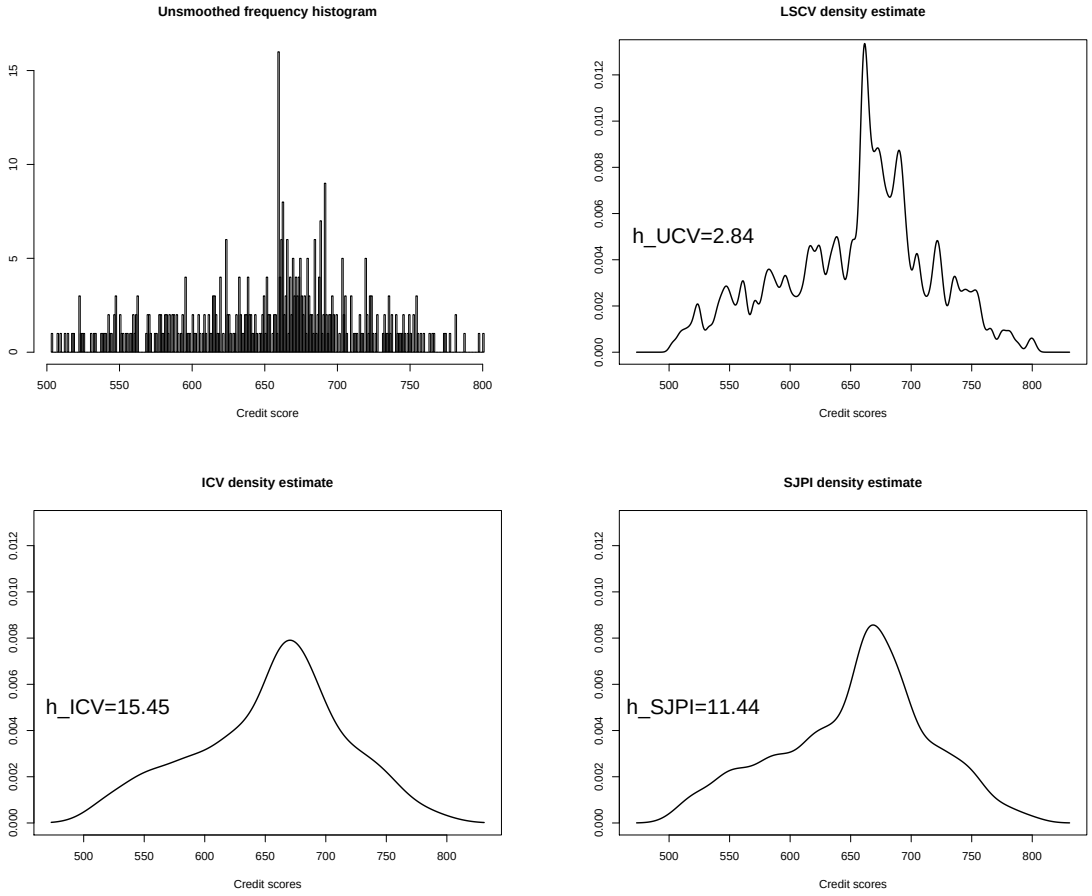


Fig. 13. Unsmoothed histogram and kernel density estimates for credit scores.

been analyzed by Sheather (2004) in the context of comparing the performances of LSCV and Sheather-Jones plug-in.

In Figure 14 we have plotted an unsmoothed frequency histogram and the LSCV, ICV and Sheather-Jones plug-in density estimates for a combined data set of 1980 and 2001 putting averages. The class interval size in the unsmoothed histogram was chosen to be 0.01, which corresponds to the accuracy to which the data have been reported. There is a clear indication of two modes in the histogram.

The estimate based on the LSCV bandwidth is apparently undersmoothed. The

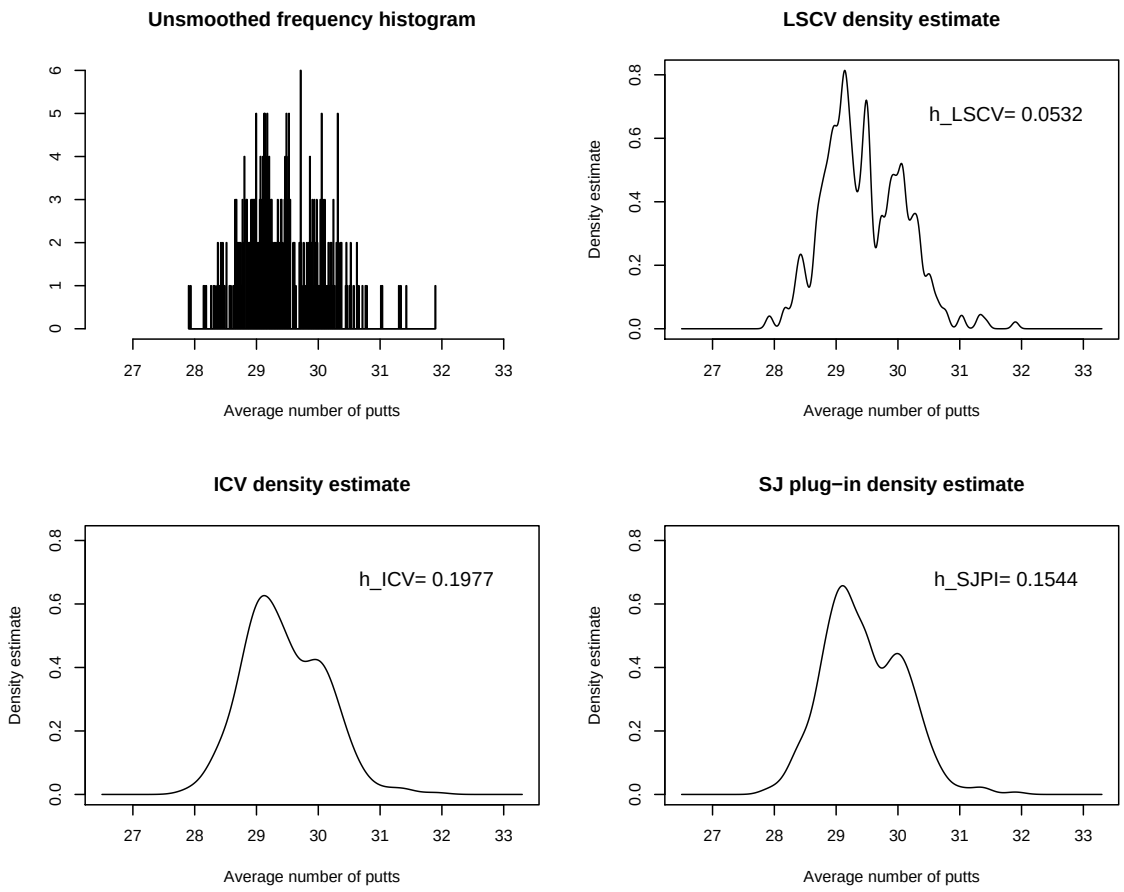


Fig. 14. Unsmoothed frequency histogram and kernel density estimates for average numbers of putts per round from 1980 and 2001 combined.

ICV and plug-in estimates look similar and have two modes, which agrees with evidence from the unsmoothed histogram and seems reasonable since the data were taken from two populations.

In Figure 15 we have plotted kernel density estimates separately for the years 1980 and 2001. ICV seems to produce a reasonable estimate in both years, whereas LSCV yields a very wiggly and apparently undersmoothed estimate in 2001.

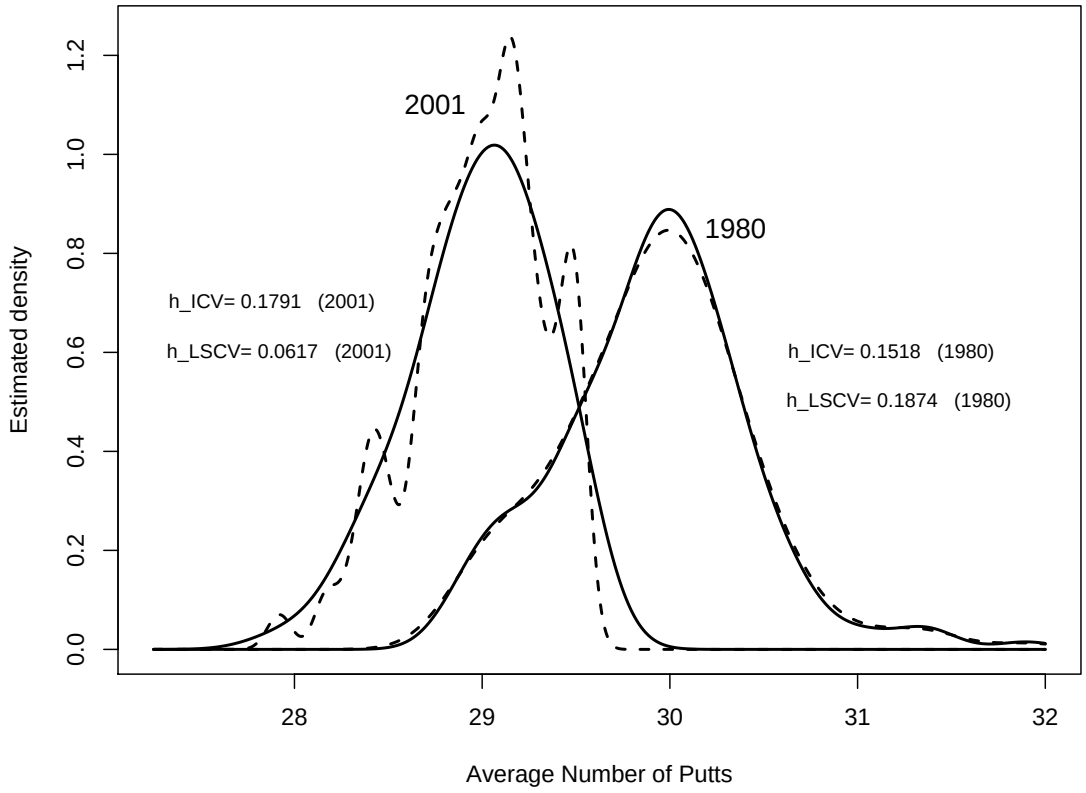


Fig. 15. Kernel density estimates based on LSCV (dashed curve) and ICV (solid curve) produced separately for the data from 1980 and 2001.

9.3. The Old Faithful geyser data

The data on the Eruption Duration of the Old Faithful geyser is a very popular example in the bandwidth selection literature. There are several versions of this data set. Our analysis deals with the data consisting of $n = 272$ observations given in Härdle (1991), which is different from the version used by Loader (1999a).

Observations in the original data set are given up to the precision of 0.001. Since our goal in this example is to show the failure of the LSCV method when the data are rounded, we rounded the observations up to the accuracy of 0.1. The $LSCV(h)$ and $ICV\left(\frac{h}{C}\right)$ curves for rounded data are plotted in Figure 16. As we can see, LSCV(h) $\to -\infty$ as $h \to 0$, and there is no local minimum in the LSCV curve, as in the example about mortgage defaulters. The $ICV\left(\frac{h}{C}\right)$ curve has two local minima of about the same size at $\hat{h}_1 = 0.0779$ and $\hat{h}_2 = 0.1253$. Notice that the LSCV bandwidth for the original data (unrounded) is equal to 0.1019 and lies almost exactly in the center of the interval $(\hat{h}_1, \hat{h}_2)$. The oversmoothed bandwidth $\hat{h}_{OS} = 0.4246$ falls above the two local minima. In this case the ICV bandwidth selection rule (2.21) will choose the bandwidth $\hat{h}^*_{ICV} = 0.0779$ which corresponds to the smaller of the two local minima. In fact, using either of the two bandwidths, $\hat{h}_1$ or $\hat{h}_2$, results in a seemingly reasonable estimate for the eruption duration density. The ICV density estimate based on the rounded data together with the LSCV estimate based on the original data are plotted in Figure 17. The two estimates are fairly close. So, for the rounded eruption duration data the ICV method yields a reasonable density estimate, whereas the LSCV method fails, selecting $h = 0$.

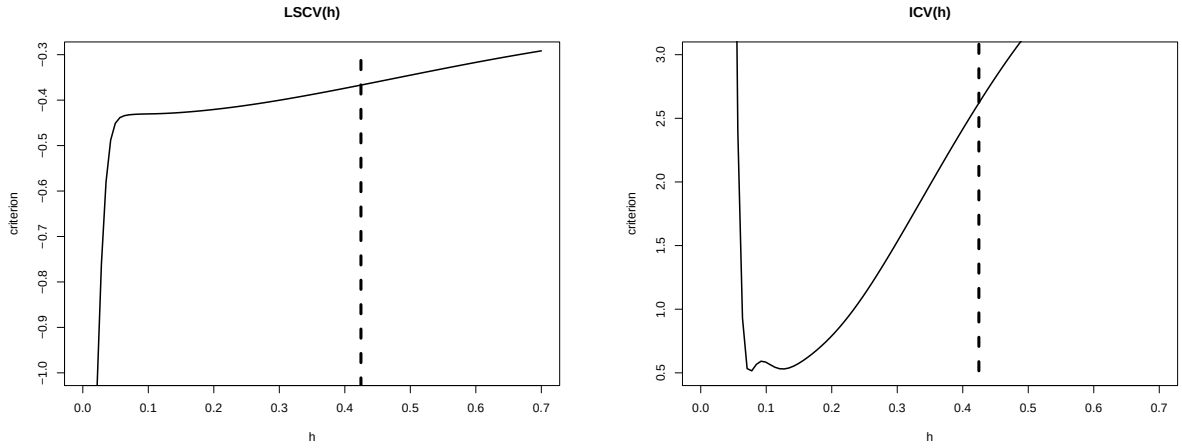


Fig. 16. The $LSCV(h)$ and $ICV\left(\frac{h}{C}\right)$ curves for the Old Faithful eruption duration data. Vertical dashed lines show the location of $\hat{h}_{OS}$.

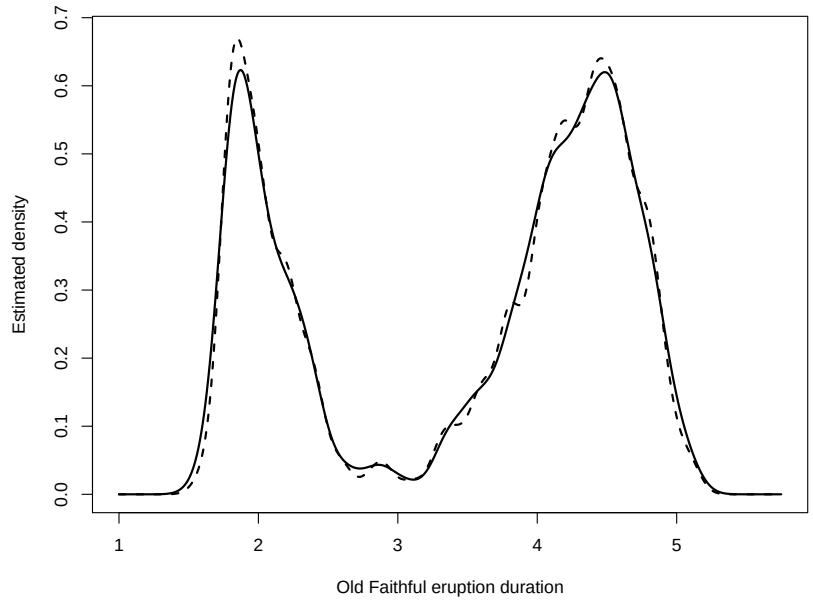


Fig. 17. LSCV density estimate based on the original data (solid curve) and ICV density estimate based on the rounded data (dashed curve).

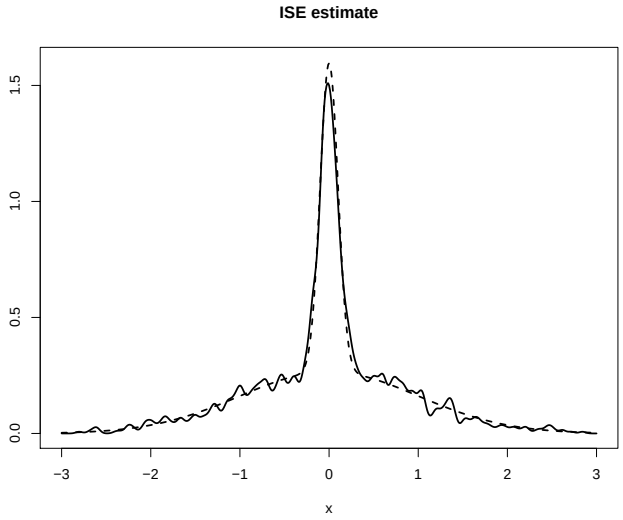


Fig. 18. The solid curve corresponds to the ISE density estimate, whereas the dashed curve shows the kurtotic unimodal density.

9.4. Local ICV: simulated example

For this example we took five samples of size $n = 1500$ from the kurtotic unimodal density defined in Marron and Wand (1992). First, we noted that even the bandwidth that minimizes $ISE(h)$ results in a density estimate that is much too wiggly in the tails. Figure 18 shows the ISE density estimate for one of the samples we considered. On the other hand, using local ICV resulted in much better density estimates.

We computed the local LSCV and ICV density estimates using four values of w ranging from 0.05 to 0.3. A selection kernel with $\alpha = 6$ and $\sigma = 6$ was used in local ICV. This (α, σ) choice performs well for global bandwidth selection when the density is unimodal, and hence seems reasonable for local bandwidth selection since locally the density should have relatively few features. For a given w, the local ICV and LSCV bandwidths were found for 61 points: $x = -3, -2.9, \ldots, 2.9, 3$, and were interpolated at other $x \in [-3, 3]$ using a spline. Average squared error (ASE) was

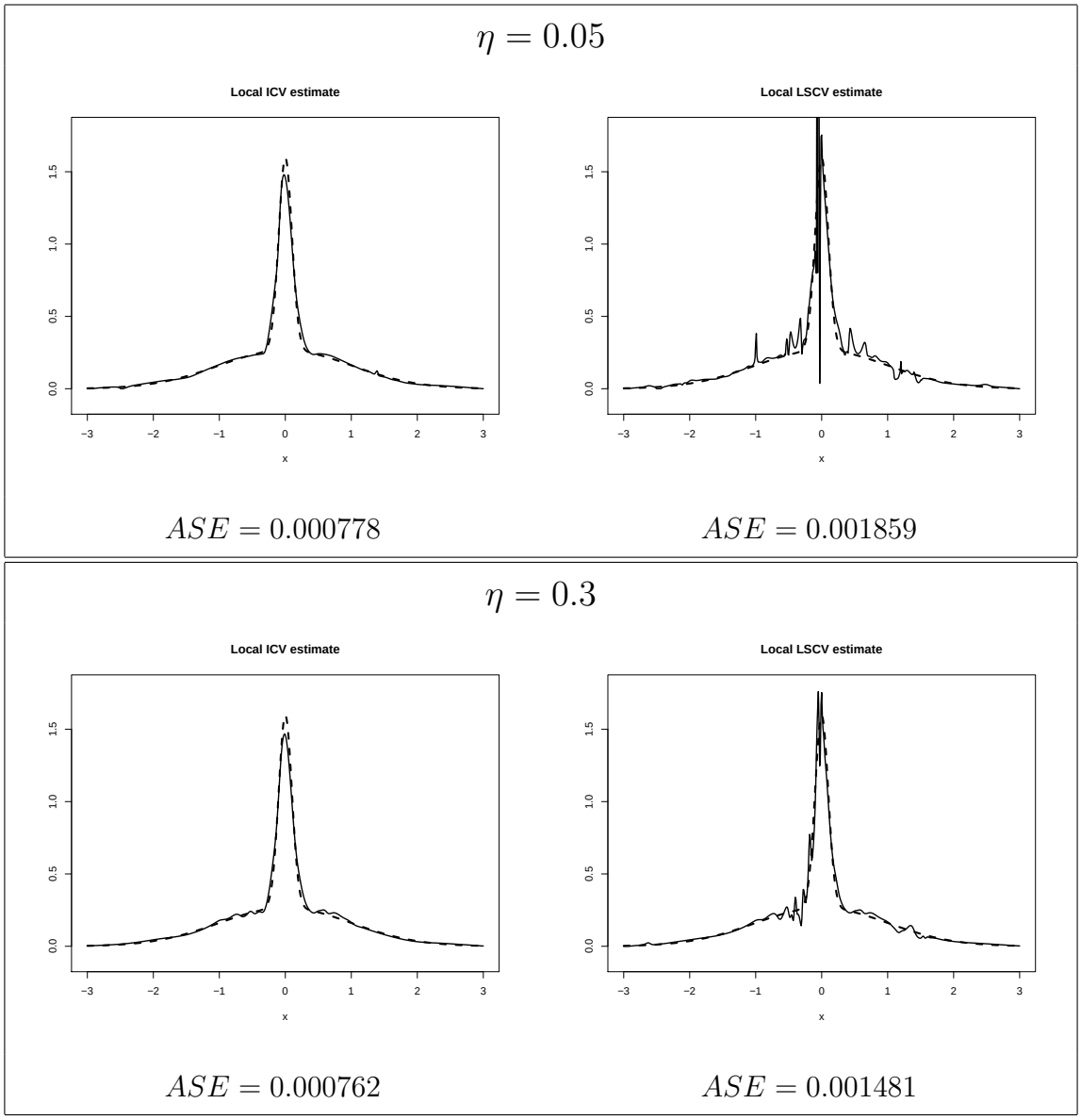


Fig. 19. The solid curves correspond to the local LSCV and ICV density estimates, whereas the dashed curves show the kurtotic unimodal density.

used to measure closeness of a local density estimate $\hat{f}_\ell$ to the true density f:

$$ASE = \frac{1}{61}\sum_{i=1}^{61}(\hat{f}_\ell(x_i) - f(x_i))^2.$$

The local ICV estimates were as smooth or smoother than the local LSCV estimates for all five samples considered. Figure 19 shows results for one of the samples, where the local LSCV method performed the worst. Estimates corresponding to the smallest and the largest values of w are provided. For this sample the local ICV method performed similarly well for all values of w considered, whereas all the local LSCV estimates were very unsmooth, albeit with some improvement in smoothness as w increased.

9.5. Local ICV: real data example

This example shows an advantage of local ICV over local LSCV. We analyze the data of size $n = 517$ on the Drought Code (DC) of the Canadian Forest Fire Weather index (FWI) system. DC is one of the explanatory variables which can be used to predict the burned area of a forest in the Forest Fires data set. This data can be downloaded from the website http://archive.ics.uci.edu/ml/datasets/Forest+Fires. The data were collected and analyzed by Cortez and Morais (2007).

We computed the LSCV, ICV and Sheather-Jones plug-in bandwidths for the DC data. The LSCV method failed by yielding $\hat{h}_{UCV} = 0$. The ICV and Sheather-Jones plug-in bandwidths were very close and produced similar density estimates. Figure 20 **(a)** gives the ICV density estimate. It shows two major modes connected with a wiggly curve, which indicates that varying the bandwidth with x may yield a smoother estimate of the underlying density. Local ICV and LSCV have been applied to the DC data. We used $w = 40$ for both methods and the selection kernel with $\alpha = 6$ and $\sigma = 6$ for local ICV. Let $x_{(i)}$, $i = 1, \ldots, n$, denote the ith member of the ordered

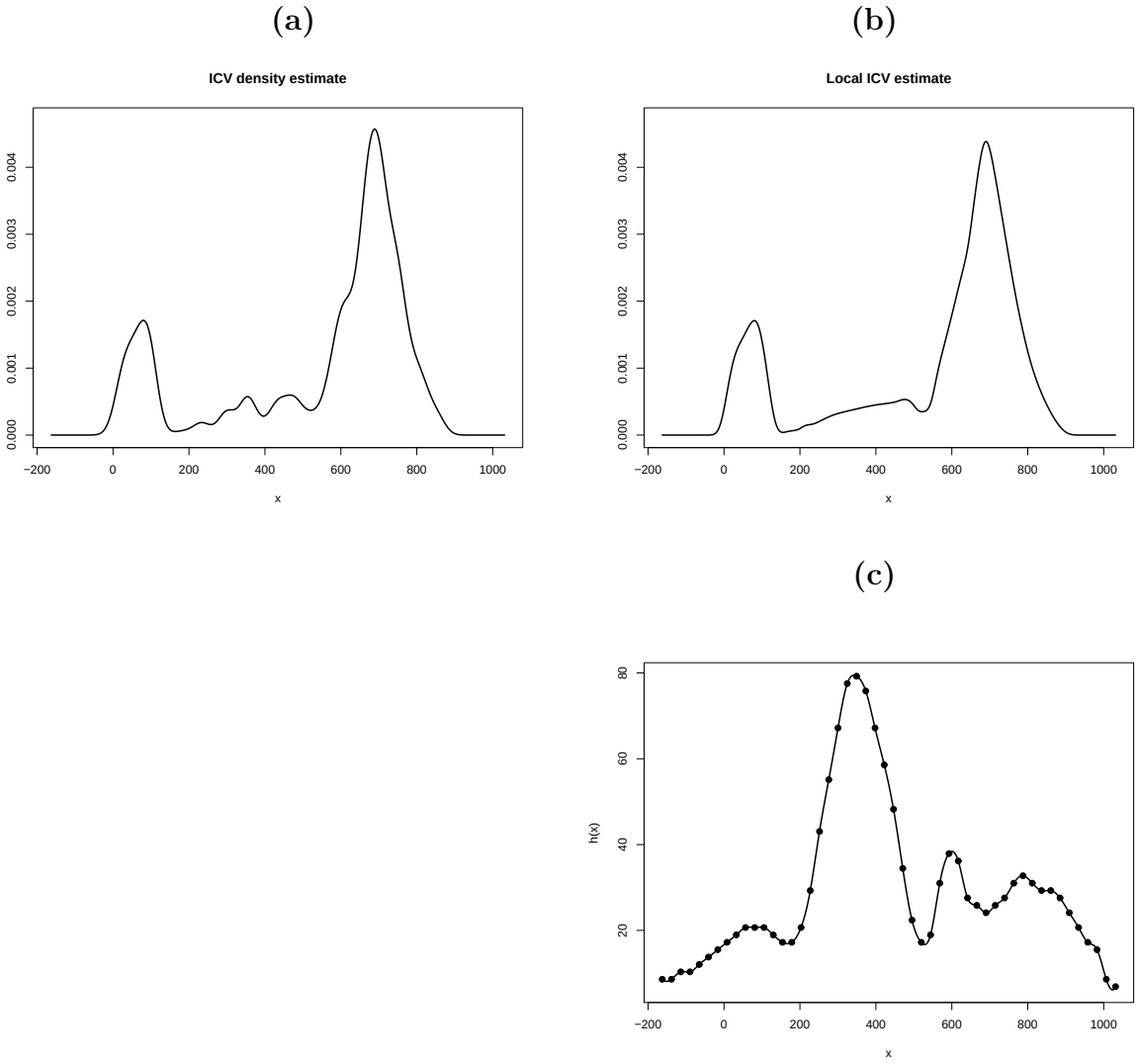


Fig. 20. Density estimates for the DC data set with **(a)** being the global ICV density estimate and **(b)** corresponding to the local ICV estimate; **(c)** Bandwidth function $\hat{h}(x)$ for Local ICV.

sequence of observations. The local ICV and LSCV bandwidths were found for 50 evenly spaced points in the interval $x_{(1)} - 0.2(x_{(n)} - x_{(1)}) \leq x \leq x_{(n)} + 0.2(x_{(n)} - x_{(1)})$. It turns out that in 45 out of 50 cases the local LSCV curve tends to $-\infty$ as $h \to 0$, which implies that the local LSCV estimate can not be computed. All 50 local ICV bandwidths were positive. A smooth bandwidth function $\hat{h}(x)$ shown in Figure 20 **(c)** was found by interpolating at other values of x via a spline. The corresponding local ICV estimate, given in Figure 20**(b)**, shows a smoother density estimate.

10. Summary

Indirect cross-validation is a method of bandwidth selection in the univariate kernel density estimation context. The method first selects the bandwidth of an L-kernel estimator by least squares cross-validation, and then rescales this bandwidth so that it is appropriate for use in a Gaussian kernel density estimator.

Selection kernels L have the form $(1 + \alpha)\phi(u) - \alpha\phi(u/\sigma)/\sigma$, where ϕ is the standard normal density and α and σ are positive constants. The interesting selection kernels in this class are of two types: unimodal, negative-tailed kernels and "cut-out the middle kernels," i.e., bimodal kernels that go negative between the modes. Large sample theory shows that the relative bandwidth error for both asymptotically optimal cut-out-the-middle kernels and negative-tailed kernels converge to 0 at a rate of $n^{-1/4}$, which is a substantial improvement over the $n^{-1/10}$ rate of LSCV. However, the best negative-tailed kernels yield bandwidths with smaller asymptotic mean squared error than do the best "cut-out-the-middle" kernels.

A practical purpose model for choosing the selection kernel parameters, α and σ, has been developed. The model was built by performing polynomial regression on the MSE-optimal values of $\log_{10}(\alpha)$ and $\log_{10}(\sigma)$ at different sample sizes for five normal mixture densities. Use of this model makes our method completely automatic.

A simulation study and examples reveal that using the model-based kernels in ICV leads to improved performance relative to ordinary LSCV.

An extensive simulation study showed that in finite samples ICV is more stable than LSCV. Although both ICV and LSCV bandwidths are asymptotically normal, the distribution of the ICV bandwidths for finite n is usually more symmetric and better concentrated in the middle of the density for ISE-optimal bandwidths. Using an oversmoothed bandwidth as an upper bound for the bandwidth search interval reduces the bias of the method and prevents selecting an impractically large value of h when the criterion curves exhibit multiple local minima.

The ICV method performs well in real data examples. ICV applied locally yields density estimates which are more smooth than estimates based on a single bandwidth. Often, local ICV estimates may be found when the local LSCV estimates do not exist.

CHAPTER III

ONE-SIDED CROSS-VALIDATION FOR NONSMOOTH REGRESSION FUNCTIONS

1. Introduction

Regression analysis is an area of statistics which studies the association between covariates and responses. In a nonparametric approach a regression function is not assumed have any specific parametric form. Nonparametric regression is studied in both fixed and random design contexts.

In the univariate fixed design case the design points $x_1 < x_2 < \cdots < x_n$ are non-random numbers, which are often specified before collecting the data. In this case the data $Y_1, \ldots, Y_n$ are assumed to come from the model

$$Y_i = r(x_i) + v(x_i)^{1/2}\varepsilon_i, \quad i = 1, \ldots, n,$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are mutually independent random variables, each having zero mean and unit variance. We call r the mean regression function, or simply the regression function, since $E(Y_i) = r(x_i)$, while v is called the variance function since $Var(Y_i) = v(x_i)$. Often it is assumed that $v(x_i) = \sigma^2$ for all i, in which case the model is called *homoscedastic*. Otherwise the model is heteroscedastic.

The random design regression model arises when we observe a bivariate sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ of random pairs, in which case the regression model can be written as

$$Y_i = r(X_i) + v(X_i)^{1/2}\varepsilon_i, \quad i = 1, \ldots, n,$$

where, conditional on $X_1, \ldots, X_n$, the ε_i are mutually independent with means equal to zero and the variances equal to one. It is also assumed that the errors $\varepsilon_1, \ldots, \varepsilon_n$

are independent of the design points $X_1, \ldots, X_n$. In the random design context

$$r(x) = E(Y|X = x) \quad \text{and} \quad v(x) = Var(Y|X = x),$$

are, respectively, the conditional mean and variance of Y given $X = x$. The marginal density of $X_1, \ldots, X_n$ will be denoted by f. In either the fixed or random design case, it may be assumed without loss of generality that the design points are distributed on the interval $[0, 1]$.

Kernel methods of estimating r include the Nadaraya-Watson estimator (see Nadaraya (1964) and Watson (1964)), Priestley-Chao estimator (see Priestley and Chao (1972)), the Gasser-Müller estimator (see Gasser and Müller (1979)), and the local linear estimator (see Fan (1992)). All the aforementioned methods require selecting a smoothing parameter, which is also called the bandwidth, as in the density estimation context.

Local linear estimators were introduced by Cleveland (1979) and studied by Fan (1992). For a given kernel K and bandwidth $h > 0$, the local linear estimator at a point x is computed as

$$\hat{r}_h(x) = \frac{\sum_{i=1}^n w_i(x) Y_i}{\sum_{i=1}^n w_i(x)}, \tag{3.1}$$

where

$$w_i(x) = K\left(\frac{x - x_i}{h}\right)\left(t_{n,2} - (x - x_i)t_{n,1}\right), \tag{3.2}$$

and

$$t_{n,j} = \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right)(x - x_i)^j, \quad j = 1, 2. \tag{3.3}$$

Most often, the kernel K is chosen to be a probability density function that is unimodal, symmetric about 0, and has finite variance. Fan (1992) showed that estimators (3.1) adapt to both fixed and random design scenarios, and have the same order of bias in the interior and boundary regions.

The bandwidth h determines the smoothness of the regression estimate $\hat{r}_h$. Inadequately small values of h produce "wiggly" estimates which follow the data too closely. Very large values of h lead to oversmoothed regression estimates which may miss some important features of the underlying regression function. An "optimal" h minimizes a measure of closeness of $\hat{r}_h$ to the true function r. Some popular measures include mean integrated squared error (MISE), average squared error (ASE), and mean average squared error (MASE). For simplicity we will consider the fixed design case below. In the random design case the ordinary expectations are replaced with the conditional expectations. The MISE function in the regression setting parallels that in the density estimation setting, and is defined in the following way:

$$MISE(h) = E\left(\int_0^1 \left(\hat{r}_h(x) - r(x)\right)^2 dx\right),$$

where $x_1, \ldots, x_n$ are the observed data values. The ASE function is given by

$$ASE(h) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{r}_h(x_i) - r(x_i)\right)^2. \tag{3.4}$$

The MASE function is defined as $E\left(ASE(h)\right)$. It can be shown that MASE is asymptotically equivalent to

$$MISE_w(h) = \int_{-\infty}^{\infty} E\left(\hat{r}_h(x) - r(x)\right)^2 f(x)\, dx. \tag{3.5}$$

Assuming that the design density f is continuous and positive in the interval $(0, 1)$, the regression function $r(x)$ has a bounded and continuous second derivative for $x \in (0, 1)$, and K is a second order kernel such that $R(K) < \infty$, the MASE function for the local linear estimator has the following asymptotic expansion:

$$MASE(h) = \frac{R(K)\sigma^2}{nh} + \frac{\mu_{2K}^2 h^4 \int_0^1 \left(r''(x)\right)^2 f(x)\, dx}{4} + o\left(h^4 + \frac{1}{nh}\right), \tag{3.6}$$

where we use the same definitions of functions $R(\cdot)$ and μ_{2K} as in (2.2).

Let h_0^* denote the bandwidth which minimizes the MASE function. From expression (3.6) it follows that h_0^* is asymptotic to

$$h_n^* = \left(\frac{R(K)\sigma^2}{\mu_{2K}^2 \int_0^1 \left(r''(x)\right)^2 f(x)\, dx} \right)^{1/5} n^{-1/5}. \tag{3.7}$$

Notice that when the design is fixed and evenly spaced or uniform, the asymptotic expansion (3.6) will hold and the formula (3.7) will be true if one takes $f(x) \equiv 1$.

One of the most frequently used data-driven bandwidth selection techniques for kernel regression estimators is the least-squares cross-validation (LSCV) method (see Stone (1977)), which parallels the LSCV method in the density estimation context. The LSCV bandwidth is the value of h which minimizes the cross-validation function defined by

$$CV(h) = \frac{1}{n} \sum_{i=1}^{n} \left(\hat{r}_h^{-i}(x_i) - Y_i \right)^2, \tag{3.8}$$

where $\hat{r}_h^{-i}$ is the leave-one-out regression estimator which is computed without using the ith observation (X_i, Y_i). The cross-validation function (3.8) is an approximately unbiased estimator of $\sigma^2 + MASE(h)$ (see Hart and Yi (1998)). It turns out that in the regression setting the cross-validation bandwidths have the same relative convergence rate of $n^{-1/10}$ (see Härdle, Hall, and Marron (1988)) as in the density estimation context. This slow convergence rate has the consequence of high variability of the LSCV bandwidths in practice. Additional details about the LSCV method may be found in the article of Hall and Johnstone (1992).

Plug-in is a popular alternative to cross-validation. The main idea of the plug-in method is to estimate the unknown terms in an expression for an asymptotically optimal bandwidth. There are different implementations of the plug-in idea, including the plug-in of Gasser, Kneip, and Köhler (1991) and the plug-in of Ruppert, Sheather,

and Wand (1995). The Gasser-Kneip-Köhler plug-in has an $O_p(n^{-1/5})$ relative rate, whereas the direct plug-in of Ruppert, Sheather, and Wand has a faster rate of $O_p(n^{-2/7})$. Although the direct plug-in has been seen to work well in practice for a wide variety of functions, it has certain shortcomings. In particular, it relies on the assumption that the regression function has four continuous derivatives, and it requires the data analyst to make a subjective choice of a nuisance parameter δ. A data example in the article of Hart and Yi (1998) illustrates how the Gasser-Kneip-Köhler plug-in local linear estimator may be sensitive to the choice of the analogous auxiliary parameter.

One of the modifications of the ordinary cross-validation method is the one-sided cross-validation method of Hart and Yi (1998). Although OSCV does not improve the LSCV convergence rate, it can achieve up to twentyfold reduction in asymptotic bandwidth variance. In a simulation study conducted by Hart and Yi (1998), the OSCV bandwidths are almost as stable as the Gasser-Kneip-Köhler plug-in bandwidths while being less biased. More simulation results for the OSCV method may be found in the article by Yi (2005). Other advantages of OSCV is that it is completely automatic, fairly robust to autocorrelation among the error terms (see Hart and Lee (2005)), and does not require more computing time than LSCV.

The OSCV theory is based on the assumption that the underlying regression function has two continuous derivatives. However, many physical, biomedical and economical processes involve nonsmooth or even discontinuous functions. For example, the speed and acceleration of a car can be interpreted as nonsmooth and discontinuous processes, respectively. Such examples motivated us to extend the OSCV methodology so that it continues to work well even if the regression function has fewer than two derivatives. We define an OSCV algorithm that produces asymptotically optimal bandwidths even when the regression function has

a discontinuous first derivative. Our methodology can be extended to deal with discontinuous functions as well, although we do not do so in this work.

The remainder of this chapter proceeds as follows. Section 2 contains a detailed description of the ordinary OSCV method and its proposed extensions. Simulation results in Section 3 and examples in Section 4 evaluate the performance of the proposed modifications of OSCV. Section 5 contains a brief summary of our findings.

2. OSCV methodology

This section is devoted to the theoretical results for OSCV. We start from a detailed description of the original OSCV method in Section 2.1. The OSCV methodology is extended for nonsmooth regression functions in Section 2.2. In Section 2.3 we propose a generic OSCV algorithm for smooth and nonsmooth functions.

2.1. OSCV for smooth regression functions

The OSCV method is very similar in spirit to the ICV method described in the previous Chapter. As in ICV, OSCV finds the bandwidth in two steps:

(**Step1**) Select the bandwidth of a kernel estimator based on a special (one-sided) kernel L using ordinary LSCV.

(**Step2**) Multiply the bandwidth obtained in Step 1 by a known constant C and use the resulting bandwidth to estimate the regression function using the K-kernel estimator.

Even though the OSCV method can be used for the Priestley-Chao and Gasser-Müller estimators, we will most often use it for the local linear estimators. An appropriate choice for L will be discussed below. The most popular choices for K include quartic,

Epanechnikov, and Gaussian kernels (see Wand and Jones (1995)). The rescaling constant C in Step 2 has the following form:

$$C = \left(\frac{R(K)}{\mu_{2K}} \cdot \frac{\mu_{2L}}{R(L)} \right)^{1/5}, \tag{3.9}$$

which is motivated by the asymptotically optimal MASE bandwidth (3.7) and the fact that the cross-validation function (3.8) is an approximately unbiased estimator of $\sigma^2 + MASE(h)$. Notice that the constant (3.9) is identical to the rescaling constant for the ICV method, defined by expression (2.7), so we can keep the same notation. Equality of the multiplicative constants for the two methods is a consequence of similarity of the MISE asymptotic expansion (2.4) in the density problem and the MASE expansion (3.6) in the regression problem.

For practical implementation of the OSCV algorithm it is proposed to perform cross-validation on a special (one-sided) estimator $\tilde{r}_b$. For each point x the one-sided estimator $\tilde{r}_b(x)$ is defined as the K-kernel local linear estimator computed from the data points (x_i, Y_i) for which $x_i \le x$. To a good approximation, $\tilde{r}_b$ is a local linear estimator with kernel L defined by

$$L(u) = 2K(u)\frac{c_2 - uc_1}{c_2 - 2c_1^2} I_{(0,\infty)}(u), \tag{3.10}$$

where $c_i = \int_0^1 u^i K(u)\, du$, $i = 1, 2$; $I_A(\cdot)$ is an indicator of a set A. Note that kernel (3.10) is a second order kernel, unless $c_2^2 = c_1 c_3$, where $c_3 = \int_0^\infty u^3 K(u)\, du$. Also note that kernel (3.10) is the same as the boundary kernel of Gasser and Müller (1979). Figure 21 shows the quartic kernel and its one-sided counterpart, which are defined as

$$\begin{aligned} K_Q(u) &= \frac{15}{16}(1-u^2)^2 I_{(-1,1)}(u), \\ L_Q(u) &= \left(\frac{160}{27} - \frac{350}{27} u \right) (1-u^2)^2 I_{(0,1)}(u). \end{aligned} \tag{3.11}$$

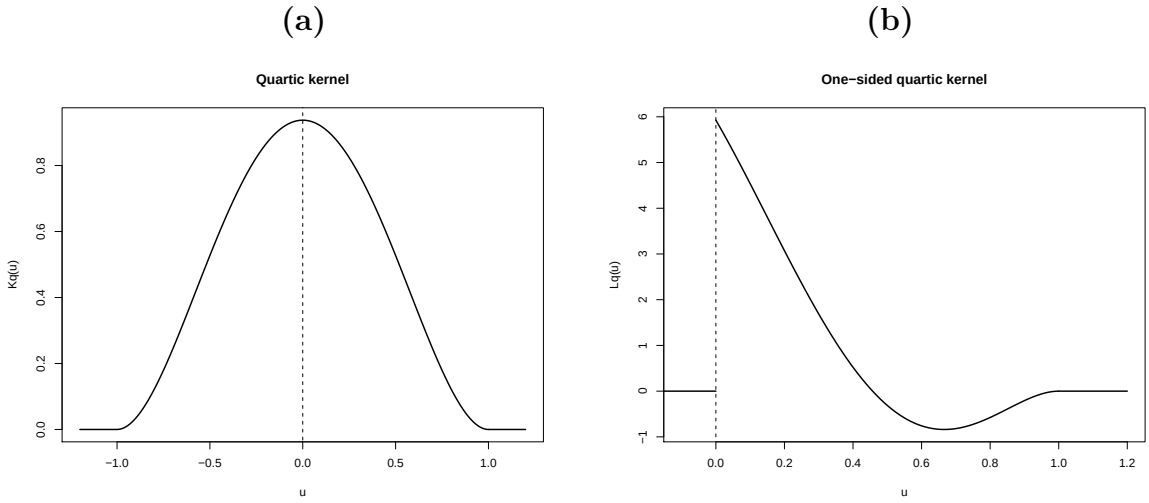


Fig. 21. **(a)**Quartic kernel K_Q; **(b)** One-sided quartic kernel L_Q.

The cross-validation function for the one-sided estimator $\tilde{r}_b$ has the following form:

$$OSCV(h) = \frac{1}{n-m} \sum_{i=m+1}^{n} \left(\tilde{r}_b^i(x_i) - Y_i \right)^2, \tag{3.12}$$

where m is an integer greater than 1; $\tilde{r}_b^i(x_i)$ is a one-sided estimator computed without the observation (x_i, Y_i). Omitting the first m points in the OSCV function (3.12) is necessary to ensure a reasonable one-sided prediction of the regression function r at the $(m+1)^{\text{st}}$ point. For practical purposes is it usually enough to take $m = 4$ (see Hart (1997)).

One-sided estimators $\tilde{r}_b$ have very low efficiency for estimating the regression function r, but are highly efficient for cross-validation purposes. Let $\hat{h}_{OSCV}$ and $\hat{h}_{CV}$ denote the OSCV and LSCV bandwidths, respectively. Hart and Yi (1998) showed that under appropriate conditions, the following result holds for the quartic kernel:

$$\lim_{n\to\infty} \frac{Var(\hat{h}_{OSCV})}{Var(\hat{h}_{CV})} \approx 0.10.$$

The above ratio may be smaller for other kernels, thus explaining the rationale for using the OSCV two-step bandwidth selection algorithm. A more detailed theoretical discussion of the OSCV method and its practical performance may be found in Yi (1996), Hart (1997), Hart and Yi (1998), and Yi (2001).

2.2. OSCV for nonsmooth regression functions

The OSCV method can be extended for nonsmooth regression functions. The same two-step procedure is appropriate with the constant C replaced by a different constant B, which depends on K and L in a different way. An expression for B follows from the MASE asymptotic expansion which is valid when the regression function r is nonsmooth. The MASE asymptotic expansion for the K-kernel local linear estimator in the case when r is nonsmooth and the design is fixed and evenly spaced or random uniform, has the following form:

$$MASE(h) = \frac{R(K)\sigma^2}{nh} + h^3 B_K \sum_{t=1}^{k} \left(r'(u_t+) - r'(u_t-)\right)^2 + o\left(\frac{1}{nh}\right) + o(h^4), \quad (3.13)$$

where $\{u_t\}$, $t = 1, \ldots, k$, are the points where the function r has cusps, and for an arbitrary function g,

$$B_g = \int_0^1 \left\{z(1 - H_g(z)) + G_g(z)\right\}^2 \, dz + \int_0^1 \left\{zH_g(-z) + G_g(-z)\right\}^2 \, dz,$$

$$H_g(z) = \int_{-\infty}^{z} g(u)\, du, \text{ and}$$

$$G_g(z) = \int_{-\infty}^{z} ug(u)\, du.$$

Derivation of expansion (3.13) and sufficient conditions for it to hold are given in the Appendix C. In particular, the kernel K in (3.13) is asumed to be supported on $[-1, 1]$, and have two continuous derivatives on its support. It is a topic of future work to prove that the result (3.13) holds when K is piecewise continuous and piecewise

twice differentiable. From expression (3.13) it follows that the MASE asymptotic minimizer in the nonsmooth case has the following form:

$$h_n^* = \left(\frac{\sigma^2}{3\sum_{t=1}^{k} \left(r'(u_t+) - r'(u_t-) \right)^2} \right)^{1/4} \left(\frac{R(K)}{B_K} \right)^{1/4} n^{-1/4}. \tag{3.14}$$

Denote by $AMASE^*(h)$ the leading terms in the asymptotic MASE expansion (3.13). It follows that in the nonsmooth case the minimum asymptotic MASE is given by

$$AMASE^*(h_n^*) = \\ \frac{4}{3} \left(\sigma^2 \int_{-\infty}^{\infty} \frac{1}{f(x)}\, dx \right)^{3/4} \left(3\sum_{t=1}^{k} \left(r'(u_t+) - r'(u_t-) \right)^2 \right)^{1/4} R(K)^{3/4} B_K^{1/4} n^{-3/4}.$$

Note the slower MASE convergence rate of $O(n^{-3/4})$ compared to the rate of $O(n^{-4/5})$ in the smooth case.

The asymptotic MASE-optimal bandwidth (3.14) implies the following formula for the rescaling constant B in the nonsmooth case:

$$B = \left(\frac{R(K)}{B_K} \right)^{1/4} \left(\frac{B_L}{R(L)} \right)^{1/4}. \tag{3.15}$$

Notice that the use of asymptotic expansion (3.13) is not yet justified for one-sided kernels L, which are often discontinuous at 0, and have support [0,1]. However, our extensive numerical experience suggests that the constant B given by (3.15) helps to remove the bandwidth bias in the case of a nonsmooth regression function. Table VI shows the values of the constants B and C for the most frequently used kernels. As we can see, most of the traditionally used kernels have $C > B$. However, the discrepancy between the constants, measured by $\left|\frac{B}{C} - 1\right| \cdot 100\%$, is less than 7% for all the kernels in Table VI except for the Gaussian, in which case the discrepancy is 14.33%. The question is "how will the bias introduced by a wrong constant (C) in the nonsmooth case impact an estimator's error?" To address this question, we introduce the penalty

Table VI. Rescaling constants B and C and the penalty for using a wrong constant in the nonsmooth case.

Kernel	C	B	$\left\lvert\frac{B}{C}-1\right\rvert \cdot 100\%$	R, %
Epanechnikov	0.5371	0.5019	6.55	0.7215
quartic	0.5573	0.5206	6.59	0.7285
triangle	0.5493	0.5133	6.55	0.7195
Gaussian	0.6168	0.5284	14.33	4.0210

measure R, defined in the following way:

$$R = \left| \frac{AMASE^*(Cb_n^*)}{AMASE^*(h_n^*)} - 1 \right| \cdot 100\%,$$

where b_n^* is the asymptotic MASE-optimal bandwidth (3.14) computed for the L-kernel estimator. The quantity Cb_n^* is a proxy to the bandwidth which would the ordinary OSCV method select if (inappropriately) applied in the case of a nonsmooth regression function. It can be shown that

$$R = \frac{3}{4}\left(x + \frac{1}{3x^3}\right),$$

where

$$x = \left(\frac{B_L}{B_K}\right)^{1/4} \left(\frac{C_K}{C_L}\right)^{1/20} \left(\frac{\mu_{2K}^2}{\mu_{2L}^2}\right)^{1/5}.$$

The last column of Table VI gives the value of R for the traditional kernels. It follows that R is less than 1% for all the kernels expect for the Gaussian, in which case $R \doteq 4\%$. This suggest that the ordinary OSCV method is fairly robust to nonsmoothness of the regression functions.

2.3. Robust OSCV

As we concluded in Section 2.2, inappropriate use of ordinary OSCV for nonsmooth functions produces biased bandwidths, since the constants of proportionality C and B are generally different. We propose to modify the OSCV method so that it has equal rescaling constants $B = C$ and, consequently, does not require the knowledge of the regression function's smoothness.

Notice that the constants C and B depend on the kernels K and L exclusively, implying that for a given kernel K one can search for a one-sided kernel L^* which will produce the desired equality of the constants. Such a kernel L^* is called *robust*, and OSCV based on a robust kernel L^* is called *robust OSCV*.

For a practical implementation of the above idea one can first define a parametric family of two-sided kernels K^*, and then find L^* using expression (3.10). The parameters in K^* and L^* are determined by requiring $B = C$.

In our initial efforts we used the quartic kernel (3.11) for K and used polynomials on the interval $[-1, 1]$ for K^*. We found 18 robust kernels in this setting, with two kernels performing better than the others in numerical studies. These winning kernels are defined below and plotted in Figure 22.

- **Kernel** K_1^*:

$$K_1^*(x) = \left(1.1393 - 2.4221x - 0.6640x^2 + 4.0368x^3 - 2.0901x^4\right) I_{[-1,1]}(x).$$

  For this kernel $B = C = 0.3030$. Both kernels K_1^* and L_1^* are continuous and smooth at $x = 1$.

- **Kernel** K_2^*:

$$K_2^* = \left(1.3822 - 0.8338x - 5.2104x^2 + 3.8913x^3 + 4.2729x^4 - 3.5022x^5\right) I_{[-1,1]}(x).$$

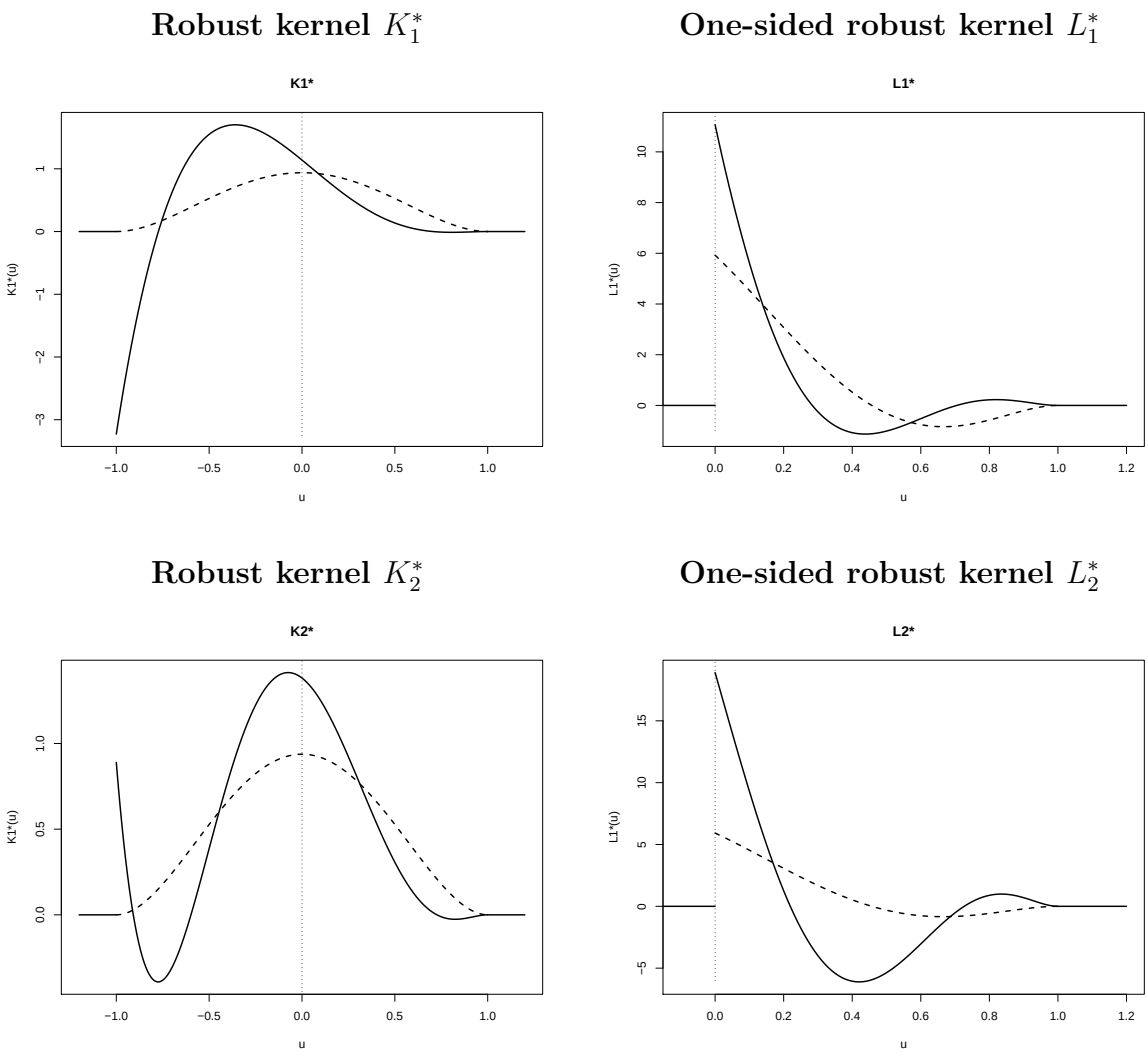


Fig. 22. Kernels K_1^* and K_2^* and the corresponding one-sided kernels L_1^* and L_2^*. The dashed lines correspond to the two-sided and one-sided quartic kernels.

For this kernel $B = C = 0.3385$. Kernels K_2^* and L_2^* are continuous and smooth at $x = 1$, and the third moment of K_2^* is zero.

Notice that the one-sided kernels L_1^* and L_2^* are discontinuous and fairly steep at $x = 0$, which matches the properties of "good" kernels of Hart and Yi (1998). In our simulation studies OSCV based on the robust kernels K_1^* and K_2^* performed comparable to ordinary OSCV based on the quartic kernel (3.11) for a wide variety of functions when the design points were fixed and evenly spaced. Even though the OSCV curves based on K_1^* and K_2^* were smooth in all our examples involving the fixed and evenly spaced design, the criterion curves for K_1^* and K_2^* in the case of the random design are usually very irregular. The following numeric example was used to illustrate this observation.

We generated $n = 100$ design points from the Uniform$(0, 1)$ distribution. Data points were produced using a smooth function with a moderate amount of added Gaussian noise. We computed the OSCV criterion curve according to expression (3.12). We used a correction for small h when computing the leave-one-out one-sided predictor $\tilde{r}_b^{-i}$ in the OSCV function (3.12), since the local linear estimator based on a kernel supported on $[-1, 1]$ is not well-defined (has a denominator equal to 0) if h is less than the largest spacing in the design points. The resulting OSCV curve plotted in Figure 23 is extremely unsmooth and behaves similarly to a discontinuous function. In fact, the OSCV curve has many spikes of a large amplitude with most of them occurring at the relatively small values of h. What is the reason for those spikes?

We found that the erratic behavior of the OSCV curve plotted in Figure 23 is due to the properties of the kernel K_1^*. The OSCV criterion (3.12) is a function of the LLE based on the kernel K_1^*. Notice that the denominator of the local linear

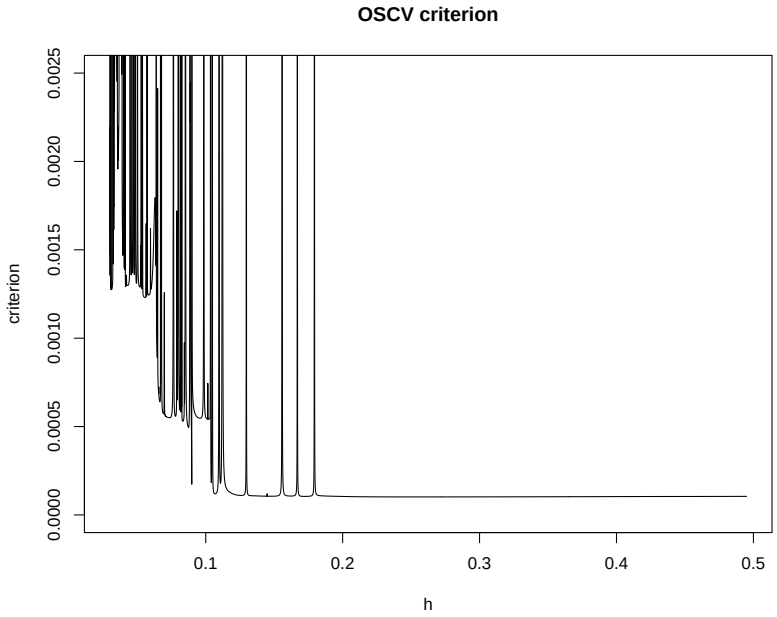


Fig. 23. An OSCV criterion function based on the kernel K_1^*.

estimator (3.1) has a sum of the weights (3.2) that depend on the kernel function. Since the kernel K_1^* is not nonnegative, the weights may add up to a small number at selected values of h, which will cause a spike in the OSCV criterion function for those h. In an attempt to solve the problem with rough criterion curves, we tried to use the one-sided Gasser-Müller estimator instead of the one-sided LLE in the cross-validation function (3.12). The Gasser-Müller estimator does not have a denominator and may not have a division by 0 problem. We noted that the OSCV curves for the one-sided Gasser-Müller estimator based on K_1^* were more smooth than those for the LLE based on K_1^*, but still unacceptably wiggly. We do not as yet have an explanation for this phenomenon.

Table VI suggests that the OSCV method does not really need correction when K is the quartic kernel. Our further efforts concentrated on K being the Gaussian

kernel, since this case corresponds to the most significant discrepancy between B and C and the largest penalty measure R. One of the parametric families we considered for K^* was

$$K^*(x) = (c_0 + c_1 x + c_2 x^2)\phi(x),$$

where $\phi(x)$ is the Gaussian kernel, and c_0, c_1, and c_2 are the parameters. We found four robust kernels from this family, none of which is nonnegative, and, similarly to the kernels K_1^* and K_2^*, produce very rough criterion curves in the random design case. Apparently, the nonnegativity is a necessary property for a "good" cross-validation kernel.

Next, we considered several parametric families of positive kernels and found eight robust kernels, all of which are bimodal. Two of the kernels are defined below and plotted in Figure 24.

- **Kernel** K_3^*:

$$K_3^*(x) = \frac{1}{2\sigma}\phi\left(\frac{x+\mu}{\sigma}\right) + \frac{1}{2\sigma}\phi\left(\frac{x-\mu}{\sigma}\right),$$

where $\alpha = 0.4121$ and $\sigma = \dfrac{1}{10}$. For this kernel $B = C = 0.1932$.

- **Kernel** K_4^*:

$$K_4^*(x) = \frac{1}{3\sigma}\phi\left(\frac{x+\frac{2}{3}\alpha}{\sigma}\right) + \frac{2}{3\sigma}\phi\left(\frac{x-\frac{\alpha}{3}}{\sigma}\right),$$

where $\alpha = 2.3729$ and $\sigma = \dfrac{1}{5}$. For this kernel $B = C = 0.3786$.

Apparently, the bimodal structure of the kernels causes the wiggles in the OSCV criterion curves. Figure 25 shows two OSCV functions computed using kernel K_3^* for $n = 100$ data points generated from a smooth function with a moderate amount of added Gaussian noise in the cases of the fixed, evenly spaced design and a random design, where the design density was a mixture of the Uniform$[0, 1]$ distribution and a beta distribution.

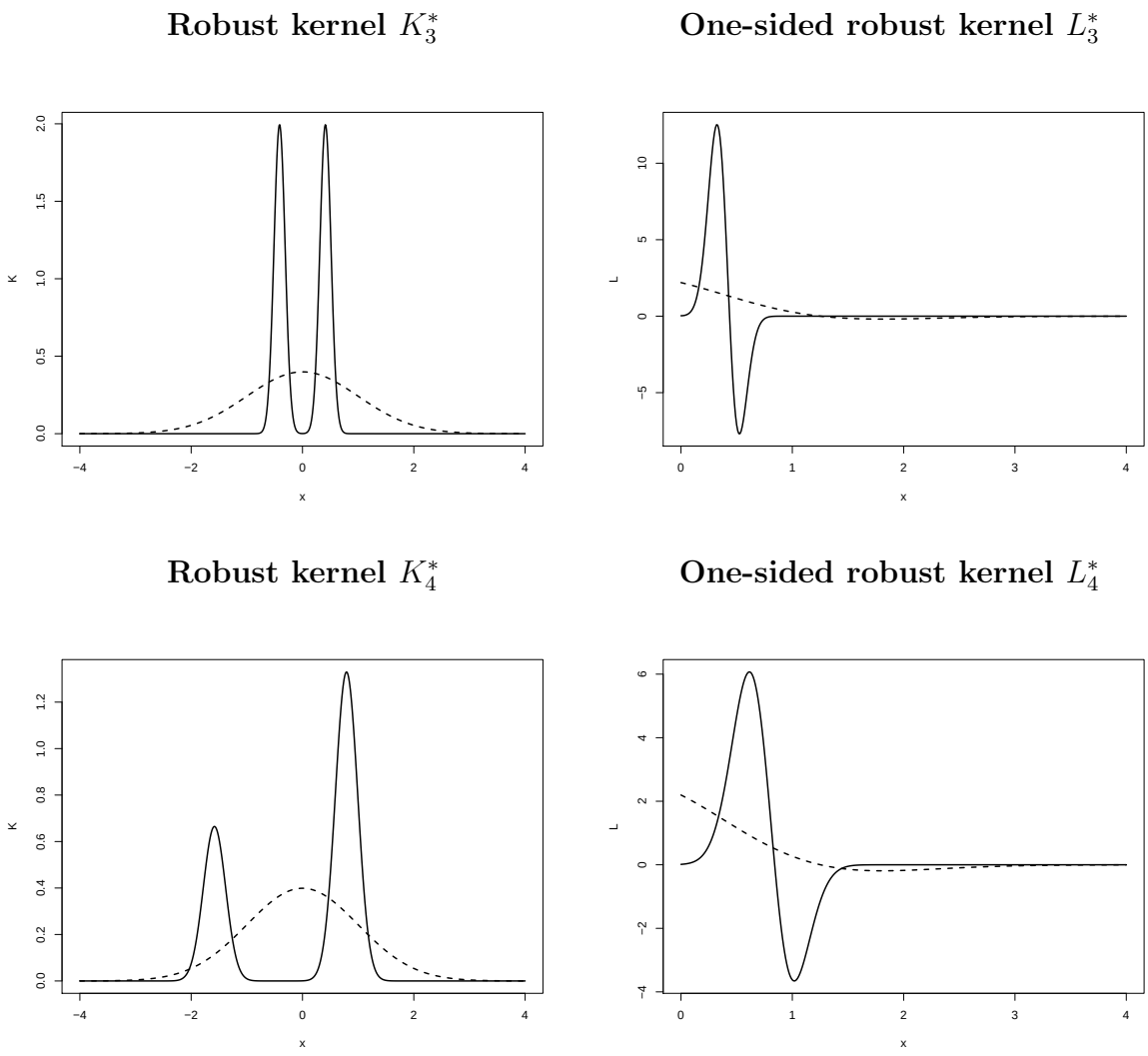


Fig. 24. Kernels K_3^* and K_4^* and the corresponding one-sided kernels L_3^* and L_4^*. The dashed lines correspond to the two-sided and one-sided Gaussian kernels.

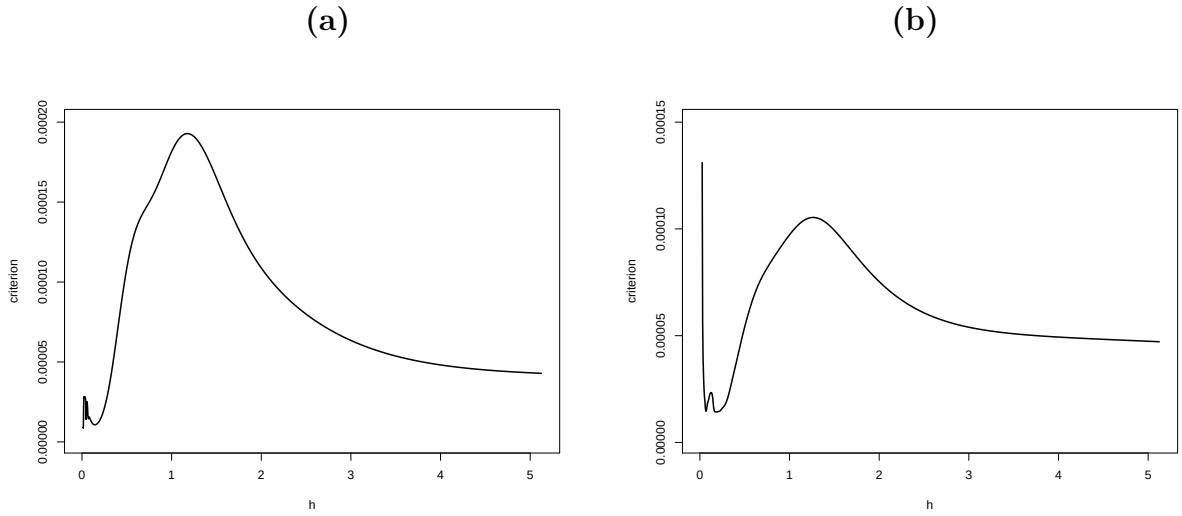


Fig. 25. OSCV criterion functions based on the kernel K_3^* in the cases of **(a)** fixed evenly spaced design and **(b)** random design.

Although none of the robust kernels so far discussed is nonnegative and unimodal, we found a kernel which is "close" to satisfying these conditions. This kernel is a member of the parametric family

$$K^*(x) = (1+\alpha)\phi(x) - \frac{\alpha}{\sigma}\phi(x), \tag{3.16}$$

and has $\alpha = 0.000088$ and $\sigma = 10$. For this kernel $B = C = 0.5217$. In what follows we will call this kernel just K^* and will denote its constant C^*. Notice that the parametric family (3.16) exactly matches the selection kernels (2.8) used in the ICV method described in the previous chapter. Although family (2.8) includes the positive kernels (family $\mathcal{L}_2$), all the robust kernels we were able to find in (3.16) were either negative-tailed or of cut-out-the-middle type. Out of all the robust kernels we found, the kernel K^* is the closest to the Gaussian kernel in the L_2-sense. The kernel K^* is a negative-tailed kernel which crosses the horizontal axes at the points $x = \pm 4.85$. As

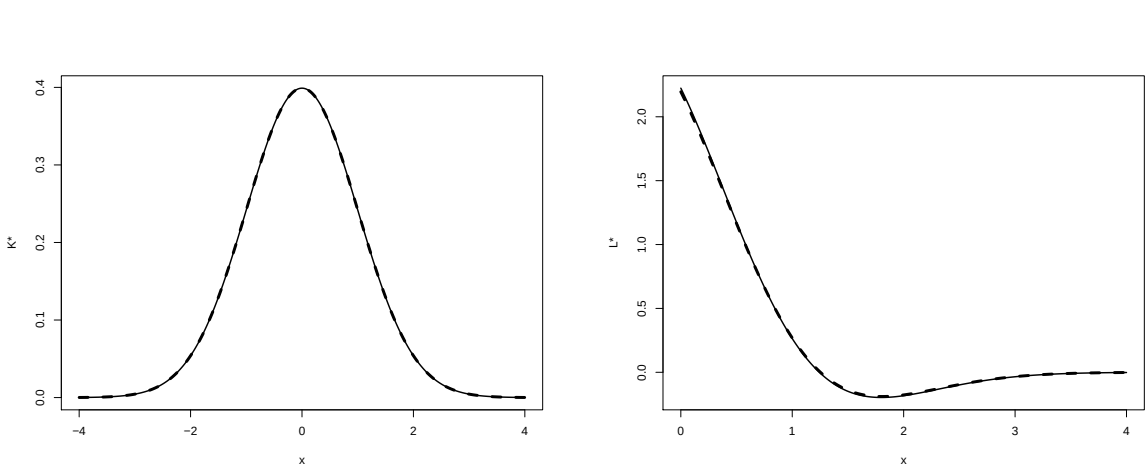


Fig. 26. Robust kernel K^* and its one-sided counterpart L^*. Dashed curves in both graphs correspond to the two-sided and one-sided Gaussian kernels.

we can see in Figure 26, the kernel K^* and its one-sided counterpart L^* are so close to the two-sided and one-sided Gaussian kernels, that they can not be distinguished by eye, at least on the intervals $(-4, 4)$ and $(0, 4)$, respectively.

We explored the OSCV curves produced by the kernel K^*. For the case of the fixed, evenly spaced design, the criterion curves were smooth for all combinations of functions and sample sizes considered. In the case of the Uniform$(0, 1)$ design we have occasionally observed some minor wiggles occurring at small values of h. In the case of the random design produced by mixing the uniform and a beta density, the criterion curves were not usually smooth for small values of h. Figure 27 shows an OSCV curve plotted for the case of the random design (f is a mixture of the uniform and a beta distribution), a nonsmooth function (with 6 cusps), $n = 100$, and a moderate amount of added Gaussian noise. Although the OSCV curve is very rough for small values of h, it is smooth in the area close to the point of its global

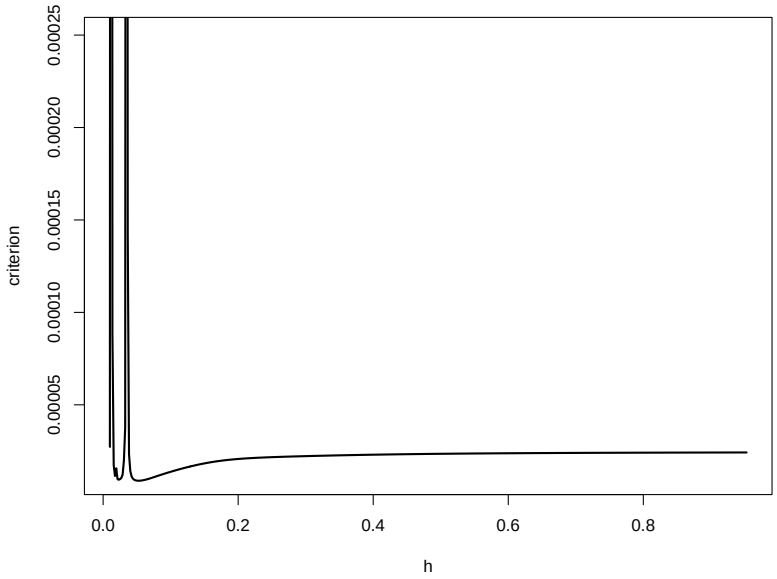


Fig. 27. An OSCV criterion function based on the kernel K^*.

minimum ($\hat{h} = 0.0516$).

The problem with the wiggly criterion curves in random design cases can be resolved by following the design transformation approach proposed by Hall, Park, and Turlach (1998). In the design transformation method the criterion curve is computed after transforming the design sequence to the fixed evenly spaced grid of points. As we noted above, the kernel K^* produces smooth OSCV curves for the fixed evenly spaced design case. One of the examples in Section 4 shows the benefit of using the OSCV method based on K^* after performing the design transformation.

The kernel K^* is the "best" robust kernel we have found thus far. A numerical study described in Section 3 and examples in Section 4 provide more information about the performance of K^*.

3. Simulation study

Here we present the results of a simulation study which compares ordinary OSCV and Robust OSCV based on kernel K^*. However, we will also provide simulation results for LSCV and Ruppert-Sheather-Wand plug-in. Our simulation setup is described next.

We used the following three functions in the study, where in each case $0 \leq x \leq 1$:

- Regression function r_1:

$$r_1(x) = 2.5(2x^{10}(1-x)^2 + x^2(1-x)^{10})$$

- Regression function r_2:

$$r_2(x) = \begin{cases} \frac{1}{20} - \frac{1}{20}|x - \frac{1}{4}|, & x < 0.5, \\ \frac{1}{20}|x - \frac{3}{4}| - \frac{1}{80}, & x \geq 0.5. \end{cases}$$

- Regression function r_3, plotted in Figure 28, which was produced by stacking 7 functions of different types including polynomials, exponential and logarithmic functions.

The function r_1 is smooth and has two peaks, r_2 has a single cusp at the point $x = 0.5$, and r_3 has six cusps. The range of each function is about the same, and hence the same values of σ, 1/250, 1/500, and 1/1000, were used with each function to represent high, moderate, and low levels of noise. We considered the three sample sizes $n = 100$, 300 and 1000. The error terms were taken to be $N(0, \sigma^2)$. We considered two designs: a fixed, evenly spaced design with

$$x_i = \frac{i - 0.5}{n}, \quad i = 1, \ldots, n,$$

and the Uniform$[0, 1]$ design. For each combination of r, σ, n, and the design we

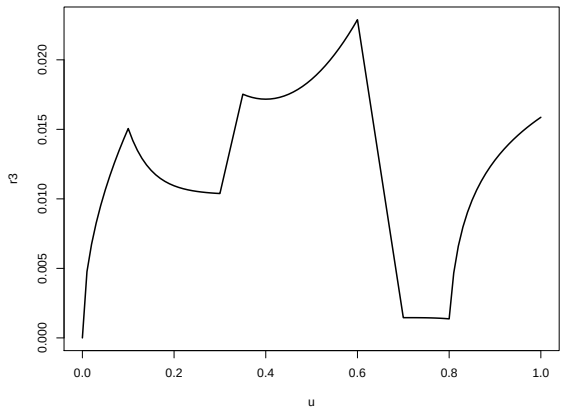


Fig. 28. Regression function r_3.

generated 1000 independent data sets. Figure 29 shows the regression functions r_1 and r_2 along with the $n = 100$ evenly spaced data points, generated using different values of σ.

We computed the bandwidth of the Gaussian local linear estimator using different data-based methods. Let $\hat{h}_0$ denote the minimizer of the ASE function (3.4) for the Gaussian local linear estimator, $\hat{h}_{PI}$ denote the Ruppert-Sheather-Wand plug-in bandwidth, $\hat{h}_{ROSCV}$ denote the bandwidth corresponding to the Robust OSCV method which uses kernel K^* for cross-validation, and, finally, $\hat{h}_{OSCV}$ stands for the ordinary OSCV bandwidth when the Gaussian kernel is used at the cross-validation stage.

Let us introduce some further notation. For each random variable Y defined in each replication of our simulation, we denote the mean, standard deviation, and the median of Y over all replications (with r, σ, and n fixed) by $\widehat{\mathrm{E}}(Y)$, $\widehat{\mathrm{SD}}(Y)$ and $\widehat{\mathrm{Median}}(Y)$. For each sample we computed $\hat{h}_{ROSCV}$, $\hat{h}_{OSCV}$, $\hat{h}_{PI}$, $\hat{h}_{CV}$, and $\hat{h}_0$. To evaluate the bandwidth selectors we computed $\widehat{\mathrm{E}}\big(ASE(\hat{h})/ASE(\hat{h}_0)\big)$ and

r_1+noise with $\sigma = 1/250$ r_2+noise with $\sigma = 1/1000$

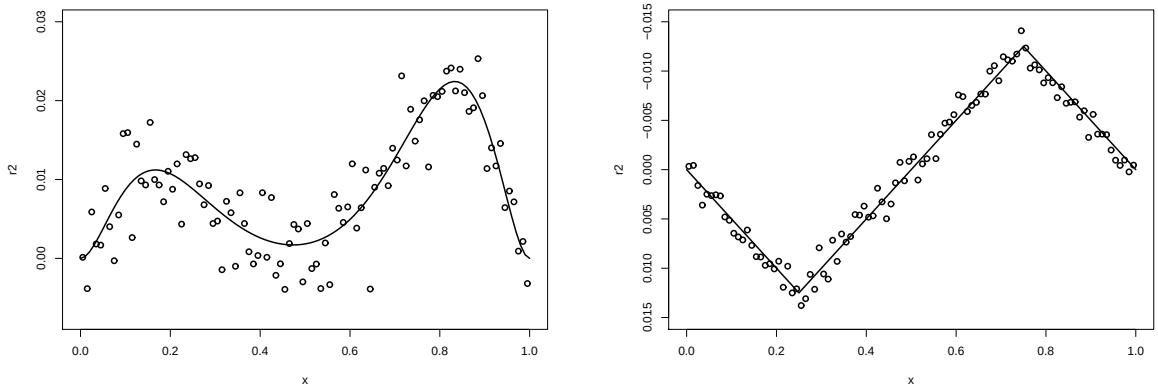


Fig. 29. Regression functions r_1 and r_2 with added noise. Design: fixed, evenly spaced; sample size: $n = 100$.

$\widehat{\text{Median}}\big(ASE(\hat{h})/ASE(\hat{h}_0)\big)$ for $\hat{h}$ equal to each of $\hat{h}_{ROSCV}$, $\hat{h}_{OSCV}$, $\hat{h}_{PI}$, $\hat{h}_{CV}$, and $\hat{h}_0$. To assess the bias of a data-driven method we define

$$\delta = \frac{|\widehat{\mathrm{E}}(\hat{h}) - \widehat{\mathrm{E}}(\hat{h}_0)|}{\widehat{\mathrm{E}}(\hat{h}_0)} \cdot 100\%, \tag{3.17}$$

which measures the relative distance between the average data-driven bandwidth $\hat{h}$ and the average ASE-optimal bandwidth $\hat{h}_0$.

Our simulation results for the functions r_1 and r_3 in the case of the fixed evenly spaced design are given in Tables VII and VIII and in Figures 30 and 31. The analogous results for the other cases considered are given in the Appendix D. Table IX contains the summary measures which were used to analyze our simulation results. The cell format in Table IX is of the form mean(standard deviation), where the mean and the standard deviation of a quantity are computed over all nine combinations of n and σ for a given regression function r and the type of design. Table IX contains

Table VII. Simulation results for r_1. Design: fixed, evenly spaced.

n	σ	**R OSCV**	**OSCV**	**PI**	**CV**	**ASE**
$\hat{\mathrm{E}}(\hat{h})$						
100	1/250	0.03199152	0.03770010	0.03595395	0.03701106	0.03769148
	1/500	0.02368298	0.02778172	0.02778078	0.02715639	0.02764088
	1/1000	0.01772376	0.02067085	0.02103406	0.01983358	0.02041447
300	1/250	0.02512314	0.02950298	0.02928123	0.02869285	0.02927278
	1/500	0.01870838	0.02184818	0.02207227	0.02112000	0.02165305
	1/1000	0.01387581	0.01613502	0.01661898	0.01552555	0.01596092
1000	1/250	0.01940042	0.02267263	0.02286507	0.02205911	0.02265662
	1/500	0.01440028	0.01675465	0.01718620	0.01627201	0.01669468
	1/1000	0.01060072	0.01230054	0.01291333	0.01195283	0.01224124
$\hat{\mathrm{SD}}(\hat{h}) \cdot 10^3$						
100	1/250	2.35334466	2.82035295	3.35478573	7.77987537	5.78855224
	1/500	1.36354160	1.64387452	1.61600631	4.91668289	3.77203463
	1/1000	0.93613228	1.11759307	0.85003210	3.44337129	2.45968021
300	1/250	1.22085401	1.47065950	1.94163024	4.65361265	4.00476245
	1/500	0.76985616	0.91728356	0.80729835	2.92937376	2.57599634
	1/1000	0.53843113	0.62557509	0.40084029	1.91421777	1.66549915
1000	1/250	0.68559950	0.81656455	0.83173119	2.83688178	2.67379715
	1/500	0.46144256	0.53254480	0.30664674	1.75928821	1.71484542
	1/1000	0.31790501	0.36135320	0.15607989	1.08983851	1.09198378
$\hat{\mathrm{E}}\big(\mathrm{ASE}(\hat{h})/\mathrm{ASE}(\hat{h}_0)\big)$						
100	1/250	1.13536606	1.07767019	1.08375306	1.19284525	
	1/500	1.10110393	1.05553616	1.05131111	1.14352015	
	1/1000	1.07483753	1.04015876	1.03352956	1.12027987	
300	1/250	1.10016853	1.05502682	1.05728897	1.13366577	
	1/500	1.07841097	1.03955511	1.03539307	1.09281417	
	1/1000	1.06379353	1.02932038	1.02536092	1.06792019	
1000	1/250	1.08059721	1.03725037	1.03386661	1.08785819	
	1/500	1.06599570	1.02781598	1.02325004	1.06119134	
	1/1000	1.05648412	1.02091674	1.02048399	1.04375999	
$\widehat{\mathrm{Median}}\big(\mathrm{ASE}(\hat{h})/\mathrm{ASE}(\hat{h}_0)\big)$						
100	1/250	1.04733054	1.03122740	1.03010994	1.07450458	
	1/500	1.04012974	1.02465166	1.02442740	1.05356328	
	1/1000	1.03192935	1.01889956	1.01429995	1.03739459	
300	1/250	1.03819936	1.02437591	1.02226036	1.04676624	
	1/500	1.03223870	1.01892777	1.01678844	1.03050059	
	1/1000	1.02907223	1.01275110	1.01339069	1.02528878	
1000	1/250	1.03842687	1.01710699	1.01601157	1.02850287	
	1/500	1.03516778	1.01304730	1.01191757	1.02260480	
	1/1000	1.03145817	1.01001448	1.01008131	1.01738506	

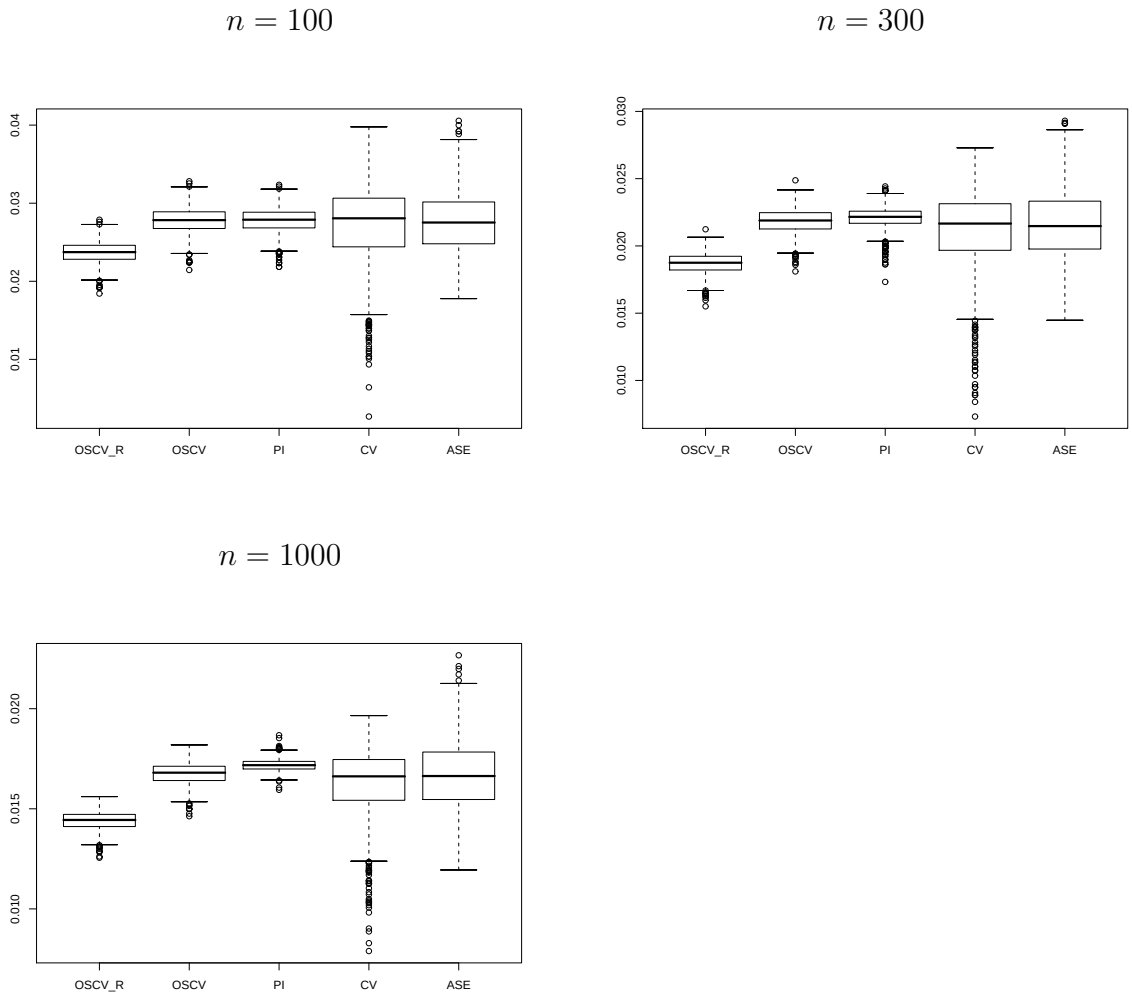


Fig. 30. Boxplots for the data-driven bandwidths in the case of the regression function r_1. The standard deviation of the added noise is $\sigma = 1/500$; the design is fixed, evenly spaced.

Table VIII. Simulation results for r_3. Design: fixed, evenly spaced.

n	σ	**R OSCV**	**OSCV**	**PI**	**CV**	**ASE**
$\hat{\mathrm{E}}(\hat{h})$						
100	1/250	0.02781345	0.03272048	0.03082594	0.02897165	0.02893465
	1/500	0.01890343	0.02219969	0.02308152	0.01987884	0.01971586
	1/1000	0.01312427	0.01539186	0.01709360	0.01358833	0.01364758
300	1/250	0.01990575	0.02340561	0.02387928	0.02076053	0.02140777
	1/500	0.01374898	0.01616434	0.01741220	0.01466136	0.01462972
	1/1000	0.00954253	0.01120754	0.01286447	0.01003337	0.01009401
1000	1/250	0.01425194	0.01676164	0.01800473	0.01515277	0.01545087
	1/500	0.00987256	0.01160366	0.01332268	0.01046123	0.01059549
	1/1000	0.00685782	0.00804998	0.00972627	0.00724627	0.00728106
$\hat{\mathrm{SD}}(\hat{h}) \cdot 10^3$						
100	1/250	2.78782735	3.30952944	3.40943951	6.38708243	4.05945129
	1/500	1.41314142	1.67788534	1.77151546	3.87946228	2.31115754
	1/1000	0.89072320	1.06116933	1.02970161	2.64514953	1.30945747
300	1/250	1.38806227	1.64994695	2.22764492	3.62993131	2.45750160
	1/500	0.76831525	0.91293330	1.11210843	2.04145693	1.46994235
	1/1000	0.44966874	0.53310432	0.49238350	1.25137106	0.82529395
1000	1/250	0.77773412	0.92418724	1.06827851	2.03710686	1.64130473
	1/500	0.41718295	0.49475848	0.51611104	1.15064153	0.87307577
	1/1000	0.24265057	0.28697854	0.21548413	0.65723075	0.51639643
$\hat{\mathrm{E}}\big(\mathrm{ASE}(\hat{h})/\mathrm{ASE}(\hat{h}_0)\big)$						
100	1/250	1.06503553	1.08304548	1.06557020	1.15361768	
	1/500	1.03829659	1.05831928	1.07487172	1.11157242	
	1/1000	1.02727808	1.04910127	1.10838061	1.10861375	
300	1/250	1.04675049	1.04979230	1.05967264	1.09435410	
	1/500	1.03235995	1.04126111	1.07397087	1.05904786	
	1/1000	1.02144663	1.03373491	1.10999261	1.04573292	
1000	1/250	1.03996338	1.03900099	1.06320398	1.06059237	
	1/500	1.02487126	1.02961526	1.09949219	1.03841928	
	1/1000	1.01736240	1.02788043	1.15148588	1.02568193	
$\widehat{\mathrm{Median}}\big(\mathrm{ASE}(\hat{h})/\mathrm{ASE}(\hat{h}_0)\big)$						
100	1/250	1.02840414	1.04389413	1.03352135	1.05743335	
	1/500	1.01679765	1.03399727	1.04636331	1.04083347	
	1/1000	1.01035092	1.02847108	1.08890399	1.02724695	
300	1/250	1.01701620	1.02479656	1.03332768	1.03646547	
	1/500	1.01291490	1.02341149	1.04784823	1.02229067	
	1/1000	1.00934112	1.01867423	1.09372865	1.01878624	
1000	1/250	1.01516733	1.02254971	1.04116087	1.02754846	
	1/500	1.01057203	1.01593085	1.08525696	1.01537137	
	1/1000	1.00769884	1.01600059	1.14179095	1.01160083	

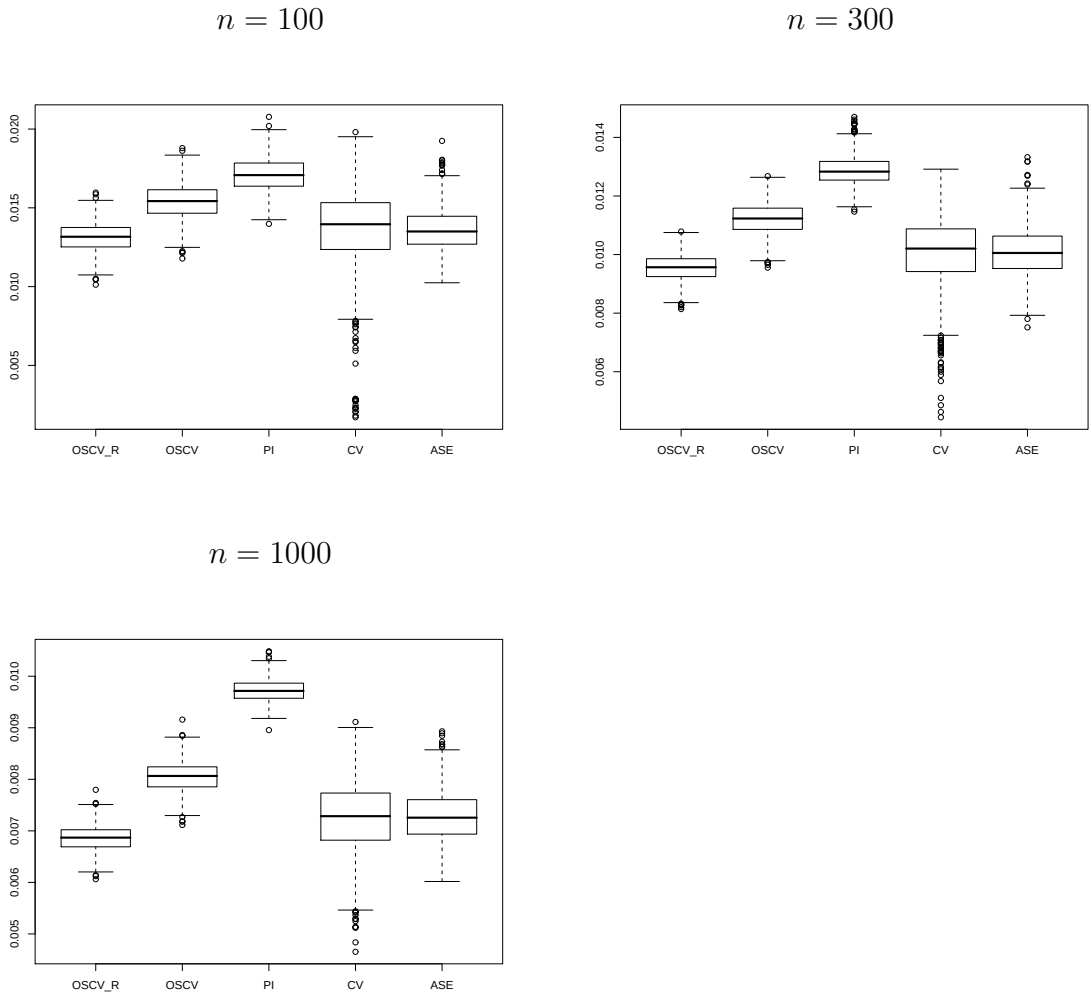


Fig. 31. Boxplots for the data-driven bandwidths in the case of regression function r_3. The standard deviation of the added noise is $\sigma = 1/1000$; the design is fixed, evenly spaced.

Table IX. Measures of performance in the numerical study.

r	R_{E}	R_{SD}	δ for ROSCV	δ for OSCV	R_{ASE}
Fixed evenly spaced design					
r_1	0.8559 (0.0043)	0.8464 (0.0180)	13.8867 (0.6686)	0.6092 (0.4323)	1.0421 (0.0082)
r_2	0.8509 (0.0025)	0.8402 (0.0043)	10.7378 (1.4757)	4.9078 (1.4342)	1.0169 (0.0085)
r_3	0.8464 (0.0108)	0.8423 (0.0017)	5.2189 (1.5324)	11.3067 (1.4370)	0.9896 (0.0075)
Uniform$(0,1)$ design					
r_1	0.8565 (0.0049)	0.8415 (0.0204)	13.3378 (1.0590)	1.1789 (0.9600)	1.0365 (0.0073)
r_2	0.8511 (0.0027)	0.8400 (0.0020)	12.3056 (1.4141)	3.0389 (1.4213)	1.0257 (0.0098)
r_3	0.8605 (0.0195)	0.8299 (0.0220)	5.8213 (2.6533)	10.0563 (2.9391)	0.9881 (0.0098)

two columns of the values of δ, defined by (3.17), for the Robust OSCV and ordinary OSCV methods. Other columns in Table IX correspond to the following ratios:

$$R_{\mathrm{E}} = \widehat{\mathrm{E}}(\hat{h}_{ROSCV})/\widehat{\mathrm{E}}(\hat{h}_{OSCV})$$

$$R_{\mathrm{SD}} = \widehat{\mathrm{SD}}(\hat{h}_{ROSCV})/\widehat{\mathrm{SD}}(\hat{h}_{OSCV})$$

$$R_{\mathrm{ASE}} = \widehat{\mathrm{E}}\big(ASE(\hat{h}_{ROSCV})/ASE(\hat{h}_0)\big)/\widehat{\mathrm{E}}\big(ASE(\hat{h}_{OSCV})/ASE(\hat{h}_0)\big).$$

Our main observations and conclusions from the numerical study are summarized below.

(C1) The average values of R_{E} and R_{SD} reported in the first two columns of Table IX

are very close to the quantity $\dfrac{C^*}{C_{Gaussian}} = \dfrac{0.5217}{0.6168} \doteq 0.8458$. Moreover, the bandwidths $\hat{h}_{ROSCV}$ and $\hat{h}_{OSCV}$ are highly correlated: the sample correlation coefficient was higher than 0.98 in 53 out of 54 considered cases. This indicates that the minimizers of the OSCV curves based on the Gaussian kernel and on the kernel K^* are typically very close, which is a consequence of the fact that the kernels are very close (see Figure 26). This suggests that at least for the fixed evenly spaced and Uniform$(0, 1)$ designs, when the criterion curves based on K^* are smooth, using kernel K^* for cross-validation purposes is like using the Gaussian kernel with the constant $C^* = 0.5217$.

(C2) Notice that $C^* = 0.5217$ is very close to the constant $B = 0.5284$ appropriate for the Gaussian kernel in the nonsmooth case. Thus, for the fixed evenly spaced and Uniform$(0, 1)$ designs, using kernel K^* is practically the same as using a version of OSCV for nonsmooth functions described in Section 2.2.

(C3) In all the considered cases $\widehat{\mathrm{E}}(\hat{h}_{ROSCV}) < \widehat{\mathrm{E}}(\hat{h}_0)$, $\widehat{\mathrm{E}}(\hat{h}_{OSCV}) > \widehat{\mathrm{E}}(\hat{h}_0)$, and $\widehat{\mathrm{SD}}(\hat{h}_{ROSCV}) < \widehat{\mathrm{SD}}(\hat{h}_{OSCV})$. In light of our conclusion **(C1)**, the reduced bandwidth variability for Robust OSCV compared to ordinary OSCV is a consequence of the fact that $C^* = 0.5217 < C_{Gaussian} = 0.6168$.

(C4) From the average values of δ reported in Table IX, it follows that the ordinary OSCV method is practically unbiased for r_1, has a very low bias (average$(\delta) <$ 5%) for r_2 and tends to have a substantial positive bias (average$(\delta) > 10\%$) for r_3. Notice that in the case of ordinary OSCV and a nonsmooth function, the quantity δ, defined by (3.17), estimates

$$\frac{|Cb_n - Bb_n|}{Bb_n} \cdot 100\% = \frac{|C - B|}{B} \cdot 100\% = \frac{|0.6168 - 0.5284|}{0.5284} \cdot 100\% = 16.73\%.$$

This implies that for sufficiently larger sample sizes the bias problem for OSCV applied to a nonsmooth function will get worse than what we observed in our study. Notice that for ordinary OSCV the average δ in the case of r_3 is closer to 16.73% than the average δ in the case of r_2. It may suggest that in finite samples the bias of the ordinary OSCV method depends on the smoothness of the regression function: the more nonsmooth the function, the more severe the bias. Notice that the Robust OSCV is practically unbiased for the case of r_3. This together with our conclusion in **(C1)** implies that it would be reasonable to use a nonsmooth version of OSCV, as described in Section 2.2, only for data sets where the underlying regression function is apparently nonsmooth.

(C5) Ruppert-Sheather-Wand plug-in is overall the best method for a smooth function r_1, but it experiences a substantial problem with the positive bias for r_2 and r_3, which is more severe in the case of the less smooth function r_3.

4. Examples

Our analysis in Section 3 reveals that the kernel K^* is not always useful. Nonetheless, the examples of using K^* are still of interest for at least two reasons. First, the conclusion **(C2)** in Section 3 suggests that performing OSCV based on K^* is practically the same as using a nonsmooth version of OSCV described in Section 2.2. Second, it is instructive to investigate the OSCV criterion curves based on K^* for the real data examples.

In this section we consider one simulated example, involving the design transformation of Hall, Park, and Turlach (1998), and one real data example which compares the performance of OSCV, Robust OSCV, LSCV and Ruppert-Sheather-Wand plug-in. In both examples the Robust OSCV method uses the kernel K^*.

4.1. Simulated example involving design transformation

For this example we generated 1000 data sets of size $n = 1000$ using the regression function

$$r(u) = \begin{cases} \frac{15}{160}u, & 0 \le u \le \frac{4}{15}; \\ -\frac{15}{80}u + \frac{3}{40}, & \frac{4}{15} < u \le \frac{2}{5}; \\ \frac{1}{4}u - \frac{1}{10}, & \frac{2}{5} < u \le \frac{1}{2}; \\ -\frac{1}{20}u + \frac{1}{20} & \frac{1}{2} < x \le 1, \end{cases} \tag{3.18}$$

which has three cusps. The error terms were taken to be $N\left(0, (1/1000)^2\right)$. The design points were generated using the density

$$f(x) = 0.3 \cdot \text{Uniform}(0, 1) + 0.7 \cdot \text{Beta}(3, 5). \tag{3.19}$$

Mixing the beta and uniform distributions ensures that the density satisfies the necessary conditions and solves the problem of missing design points close to 0 and 1. Design is irregular enough for Robust OSCV to yield unacceptably rough criterion curves. However, the Robust OSCV method can still be used after we perform the design transformation proposed by Hall, Park, and Turlach (1998).

For each generated data set we first transform the ordered design sequence x_i, $i = 1, \ldots, n$, to the fixed evenly spaced grid of points $u_i = i/n$. Using data pairs (u_i, Y_i) we select the bandwidth $\hat{h}$ and compute the estimate $\widehat{r_Q}(u)$ of the so called regression function of quantiles $r_Q(u) = r(F^{-1}(u))$. The final regression estimate $\hat{r}_{tr}(x)$ is computed as

$$\hat{r}_{tr}(x) = \widehat{r_Q}\big(\hat{F}(x)\big),$$

where $\hat{F}$ is the empirical distribution function (edf) computed for the design sequence. Notice that $\widehat{r_Q}(u_i) \equiv \hat{r}_{tr}(x_i)$ which implies the identity of the ASE values computed

for the estimates $\widehat{r_Q}(u)$ and $\hat{r}_{tr}(x)$.

In the transformed scale the Robust OSCV method yields smooth criterion curves with a single easily detectable minimum. The average ASE values computed over 1000 replication runs for the Robust OSCV, ordinary OSCV and the Ruppert-Sheather-Wand plug-in methods are given in Table X. Robust OSCV outperforms the other

Table X. Average ASE values corresponding to different bandwidth selection methods.

	Robust OSCV	OSCV	R-S-W plug-in
Average$(ASE) \cdot 10^8$	6.5460	6.7274	7.6068

two methods. This happened because of the positive bias problem for ordinary OSCV and Ruppert-Sheather-Wand plug-in. This suggests that the regression function is nonsmooth to such an extent that a nonsmooth version of OSCV should be preferred to ordinary OSCV.

Figure 32**(a)** shows the Robust OSCV regression estimate which was computed involving the design transformation for the data set with the median value of ASE among 1000 replicated data sets. The corresponding estimate of the regression function of quantiles is plotted in Figure 32**(b)**.

We noticed that using the edf for back transformation in the design transformation method yields a staircase regression estimate which is noticeable for smaller sample sizes and/or more irregular designs compared to what we used in our example. A smooth approximation of the edf is needed to get a smooth regression estimate.

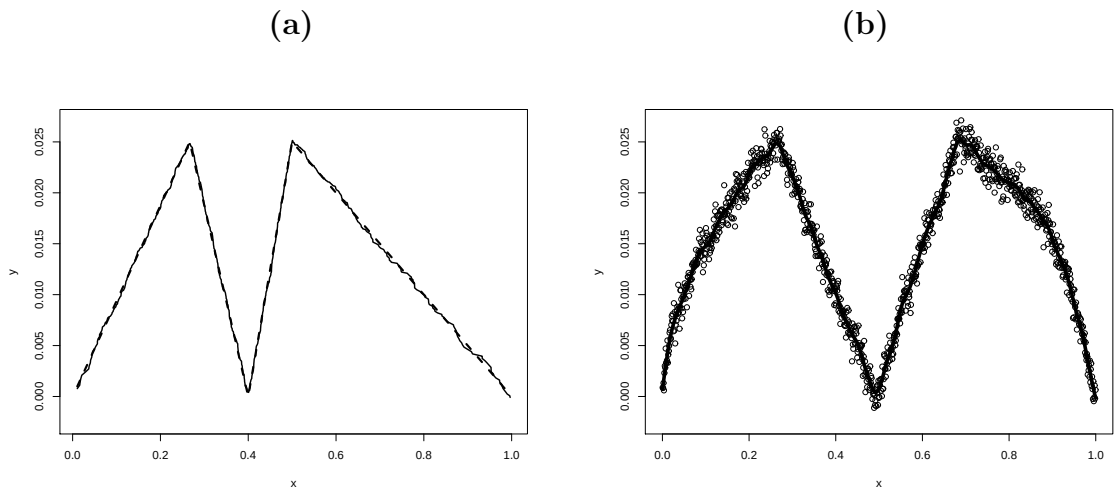


Fig. 32. **(a)** Robust OSCV regression estimate. Dashed line shows the true regression function; **(b)** Robust OSCV estimate of the regression function of quantiles. Circles show the data values in a transformed scale.

4.2. Electricity consumption and temperature in the building data

In this example we analyze the data on measurements of electricity consumption in KWH and mean temperature in degrees F for one building on the University of Minnesota's Twin City campus for $n = 39$ months in 1988-1992. The goal is to model consumption as a function of temperature. The data were taken from the website http://www.stat.umn.edu/alr/data.html associated with the book of Weisberg (2005). The author argues that the high temperature should mean high consumption, since the higher temperature causes the use of air conditioning. Also it is known that the building was steam heated, so electricity was not used for heating. Weisberg (2005) uses a so-called broken-stick model, which has a mean function consisting of two stacked lines, and estimates the location of the band to be around 42°F.

Figure 33 shows the criterion curves for the LSCV, OSCV, and Robust OSCV

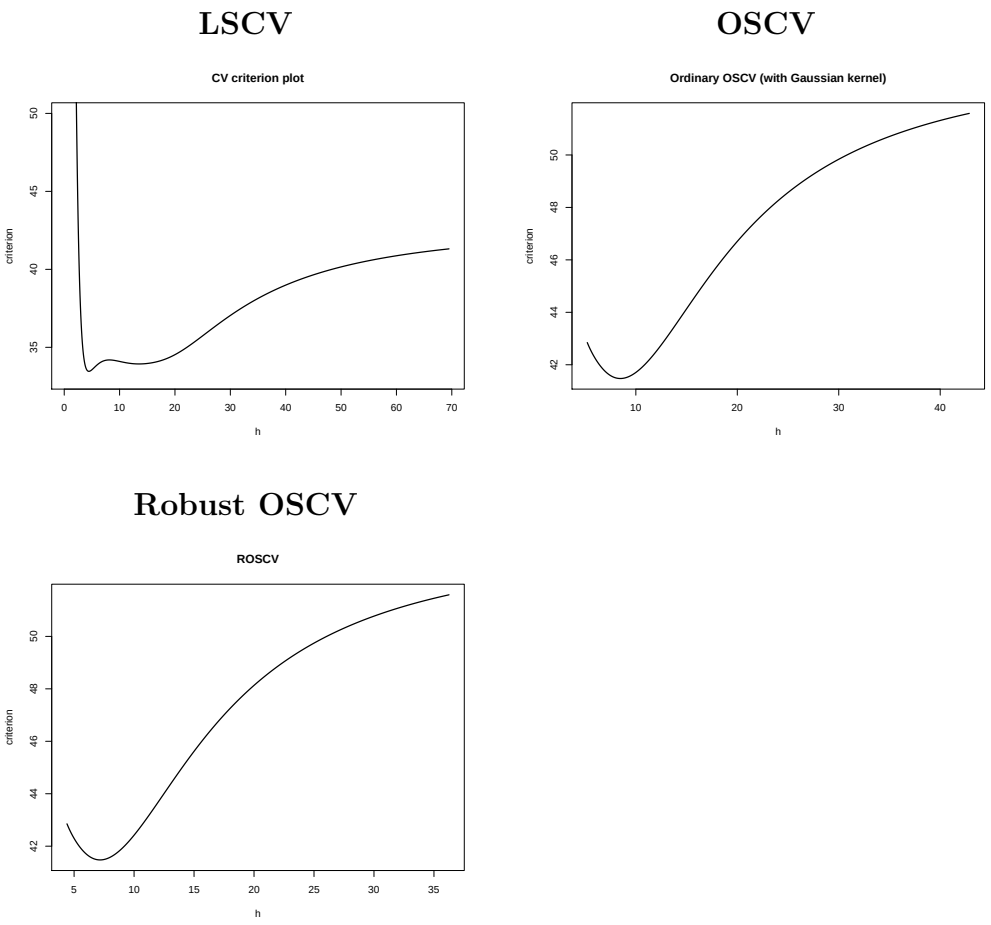


Fig. 33. LSCV, OSCV, and Robust OSCV criterion curves for the electricity and temperature data.

methods. To make the scales of the graphs comparable, we plotted the functions $CV(h)$, $OSCV\left(\frac{h}{C_{Gaussian}}\right)$, and $OSCV^*\left(\frac{h}{C^*}\right)$, where $OSCV^*(\cdot)$ denotes the OSCV function (3.12) based on the kernel K^*. We considered the values of h ranging between the largest spacing and the range of the design data. The Robust OSCV curve is smooth and has a single easily detectable minimum in this interval. Notice that the LSCV curve has two local minima with the largest minimum occurring at $\hat{h} = 13.51$.

The LSCV, OSCV, and Robust OSCV Gaussian local linear estimates are shown

in Figure 34 along with the data points. The OSCV and Robust OSCV estimates are consistent with the Weisberg analysis, whereas the LSCV and the Ruppert-Sheather-Wand plug-in estimates are quite wiggly.

5. Conclusions

The ordinary OSCV method was shown to be fairly robust to lack of smoothness in the regression function. Even though the use of ordinary OSCV for nonsmooth functions produces biased bandwidths, the asymptotic MASE of the K-kernel regression estimator does not increase by more than 1% if K is the Epanechnikov, quartic, or triangle kernel, and it increases by about 4% when K is the Gaussian kernel. This implies that ordinary OSCV does not need much adjustment for the case of discontinuous derivative when K is Epanechnikov, quartic, or triangle. This conclusion is supported by results of our numerous simulation studies. It also explains the good performance of OSCV applied to a function with discontinuous derivative in a numerical study of Hart and Yi (1998), which involved the quartic kernel.

We noted in our simulation study that OSCV based on the Gaussian kernel produces biased upwards bandwidths for nonsmooth functions, with the bias problem getting worse as the smoothness of the regression function decreases. Should the Gaussian kernel be used in OSCV at all? The utility of the Gaussian kernel in OSCV is justified by the following. From the paper of Seifert and Gasser (1996) and from our numerical examples it follows that for the irregular designs containing sparse regions, the OSCV criterion plots based on the Gaussian kernel are smoother than those based on the compactly supported kernels. Below we propose two approaches to reduce the bias of the OSCV method based on the Gaussian kernel for functions with discontinuous derivative.

A first approach consists of using a different constant B in place of C for

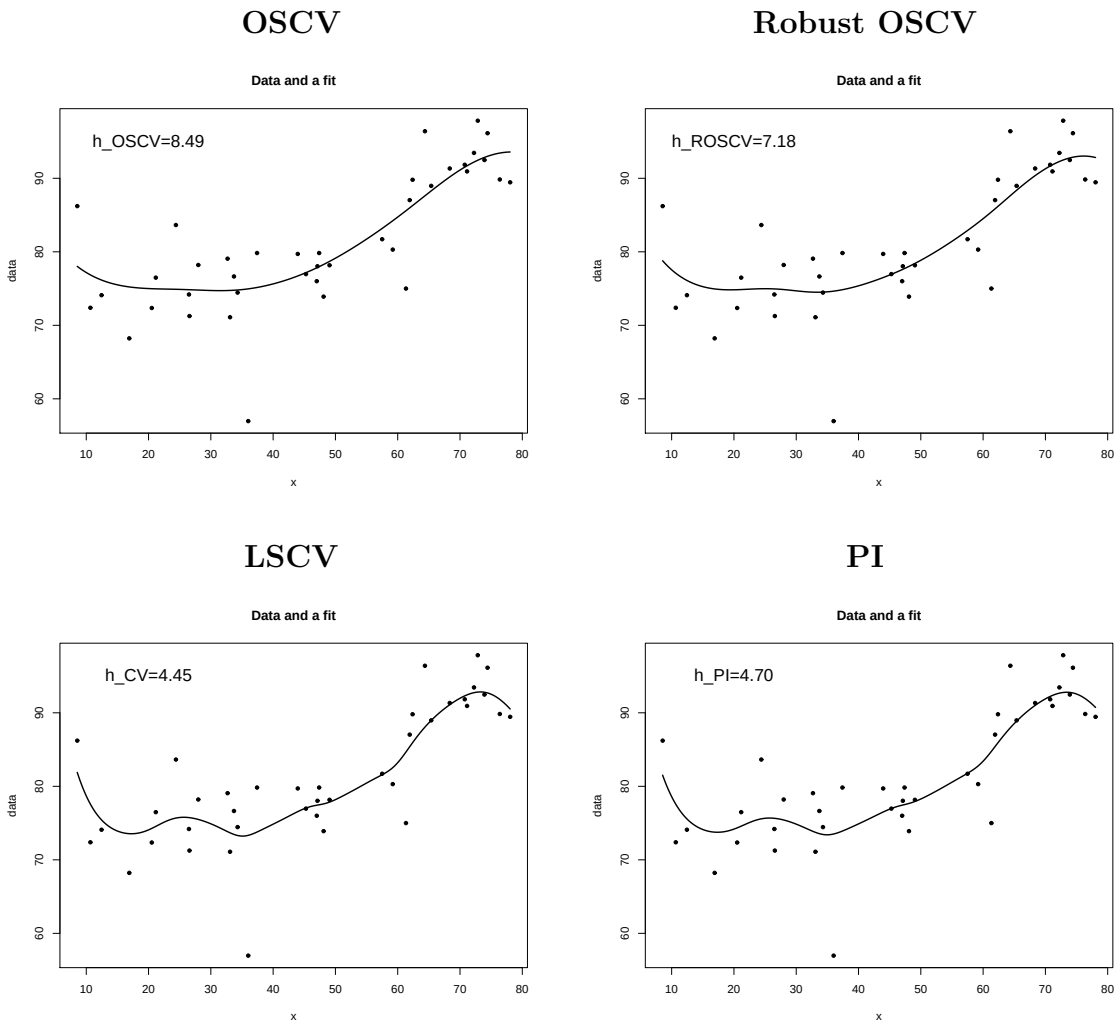


Fig. 34. Regression estimates for the electricity and temperature data.

nonsmooth functions. Our numerical experience suggests that this approach should be followed in practice for functions which have more than one cusp. Another method, called Robust OSCV, consists in equating the method's constants B and C by means of the appropriate cross-validation kernel, which is called robust. A "good" robust kernel should be unimodal and nonnegative. Robust kernels which are not nonnegative may produce very rough criterion curves. Bimodality of the robust kernel should not be allowed since it may lead to criterion curves which have multiple local minima. So far we did not find an entirely suitable robust kernel, so this question remains open.

CHAPTER IV

SUMMARY

A widely help view is that kernel choice is not terribly important when it comes to estimation of the underlying curve. In this dissertation we have shown that the kernel can have a dramatic effect on the properties of cross-validation. In particular, properly chosen kernels can eliminate bandwidth bias, which is the case with robust kernels in the Robust OSCV method. Kernels can also substantially reduce the asymptotic bandwidth variance, which happens with one-sided kernels in the OSCV method. Finally, we found kernels which improve the bandwidth error rate. In particular, in the ICV framework we showed that the asymptotically optimal kernels of the form $(1+\alpha)\phi(u) - \alpha\phi(u/\sigma)/\sigma$, where α and σ are positive constants, produce bandwidths that converge to 0 at a rate of $n^{-1/4}$, which is substantially better than the $n^{-1/10}$ rate of the ordinary LSCV method.

There is a lot of room for further research on the topic of choosing a kernel for cross-validation. In particular, it is entirely possible that there exists another class of kernels that can improve the relative convergence rate of $n^{-1/4}$. Another open problem is to find a robust kernel for the Robust OSCV method which will provide bandwidth variance reduction and smooth criterion curves. One more problem mentioned in the dissertation of Yi (1996) and in the paper of Hart and Yi (1998), which still remains unsolved, is to find an asymptotically optimal OSCV kernel. Finally, it is of interest to generalize the ICV and OSCV methods to multiple dimensions.

REFERENCES

Ahmad, I. A., and Ran, I. S. (2004), "Kernel Contrasts: A Data-based Method of Choosing Smoothing Parameters in Nonparametric Density Estimation," *J. Nonparametr. Stat.,* 16(5), 671–707.

Bowman, A. W. (1984), "An Alternative Method of Cross-validation for the Smoothing of Density Estimates," *Biometrika,* 71(2), 353–360.

Cao, R., Quintela del Rio, A., and Vilar Fernandez, J.M. (1993), "Bandwidth Selection in Nonparametric Density Estimation under Dependence: A Simulation Study," *Computational Statistics,* 8, 313–332.

Chiu, S.-T. (1991a), "Bandwidth Selection for Kernel Density Estimation," *Ann. Statist.,* 19(4), 1883–1905.

Chiu, S.-T. (1991b), "The Effect of Discretization Error on Bandwith Selection for Kernel Density Estimation," *Biometrika,* 78(2), 436–441.

Cleveland, W. S. (1979), "Robust Locally Weighted Regression and Smoothing Scatterplots," *J. Amer. Statist. Assoc.,* 74(368), 829–836.

Cortez, P., and Morais, A. (2007), "A Data Mining Approach to Predict Forest Fires Using Meteorological Data," *in New Trends in Artificial Intelligence, Proceedings of the 13th EPIA 2007 - Portuguese Conference on Artificial Intelligence,* eds. J. Neves, M. F. Santos and J. Machado, December, Guimaraes, Portugal, 512–523.

Desmond, M. (2008), "Lipstick on a Pig," *Forbes*, available at `www.forbes.com/2008/06/06/credit-optimizer-expert-markets-bonds-cx_md_markets46.html`, (retrieved in May 2009).

Fan, J. (1992), "Design-adaptive Nonparametric Regression," *J. Amer. Statist. Assoc.,* 87(420), 998–1004.

Fan, J., Hall, P., Martin, M. A., and Patil, P. (1996), "On Local Smoothing of Nonparametric Curve Estimators," *J. Amer. Statist. Assoc.,* 91(433), 258–266.

Feluch, W., and Koronacki, J. (1992), "A Note on Modified Cross-validation in Density Estimation," *Comput. Statist. Data Anal.,* 13(2), 143–151.

Gasser, T., Kneip, A., and Köhler, W. (1991), "A Flexible and Fast Method for Automatic Smoothing," *J. Amer. Statist. Assoc.,* 86(415), 643–652.

Gasser, T., and Müller, H.-G. (1979), "Kernel Estimation of Regression Functions," In *Smoothing Techniques for Curve Estimation, Proceedings of a Workshop Held in Heidelberg, April 2–4, 1979*, Vol. 757 of *Lecture Notes in Math.*, eds. Th. Gasser and M. Rosenblatt, pp. 23–68. Berlin: Springer.

Hall, P. (1983), "Large Sample Optimality of Least Squares Cross-validation in Density Estimation," *Ann. Statist.,* 11(4), 1156–1174.

Hall, P. (1984), "Central Limit Theorem for Integrated Square Error of Multivariate Nonparametric Density Estimators," *J. Multivariate Anal.,* 14(1), 1–16.

Hall, P., and Johnstone, I. (1992), "Empirical Functional and Efficient Smoothing Parameter Selection," *J. Roy. Statist. Soc. Ser. B,* 54(2), 475–530. With discussion and a reply by the authors.

Hall, P., and Marron, J. S. (1987), "Extent to Which Least-squares Cross-validation Minimises Integrated Square Error in Nonparametric Density Estimation," *Probab. Theory Related Fields,* 74(4), 567–581.

Hall, P., and Marron, J. S. (1991), "Local Minima in Cross-Validation Functions," *J. Roy. Statist. Soc. Ser. B,* 53(1), 245–252.

Hall, P., Park, B. U., and Turlach, B. A. (1998), "A Note on Design Transformation and Binning in Nonparametric Curve Estimation," *Biometrika,* 85(2), 469–476.

Hall, P., and Schucany, W. R. (1989), "A Local Cross-validation Algorithm," *Statist. Probab. Lett.,* 8(2), 109–117.

Härdle, W. (1991), *Smoothing Techniques: With Implementation in S*. New York: Springer.

Härdle, W., Hall P., and Marron, J. S. (1988), "How Far are Automatically Chosen Regression Smoothing Parameters from Their Optimum?" *J. Amer. Statist. Assoc.,* 83(401), 86–101. With comments by David W. Scott and Iain Johnstone and a reply by the authors.

Hart, J. D. (1997), *Nonparametric Smoothing and Lack-of-fit Tests*. Springer Series in Statistics. New York: Springer-Verlag.

Hart, J. D., and Lee, C.-L. (2005), "Robustness of One-sided Cross-validation to Autocorrelation," *J. Multivariate Anal.,* 92(1), 77–96.

Hart, J. D., and Vieu, P. (1990), "Data-driven Bandwidth Choice for Density Estimation Based on Dependent Data," *Ann. Statist.,* 18(2), 873–890.

Hart, J. D., and S. Yi (1998), "One-sided Cross-validation," *J. Amer. Statist. Assoc.,* 93(442), 620–631.

Jones, M. C., Marron, J. S., and Sheather, S. J. (1996a), "A Brief Survey of Bandwidth Selection for Density Estimation," *J. Amer. Statist. Assoc.,* 91(433), 401–407.

Jones, M. C., Marron, J. S., and Sheather, S. J. (1996b), "Progress in Data-Based Bandwidth Selection For Kernel Density Estimation," *Comput. Statist.,* 11(3), 337–381.

Loader, C. (1999a), *Local Regression and Likelihood*. Statistics and Computing. New York: Springer-Verlag.

Loader, C. R. (1999b), "Bandwidth Selection: Classical or Plug-in?" *Ann. Statist.,* 27(2), 415–438.

Marron, J. S., and Wand, M. P. (1992), "Exact Mean Integrated Squared Error," *Ann. Statist.,* 20(2), 712–736.

Marti'nez Miranda, M. D., Nielsen, J. P., and Sperlich, S. (2009), "One Sided Cross Validation for Density Estimation," *Operational Risk Towards Basel III: Best Practices and Issues in Modeling, Management and Regulation, ed. G.N.Gregoriou; John Wiley and Sons, Hoboken, New Jersey*, 177–196.

Mielniczuk, J., Sarda, P., and Vieu, P. (1989), "Local Data-driven Bandwidth Choice for Density Estimation," *J. Statist. Plann. Inference,* 23(1), 53–69.

Nadaraya, E. A. (1964), "On Estimating Regression," *Theory of Probability and its Applications,* 9(1), 141–142.

Park, B. U., and Marron, J. S. (1990), "Comparison of Data-driven Bandwidth Selectors," *Journal of the American Statistical Association, 85*(409), 66–72.

Parzen, E. (1962), "On Estimation of a Probability Density Function and Mode," *Ann. Math. Statist.,* 33, 1065–1076.

Priestley, M. B., and Chao, M. T. (1972), "Non-parametric Function Fitting," *J. Roy. Statist. Soc. Ser. B,* 34, 385–392.

Rosenblatt, M. (1956), "Remarks on Some Nonparametric Estimates of a Density Function," *Ann. Math. Statist.,* 27, 832–837.

Rudemo, M. (1982), "Empirical Choice of Histograms and Kernel Density Estimators," *Scand. J. Statist.,* 9(2), 65–78.

Ruppert, D., Sheather, S. J., and Wand, M. P. (1995), "An Effective Bandwidth Selector for Local Least Squares Regression," *J. Amer. Statist. Assoc.,* 90(432), 1257–1270.

Scott, D. W., and Terrell, G. R. (1987), "Biased and Unbiased Cross-validation in Density Estimation," *J. Amer. Statist. Assoc.,* 82(400), 1131–1146.

Seifert, B., and Gasser, T. (1996), "Finite-sample Variance of Local Polynomials: Analysis and Solutions," *J. Amer. Statist. Assoc.,* 91(433), 267–275.

Sheather, S. J. (2004), "Density Estimation," *Statist. Sci.,* 19(4), 588–597.

Sheather, S. J., and Jones, M. C. (1991), "A Reliable Data-based Bandwidth Selection Method for Kernel Density Estimation," *J. Roy. Statist. Soc. Ser. B,* 53(3), 683–690.

Shmueli, G., Patel, N. R., and Bruce, P. C. (2006), *Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner*. New York: Wiley.

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*. Monographs on Statistics and Applied Probability. London: Chapman & Hall.

Stone, C. J. (1977), "Consistent Nonparametric Regression," *Ann. Statist.,* 5(4), 595–645. With discussion and a reply by the author.

Stute, W. (1992), "Modified Cross-validation in Density Estimation," *J. Statist. Plann. Inference,* 30(3), 293–305.

Terrell, G. R. (1990), "The Maximal Smoothing Principle in Density Estimation," *J. Amer. Statist. Assoc.,* 85(410), 470–477.

van Es, B. (1992), "Asymptotics for Least Squares Cross-validation Bandwidths in Nonsmooth Cases," *Ann. Statist.,* 20(3), 1647–1657.

Wand, M. P., and Jones, M. C. (1995), *Kernel Smoothing*, Vol. 60 of *Monographs on Statistics and Applied Probability*. London: Chapman and Hall Ltd.

Watson, G. S. (1964), "Smooth Regression Analysis," *Sankhyā Ser. A,* 26, 359–372.

Weisberg, S. (2005), *Applied Linear Regression* (Third ed.). Wiley Series in Probability and Statistics. Hoboken, NJ: Wiley-Interscience [John Wiley & Sons].

Yi, S. (1996), *On One-Sided Cross-validation in Nonparametric Regression*. Ph.D. dissertation, Texas A&M University.

Yi, S. (2001), "Asymptotic Stability of the OSCV Smoothing Parameter Selection," *Comm. Statist. Theory Methods,* 30(10), 2033–2044.

Yi, S. (2005), "A Comparison of two Bandwidth Selectors OSCV and AICc in Nonparametric Regression," *Comm. Statist. Simulation Comput.,* 34(3), 585–594.

APPENDIX A

PROOF OF THE THEOREM II.1

Here we outline the proof of our Theorem II.1. A much more detailed proof is available from the authors.

We start by writing

$$\begin{aligned} T_n(b_0) &= T_n(\hat{b}_{UCV}) + (b_0 - \hat{b}_{UCV})T_n^{(1)}(b_0) + \frac{1}{2}(b_0 - \hat{b}_{UCV})^2 T_n^{(2)}(\tilde{b}) \\ &= -nR(L)/2 + (b_0 - \hat{b}_{UCV})T_n^{(1)}(b_0) + \frac{1}{2}(b_0 - \hat{b}_{UCV})^2 T_n^{(2)}(\tilde{b}), \end{aligned}$$

where $\tilde{b}$ is between b_0 and $\hat{b}_{UCV}$, and so

$$(\hat{b}_{UCV} - b_0)\left(1 - (\hat{b}_{UCV} - b_0)\frac{T_n^{(2)}(\tilde{b})}{2T_n^{(1)}(b_0)}\right) = \frac{T_n(b_0) + nR(L)/2}{-T_n^{(1)}(b_0)}.$$

Using condition (2.10) we may write the last equation as

$$(\hat{b}_{UCV} - b_0) = \frac{T_n(b_0) + nR(L)/2}{-T_n^{(1)}(b_0)} + o_p\left(\frac{T_n(b_0) + nR(L)/2}{-T_n^{(1)}(b_0)}\right). \tag{A.1}$$

Defining $s_n^2 = \mathrm{Var}(T_n(b_0))$ and $\beta_n = E(T_n(b_0)) + nR(L)/2$, we have

$$\frac{T_n(b_0) + nR(L)/2}{-T_n^{(1)}(b_0)} = \frac{T_n(b_0) - ET_n(b_0)}{s_n} \cdot \frac{s_n}{-T_n^{(1)}(b_0)} + \frac{\beta_n}{-T_n^{(1)}(b_0)}.$$

Using the central limit theorem of Hall (1984), it can be verified that

$$Z_n \equiv \frac{T_n(b_0) - ET_n(b_0)}{s_n} \xrightarrow{\mathcal{D}} N(0,1).$$

Computation of the first two moments of $T_n^{(1)}(b_0)$ reveals that

$$\frac{-T_n^{(1)}(b_0)}{5R(f'')b_0^4\mu_{2L}^2 n^2/2} \xrightarrow{p} 1,$$

and so

$$\frac{T_n(b_0)+nR(L)/2}{-T_n^{(1)}(b_0)} = Z_n \cdot \frac{2s_n}{5R(f'')b_0^4\mu_{2L}^2 n^2} + \frac{2\beta_n}{5R(f'')b_0^4\mu_{2L}^2 n^2} + o_p\left(\frac{s_n+\beta_n}{b_0^4\mu_{2L}^2 n^2}\right).$$

At this point we need the first two moments of $T_n(b_0)$. A fact that will be used frequently from this point on is that $\mu_{2k,L} = O(\sigma^{2k})$, $k = 1, 2, \ldots$. Using our assumptions on the smoothness of f, Taylor series expansions, symmetry of γ about 0 and $\mu_{2\gamma} = 0$,

$$ET_n(b_0) = -\frac{n^2}{12}b_0^5\mu_{4\gamma}R(f'') + \frac{n^2}{240}b_0^7\mu_{6\gamma}R(f''') + O(n^2b_0^8\sigma^7).$$

Recalling the definition of b_n from (2.5), we have

$$\begin{aligned}\beta_n &= -\frac{n^2}{12}b_0^5\mu_{4\gamma}R(f'') + \frac{n^2}{240}b_0^7\mu_{6\gamma}R(f''') \\ &\quad + \frac{n^2}{2}b_n^5\mu_{2L}^2R(f'') + O(n^2b_0^8\sigma^7). \qquad \text{(A.2)}\end{aligned}$$

Let $MISE_L(b)$ denote the MISE of an L-kernel estimator with bandwidth b. Then $MISE_L'(b_n) = (b_n - b_0)MISE_L''(b_0) + o\left[(b_n - b_0)MISE_L''(b_0)\right]$, implying that

$$b_n^5 = b_0^5 + 5b_0^4\frac{MISE_L'(b_n)}{MISE_L''(b_0)} + o\left[b_0^4\frac{MISE_L'(b_n)}{MISE_L''(b_0)}\right]. \qquad \text{(A.3)}$$

Using a second order approximation to $MISE_L'(b)$ and a first order approximation to $MISE_L''(b)$, we then have

$$b_n^5 = b_0^5 - b_0^7\frac{\mu_{2L}\mu_{4L}R(f''')}{4\mu_{2L}^2R(f'')} + o(b_0^7\sigma^2).$$

Substitution of this expression for b_n into (A.2) and using the facts $\mu_{4\gamma} = 6\mu_{2L}^2$, $\mu_{6\gamma} = 30\mu_{2L}\mu_{4L}$ and $b_0\sigma = o(1)$, it follows that $\beta_n = o(n^2b_0^7\sigma^6)$. Later in the proof we will see that this last result implies that the first order bias of $\hat{h}_{ICV}$ is due only to the difference $Cb_0 - h_0$.

Tedious but straightforward calculations show that $s_n^2 \sim n^2 b_0 R(f) A_\alpha/2$, where A_α is as defined in Section 3.1. It is worth noting that $A_\alpha = R(\rho_\alpha)$, where $\rho_\alpha(u) = u\gamma'_\alpha(u)$ and $\gamma_\alpha(u) = (1+\alpha)^2 \int \phi(u+v)\phi(v)\,dv - 2(1+\alpha)\phi(u)$. One would expect from Theorem 4.1 of Scott and Terrell (1987) that the factor $R(\rho)$ would appear in $\mathrm{Var}(T_n(b_0))$. Indeed it does implicitly, since $R(\rho_\alpha) \sim R(\rho)$ as $\sigma \to \infty$. Our point is that, when $\sigma \to \infty$, the part of L depending on σ is negligible in terms of its effect on $R(\rho)$ and also $R(L)$.

To complete the proof write

$$\begin{aligned}\frac{\hat{h}_{ICV} - h_0}{h_0} &= \frac{\hat{h}_{ICV} - h_0}{h_n} + o_p\left[\frac{\hat{h}_{ICV} - h_0}{h_n}\right] \\ &= \frac{\hat{b}_{UCV} - b_0}{b_n} + \frac{(Cb_0 - h_0)}{h_n} + o_p\left[\frac{\hat{h}_{ICV} - h_0}{h_n}\right].\end{aligned}$$

Applying the same approximation of b_0 that led to (A.3), and the analogous one for h_0, we have

$$\begin{aligned}\frac{Cb_0 - h_0}{h_n} &= b_n^2 \frac{\mu_{2L}\mu_{4L} R(f''')}{20\mu_{2L}^2 R(f'')} - h_n^2 \frac{\mu_{2\phi}\mu_{4\phi} R(f''')}{20\mu_{2\phi}^2 R(f'')} + o(b_n^2\sigma^2 + h_n^2) \\ &= \frac{R(L)^{2/5}\mu_{2L}\mu_{4L} R(f''')}{20(\mu_{2L}^2)^{7/5} R(f'')^{7/5}}\, n^{-2/5} + o(b_n^2\sigma^2).\end{aligned}$$

It is easily verified that, as $\sigma \to \infty$, $R(L) \sim (1+\alpha)^2/(2\sqrt{\pi})$, $\mu_{2L} \sim -\alpha\sigma^2$ and $\mu_{4L} \sim -3\alpha\sigma^4$, and hence

$$\frac{Cb_0 - h_0}{h_n} = \left(\frac{\sigma}{n}\right)^{2/5} \frac{R(f''')}{R(f'')^{7/5}} D_\alpha + o\left[\left(\frac{\sigma}{n}\right)^{2/5}\right].$$

The proof is now complete upon combining all the previous results.

APPENDIX B

MORE SIMULATION RESULTS FOR THE ICV METHOD

Simulation results for the "skewed unimodal," "separated bimodal," and "skewed bimodal" densities, as defined in Section 4.1, are given in Tables XII, XIII, XIV and Figures 35, 36, 37. Table XI shows the percentage of times in 1000 replications that $\hat{h}^*_{ICV} = \hat{h}_{OS}$ for each combination of density and sample size.

Table XI. Percent of times when $\hat{h}^*_{ICV} = \hat{h}_{OS}$ for each combination of f and n.

	Sample size, n			
Density, f	100	250	500	5000
Gaussian	47.8	45.9	46.8	32.2
Skewed Unimodal	21	11.4	5	0
Bimodal	20.4	6.5	1.3	0
Separated Bimodal	0	0	0	0
Skewed Bimodal	32.4	7.4	1	0

Table XII. Simulation results for the Skewed Unimodal density.

n	**LSCV**	**SJPI**	**ICV**	**ISE**
$\hat{\mathrm{E}}(\hat{h})$				
100	0.3101	0.2792	0.3049	0.2996
250	0.2466	0.2353	0.2506	0.2459
500	0.2111	0.2063	0.2180	0.2098
5000	0.1281	0.1299	0.1345	0.1317
$\hat{\mathrm{SD}}(\hat{h}) \cdot 10^2$				
100	8.6639	4.6566	5.6775	5.5223
250	5.8481	2.6449	4.0248	4.0890
500	4.7755	1.8173	2.7911	3.7376
5000	2.2291	0.4058	0.6675	2.0831
$\hat{\mathrm{E}}(\hat{h} - \hat{\mathrm{E}}(\hat{h}_0))^2 \cdot 10^4$				
100	76.0793	25.8405	32.4775	
250	34.1710	8.1077	16.4049	
500	22.8000	3.4228	8.4560	
5000	5.0947	0.1965	0.5218	
$\hat{\mathrm{E}}\big(\mathrm{ISE}(\hat{h})/\mathrm{ISE}(\hat{h}_0)\big)$				
100	2.4546	1.8565	1.7177	
250	1.7447	1.4059	1.4441	
500	1.7162	1.3186	1.3189	
5000	1.2833	1.1135	1.1121	
$\widehat{\mathrm{Median}}\big(\mathrm{ISE}(\hat{h})/\mathrm{ISE}(\hat{h}_0)\big)$				
100	1.3074	1.1456	1.1640	
250	1.1818	1.0941	1.1292	
500	1.1912	1.0880	1.1200	
5000	1.0845	1.0450	1.0472	

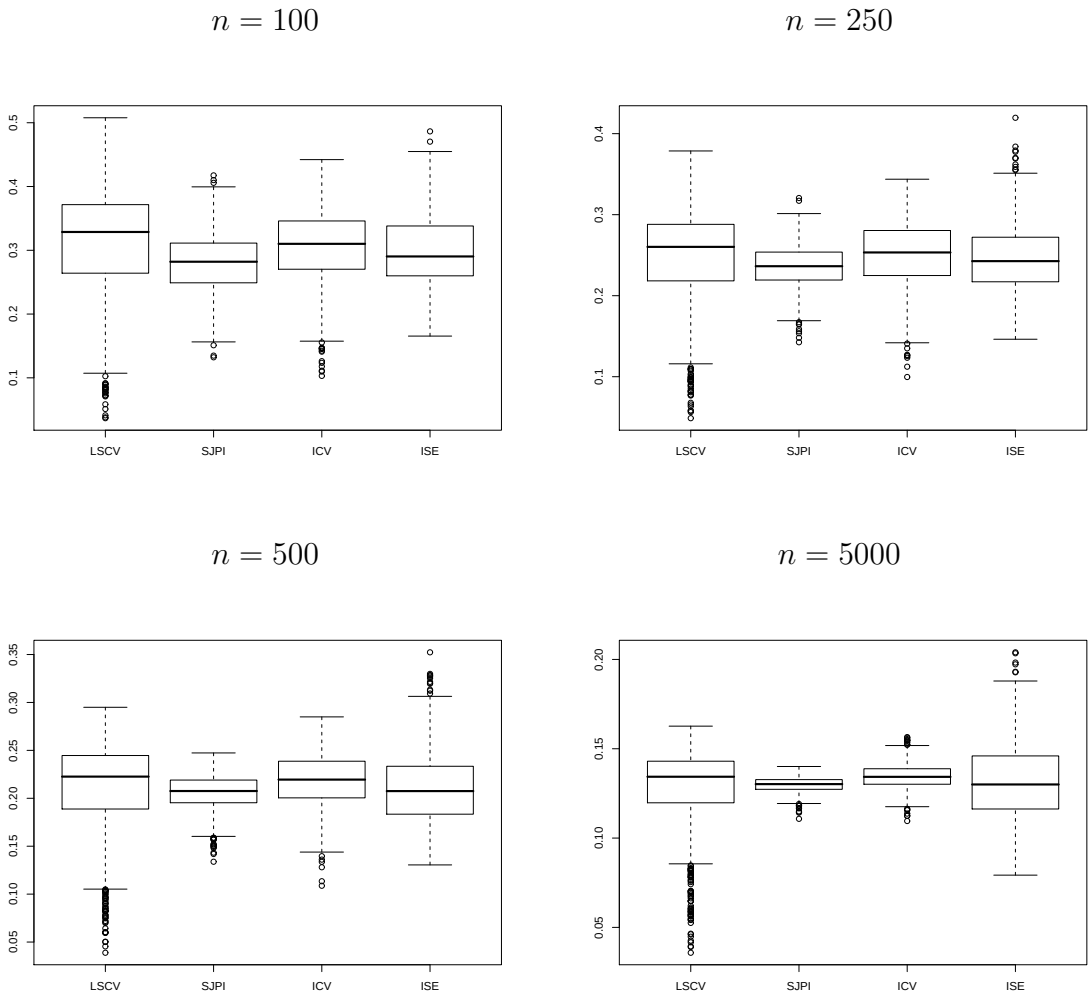


Fig. 35. Boxplots for the data-driven bandwidths in case of the Skewed Unimodal density.

Table XIII. Simulation results for the Separated Bimodal density.

n	**LSCV**	**SJPI**	**ICV**	**ISE**
$\hat{\mathrm{E}}(\hat{h})$				
100	0.2657	0.2796	0.2717	0.2563
250	0.2099	0.2248	0.2178	0.2094
500	0.1794	0.1919	0.1876	0.1798
5000	0.1105	0.1153	0.1152	0.1116
$\hat{\mathrm{SD}}(\hat{h}) \cdot 10^2$				
100	6.0692	2.0866	4.5530	3.4093
250	4.6362	1.2296	2.9603	2.7998
500	3.5225	0.7980	2.0031	2.3275
5000	1.5237	0.1856	0.4645	1.2736
$\hat{\mathrm{E}}(\hat{h} - \hat{\mathrm{E}}(\hat{h}_0))^2 \cdot 10^4$				
100	37.6741	9.7694	23.0845	
250	21.4759	3.8939	9.4640	
500	12.3971	2.1099	4.6171	
5000	2.3313	0.1722	0.3419	
$\hat{\mathrm{E}}\big(\mathrm{ISE}(\hat{h})/\mathrm{ISE}(\hat{h}_0)\big)$				
100	1.4311	1.1250	1.2008	
250	1.3811	1.1047	1.1653	
500	1.3087	1.0833	1.1178	
5000	1.1266	1.0443	1.0510	
$\widehat{\mathrm{Median}}\big(\mathrm{ISE}(\hat{h})/\mathrm{ISE}(\hat{h}_0)\big)$				
100	1.1308	1.0609	1.0752	
250	1.1067	1.0491	1.0687	
500	1.0885	1.0421	1.0521	
5000	1.0432	1.0217	1.0247	

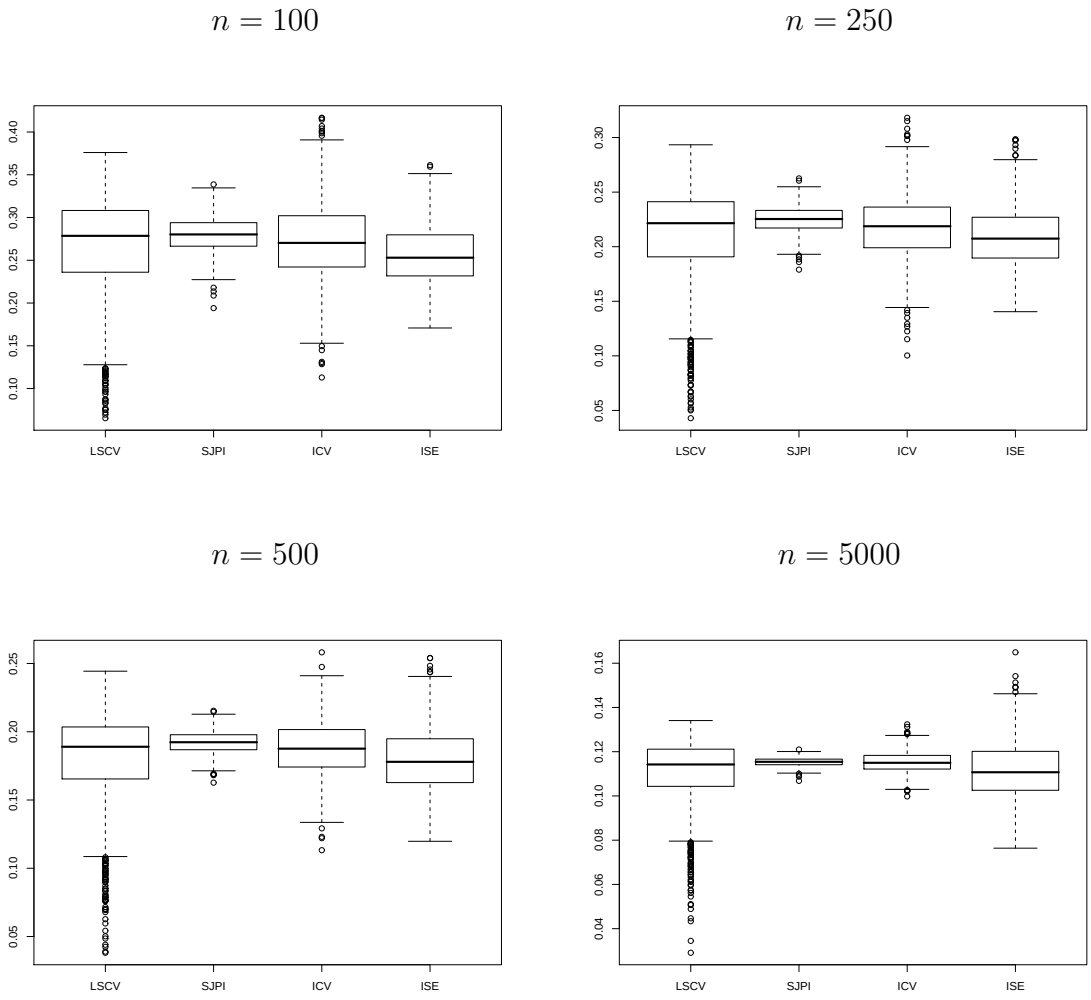


Fig. 36. Boxplots for the data-driven bandwidths in case of the Separated Bimodal density.

Table XIV. Simulation results for the Skewed Bimodal density.

n	**LSCV**	**SJPI**	**ICV**	**ISE**
$\hat{\mathrm{E}}(\hat{h})$				
100	0.3641	0.3530	0.3903	0.3217
250	0.2552	0.2689	0.2814	0.2368
500	0.2046	0.2201	0.2263	0.1990
5000	0.1143	0.1227	0.1259	0.1171
$\hat{\mathrm{SD}}(\hat{h}) \cdot 10^2$				
100	12.8290	7.0324	10.2930	8.2392
250	6.9456	3.8104	6.7463	3.7563
500	4.7935	2.3689	4.2356	2.6883
5000	1.8091	0.4914	0.7461	1.4445
$\hat{\mathrm{E}}(\hat{h} - \hat{\mathrm{E}}(\hat{h}_0))^2 \cdot 10^4$				
100	182.4215	59.2292	152.9251	
250	51.5816	24.8270	65.3387	
500	23.2667	10.0500	25.3556	
5000	3.3476	0.5544	1.3329	
$\hat{\mathrm{E}}\big(\mathrm{ISE}(\hat{h})/\mathrm{ISE}(\hat{h}_0)\big)$				
100	1.5790	1.2198	1.3989	
250	1.3816	1.1550	1.2644	
500	1.2872	1.1179	1.1867	
5000	1.1685	1.0636	1.0745	
$\widehat{\mathrm{Median}}\big(\mathrm{ISE}(\hat{h})/\mathrm{ISE}(\hat{h}_0)\big)$				
100	1.1921	1.1028	1.1852	
250	1.1348	1.0820	1.1363	
500	1.1015	1.0597	1.0894	
5000	1.0514	1.0328	1.0445	

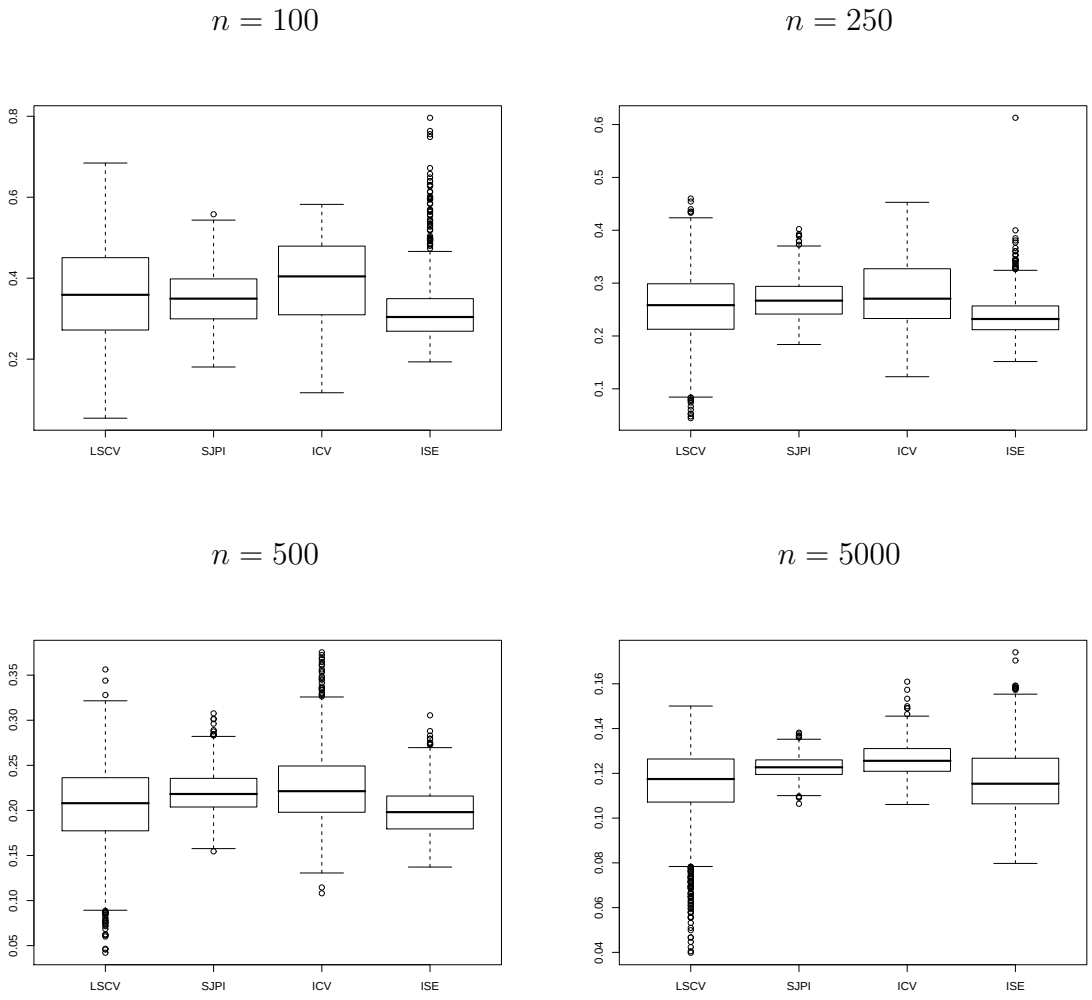


Fig. 37. Boxplots for the data-driven bandwidths in case of the Skewed Bimodal density.

APPENDIX C

ASYMPTOTIC MISE EXPANSION FOR THE LLE IN THE NONSMOOTH CASE

Our goal in this section is to find an asymptotic expansion for the MASE of the LLE (3.1) in the case when the regression function r is nonsmooth. We state our assumptions next.

Assumptions about the regression function:

(R1) r is continuous on $[0, 1]$.

(R2) Second derivative of r exists and is bounded on $[0, 1]$, except at a finite set of points $\{u_t\}$, $t = 1, \ldots, k$, at which $r'(u_t-)$, $r''(u_t-)$ and $r'(u_t+)$ and $r''(u_t+)$ exist with $r'(u_t-) \neq r'(u_t+)$.

Assumptions about kernel K:

(K1) $\int_{-1}^{1} K(u)\,du = 1$;

(K2) $\int_{-1}^{1} uK(u)\,du = 0$;

(K3) $\int u^2 K(u)\,du = \sigma_K^2 \neq 0$;

(K4) K vanishes outside $(-1, 1)$

(K5) K is twice continuously differentiable on $[-1, 1]$.

As $n \to \infty$, we assume that $h \to 0$ and

$$n^2 h^3 \to \infty. \tag{C.1}$$

Now we consider the case of the fixed, evenly spaced design:

$$x_i = \frac{i - \frac{1}{2}}{n}.$$

For notational convenience, we assume that $k = 1$ with the point of discontinuity x_0. We choose h such that $h < \min(x_0, 1 - x_0)$.

We will use the fact that MASE is asymptotically equivalent to the function (3.5), which is equal to MISE (3.6) in the case of the evenly spaced design. We have:

$$MISE_w(h) = MISE(h) = \int_0^1 \text{Bias}^2(\hat{r}_h(x))\, dx + \int_0^1 Var(\hat{r}_h(x))\, dx = ISB(h) + IV(h),$$

where $\text{Bias}(\hat{r}_h(x)) = E(\hat{r}_h(x)) - r(x)$, and ISB and IV stand for the integrated squared bias and integrated variance for $\hat{r}_h$, respectively.

The variance of $\hat{r}_h$ remains the same as in the smooth case. In fact, the local linear estimator has the form $\sum_{i=1}^n W_i Y_i$, where W_i, $i = 1, \ldots, n$, are the fixed weights. Then the variance in either smooth or nonsmooth case is equal to $\sigma^2 \sum_{i=1}^n W_i^2$. However, the bias of $\hat{r}_h$ in the nonsmooth case will change compared to the smooth case. Thus, our goal is to find a new expansion for the ISB term. The ISB derivation in the nonsmooth case relies on the following Lemma.

Lemma IV.1. *For any point $h < x < 1 - h$ the following is true:*

$$E\left(\hat{r}_h(x)\right) = \frac{1}{h} \int_0^1 r(u) K\left(\frac{x - u}{h}\right) du + O\left(\frac{1}{n^2 h^2}\right).$$

The proof of the above lemma relies on the error bound for the middle sum

approximation of an integral. In a simple case when a function $g(x)$ is twice differentiable on the interval $[a, b]$ and the absolute value of its second derivative is bounded by a constant M_2, the error bound for the middle sum approximation of the integral $\int_a^b g(x)\,dx$ when using the value of the function $g(x)$ at n points is given by

$$E = \frac{(b-a)^3}{24n^2} M_2. \tag{C.2}$$

However, some of our functions are not twice differentiable on the intervals where they are defined. The lemma below extends the error bound (C.2) to a more general case we need.

Lemma IV.2. *Suppose the function $g(x)$ satisfies the following conditions:*

(G1) *g is continuous on $[a, b]$.*

(G2) *The second derivative of $g(x)$ exists and is bounded on $[a, b]$, except at a finite set of points $\{x_t\}$, $t = 1, \ldots, k$, at which $g'(x_t-)$, $g''(x_t-)$ and $g'(x_t+)$ and $g''(x_t+)$ exist with $g'(x_t-) \neq g'(x_t+)$.*

Then the error bound for the middle sum approximation of the integral $\int_a^b g(x)\,dx$ is given by

$$E_1 = \frac{(b-a)^3}{24n^2} M_2 + \frac{(b-a)^2}{2n^2} \cdot k \cdot M_1$$

where

$$M_1 = \max_{x \in [a,b]} \{g'(x-), g'(x+)\},$$
$$M_2 = \max_{x \in [a,b]} \{g''(x-), g''(x+)\}.$$

Proof of Lemma IV.1. Find the expectation of $\hat{r}_h$:

$$E(\hat{r}_h(x)) = \frac{t_{n,2} \sum_{i=1}^n r(x_i) K\left(\frac{x-x_i}{h}\right) - t_{n,1} \sum_{i=1}^n r(x_i) K\left(\frac{x-x_i}{h}\right)(x-x_i)}{t_{n,0} t_{n,2} - t_{n,1}^2}, \tag{C.3}$$

where the definition of $t_{n,j}$, $j = 0, 1, 2$, is given by (3.3). Notice that the function $K\left(\frac{x-u}{h}\right)$ is supported on the interval $[x-h,\, x+h]$, which contains $O(2nh)$ points. Denote by i_1 the index of the largest design point such that $x_{i_1} \leq (x-h)$, and let i_2 be the index of the smallest design point, such that $x_{i_2} \geq (x+h)$. In other words

$$i_1 = \lfloor (x-h)n \rfloor / n,$$

$$i_2 = \lceil (x+h)n \rceil / n.$$

Consider the term $t_{n,0}$. Since $K\left(\frac{x-u}{h}\right)$ vanishes outside $[x-h, x+h]$, the following is true:

$$\frac{1}{n} t_{n,0} = \frac{1}{n} \sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) = \frac{1}{n} \sum_{i=i_1}^{i_2} K\left(\frac{x-x_i}{h}\right).$$

Notice that the right part of the above equation is the middle sum approximation to the integral $\displaystyle\int_{x_{i_1}}^{x_{i_2}} K\left(\frac{x-u}{h}\right) du$. The error bound for this approximation can be found using (C.2). In order to compute it, we need to find the second derivative of $K\left(\frac{x-u}{h}\right)$:

$$\frac{\partial^2}{\partial u^2} K\left(\frac{x-u}{h}\right) = \frac{1}{h^2} K''\left(\frac{x-u}{h}\right).$$

From assumption **(K5)** it follows that the maximum of the above function is $O\left(\frac{1}{h^2}\right)$. Therefore the error is

$$E = (x_{i_2} - x_{i_1})^3 \cdot O\left(\frac{1}{24(2nh)^2} \cdot \frac{1}{h^2}\right) = O\left(\frac{(2h)^3}{24(2nh)^2} \frac{1}{h^2}\right) = O\left(\frac{1}{n^2 h}\right).$$

Notice that

$$\int_{x_{i_1}}^{x_{i_2}} K\left(\frac{x-u}{h}\right) du = h \int_{-1}^{1} K(z)\, dz = h,$$

which follows from assumptions **(K1)** and **(K4)**. Combining all steps, we get

$$t_{n,0} = nh + O\left(\frac{1}{nh}\right). \tag{C.4}$$

Next, consider $t_{n,1}$. It follows that

$$\frac{1}{n}t_{n,1} = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)(x-x_i) = \frac{1}{n}\sum_{i=i_1}^{i_2} K\left(\frac{x-x_i}{h}\right)(x-x_i).$$

The righthand side is the middle sum approximation to the integral $\int_{x_{i_1}}^{x_{i_2}} K\left(\frac{x-u}{h}\right)(x-u)\,du$. The second derivative is

$$\frac{\partial^2}{\partial u^2}\left\{K\left(\frac{x-u}{h}\right)(x-u)\right\} = \frac{1}{h^2}K''\left(\frac{x-u}{h}\right)(x-u) + \frac{2}{h}K'\left(\frac{x-u}{h}\right).$$

From assumptions **(K5)** and **(K4)** it follows that the maximum of the above function above is $O\left(\frac{1}{h}\right)$. The error bound is

$$E = (x_{i_2} - x_{i_1})^3 \cdot O\left(\frac{1}{24(2nh)^2}\cdot\frac{1}{h}\right) = O\left(\frac{(2h)^3}{24(2nh)^2}\frac{1}{h}\right) = O\left(\frac{1}{n^2}\right).$$

The integral of interest is

$$\int_{x_{i_1}}^{x_{i_2}} K\left(\frac{x-u}{h}\right)(x-u)\,du = h^2\int_{-1}^{1} uK(u)\,du = 0,$$

which follows from assumptions **(K2)** and **(K4)**. Putting all steps together, we get

$$t_{n,1} = O\left(\frac{1}{n}\right). \tag{C.5}$$

Finally, consider $t_{n,2}$ and observe that

$$\frac{1}{n}t_{n,2} = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)(x-x_i)^2 = \frac{1}{n}\sum_{i=i_1}^{i_2} K\left(\frac{x-x_i}{h}\right)(x-x_i)^2.$$

The righthand side is the middle sum approximation to the integral

$\int_{x_{i_1}}^{x_{i_2}} K\left(\frac{x-u}{h}\right)(x-u)^2\,du$. The second derivative is

$$\frac{\partial^2}{d\partial u^2}\left\{K\left(\frac{x-u}{h}\right)(x-u)^2\right\} = \frac{1}{h^2}K''\left(\frac{x-u}{h}\right)(x-u)^2 + \frac{4}{h}(x-u)K'\left(\frac{x-u}{h}\right) + 2K\left(\frac{x-u}{h}\right),$$

and the maximum of the above function is $O(1)$, which follows from assumptions **(K5)** and **(K4)**. The error bound is

$$E = (x_{i_2} - x_{i_1})^3 \cdot O\left(\frac{1}{24(2nh)^2}\right) = O\left(\frac{(2h)^3}{24(2nh)^2}\right) = O\left(\frac{h}{n^2}\right).$$

We have

$$\int_{x_{i_1}}^{x_{i_2}} K\left(\frac{x-u}{h}\right)(x-u)^2\,du = h^3\int_{-1}^{1} u^2K(u)\,du = h^3\sigma_K^2.$$

Finally, we get

$$t_{n,\,2} = nh^3\sigma_K^2 + O\left(\frac{h}{n}\right). \tag{C.6}$$

The next step is to consider the term $\sum_{i=1}^{n} r(x_i)K\left(\frac{x-x_i}{h}\right)$. Notice that

$$\frac{1}{n}\sum_{i=1}^{n} r(x_i)K\left(\frac{x-x_i}{h}\right) = \frac{1}{n}\sum_{i=i_1}^{i_2} r(x_i)K\left(\frac{x-x_i}{h}\right).$$

The right side of the above equation is the middle sum approximation for the integral $\int_{x_{i_1}}^{x_{i_2}} r(u)K\left(\frac{x-u}{h}\right)du$. Notice that the integrand $r(u)K\left(\frac{x-u}{h}\right)$ is a function described by the conditions **(G1)** and **(G2)** of Lemma IV.2. Let $\frac{\partial}{\partial u_{\pm}}(\cdot)$ denote the

right or left derivative of a function. Then

$$\frac{\partial}{\partial u_{\pm}}\left(r(u)K\left(\frac{x-u}{h}\right)\right) = r'(u_{\pm})K\left(\frac{x-u}{h}\right) - \frac{1}{h}r(u)K'\left(\frac{x-u}{h}\right);$$

$$\frac{\partial^2}{\partial u_{\pm}^2}\left\{r(u)K\left(\frac{x-u}{h}\right)\right\} = r''(u_{\pm})K\left(\frac{x-u}{h}\right) - \frac{2}{h}r'(u_{\pm})K'\left(\frac{x-u}{h}\right) + \frac{1}{h^2}r(u)K''\left(\frac{x-u}{h}\right).$$

From assumptions **(R1)**, **(R2)**, and **(K5)** it follows that $M_1 = O\left(\frac{1}{h}\right)$ and $M_2 = O\left(\frac{1}{h^2}\right)$. The error bound for the integral approximation is

$$E_1 = (x_{i_2} - x_{i_1})^3 \cdot O\left(\frac{1}{24(2nh)^2}\cdot\frac{1}{h^2}\right) + (x_{i_2} - x_{i_1})^2 \cdot O\left(\frac{1}{2(2nh)^2}\cdot\frac{1}{h}\right) = O\left(\frac{(2h)^3}{24(2nh)^2}\cdot\frac{1}{h^2} + \frac{(2h)^2}{2(2nh)^2}\cdot\frac{1}{h}\right) = O\left(\frac{1}{n^2h}\right).$$

Notice that

$$\int_{x_{i_1}}^{x_{i_2}} r(u)K\left(\frac{x-u}{h}\right)\,du = \int_0^1 r(u)K\left(\frac{x-u}{h}\right)\,du.$$

The asymptotic order of the above integral is

$$\int_0^1 r(u)K\left(\frac{x-u}{h}\right)\,du = h\int_{-1}^1 r(x-hz)K(z)\,dz \le hA,$$

where $A = \max_{u\in[0,1]} r(u)$. Hence, $\int_0^1 r(u)K\left(\frac{x-u}{h}\right)\,du = O(h)$. Combining all work, we get the following:

$$\sum_{i=1}^n r(x_i)K\left(\frac{x-x_i}{h}\right) = n\int_0^1 r(u)K\left(\frac{x-u}{h}\right)\,du + O\left(\frac{1}{nh}\right). \tag{C.7}$$

The last term to consider in (C.3) is $\sum_{i=1}^{n} r(x_i)K\left(\frac{x-x_i}{h}\right)(x-x_i)$. Notice that

$$\frac{1}{n}\sum_{i=1}^{n} r(x_i)K\left(\frac{x-x_i}{h}\right)(x-x_i) = \frac{1}{n}\sum_{i=i_1}^{i_2} r(x_i)K\left(\frac{x-x_i}{h}\right)(x-x_i).$$

The righthand term is the middle sum approximation to the integral $\int_{x_{i_1}}^{x_{i_2}} r(u)(x-u)K\left(\frac{x-u}{h}\right)du$. The integrand $r(u)(x-u)K\left(\frac{x-u}{h}\right)$ is a type of function described by conditions **(G1)** and **(G2)** of Lemma IV.2. The derivatives of the integrand are

$$\frac{\partial}{\partial u_\pm}\left(r(u)(x-u)K\left(\frac{x-u}{h}\right)\right) = r'(u_\pm)(x-u)K\left(\frac{x-u}{h}\right) - r(u)K\left(\frac{x-u}{h}\right) - \frac{1}{h}r(u)(x-u)K'\left(\frac{x-u}{h}\right),$$

and

$$\frac{\partial^2}{\partial u_\pm^2}\left\{r(u)(x-u)K\left(\frac{x-u}{h}\right)\right\} = r''(u_\pm)(x-u)K\left(\frac{x-u}{h}\right) - 2r'(u_\pm)K\left(\frac{x-u}{h}\right) - \frac{2}{h}r'(u_\pm)(x-u)K'\left(\frac{x-u}{h}\right) + \frac{2}{h}r(u)K'\left(\frac{x-u}{h}\right) + \frac{1}{h^2}r(u)(x-u)K''\left(\frac{x-u}{h}\right).$$

Using assumptions **(R1)**, **(R2)**, and **(K5)**, we find $M_1 = O(1)$ and $M_2 = O\left(\frac{1}{h}\right)$. The error bound for the integral approximation is

$$E_1 = (x_{i_2}-x_{i_1})^3 \cdot O\left(\frac{1}{24(2nh)^2}\cdot\frac{1}{h}\right) + (x_{i_2}-x_{i_1})^2 \cdot O\left(\frac{1}{2(2nh)^2}\right) = O\left(\frac{(2h)^3}{24(2nh)^2}\cdot\frac{1}{h} + \frac{(2h)^2}{2(2nh)^2}\right) = O\left(\frac{1}{n^2}\right).$$

Consider

$$\int_{x_{i_1}}^{x_{i_2}} r(u)(x-u)K\left(\frac{x-u}{h}\right)du = \int_0^1 r(u)(x-u)K\left(\frac{x-u}{h}\right)du.$$

Find the asymptotic order of the above integral:

$$\int_0^1 r(u)(x-u)K\left(\frac{x-u}{h}\right)\,du = h^2\int_{-1}^1 r(x-hz)zK(z)\,dz = O(h^2),$$

since the integrand is a bounded function, which follows from assumptions **(R1)** and **(K5)**. Finally, we get:

$$\sum_{i=1}^n r(x_i)K\left(\frac{x-x_i}{h}\right)(x-x_i) = O(nh^2) + O\left(\frac{1}{n}\right) = O(nh^2). \tag{C.8}$$

Now we will combine (C.4), (C.5), (C.6), (C.7), and (C.8) to evaluate the expected value (C.3). First, we will consider the numerator and the denominator separately.

$$\text{Numerator} = t_{n,2}\sum_{i=1}^n r(x_i)K\left(\frac{x-x_i}{h}\right) - t_{n,1}\sum_{i=1}^n r(x_i)(x-x_i)K\left(\frac{x-x_i}{h}\right) =$$
$$\left(nh^3\sigma_K^2 + O\left(\frac{h}{n}\right)\right)\left(n\int_0^1 r(u)K\left(\frac{x-u}{h}\right)\,du + O\left(\frac{1}{nh}\right)\right) - O\left(\frac{1}{n}\right)O(nh^2) =$$
$$n^2h^3\sigma_K^2\int_0^1 r(u)K\left(\frac{x-u}{h}\right)\,du + O(h^2) + O\left(\frac{1}{n^2}\right) =$$
$$n^2h^4\left\{\sigma_K^2\cdot\frac{1}{h}\int_0^1 r(u)K\left(\frac{x-u}{h}\right)\,du + O\left(\frac{1}{n^2h^2}\right)\right\}.$$

$$\text{Denominator} = t_{n,0}t_{n,2} - t_{n,1}^2 = \left(nh + O\left(\frac{1}{nh}\right)\right)\left(nh^3\sigma_K^2 + O\left(\frac{h}{n}\right)\right) - O\left(\frac{1}{n^2}\right) =$$
$$n^2h^4\sigma_K^2 + O(h^2) + O\left(\frac{1}{n^2}\right) = n^2h^4\left(\sigma_K^2 + O\left(\frac{1}{n^2h^2}\right)\right).$$

Finally, we get

$$E\left(\hat{r}_h(x)\right) = \frac{n^2h^4\left\{\sigma_K^2 \cdot \frac{1}{h}\int_0^1 r(u)K\left(\frac{x-u}{h}\right)\,du + O\left(\frac{1}{n^2h^2}\right)\right\}}{n^2h^4\left(\sigma_K^2 + O\left(\frac{1}{n^2h^2}\right)\right)} = \frac{1}{h}\int_0^1 r(u)K\left(\frac{x-u}{h}\right)\,du + O\left(\frac{1}{n^2h^2}\right),$$

which finishes the proof of Lemma IV.1. □

Now we will proceed with computing the ISB for the LLE in the nonsmooth case. Notice that

$$ISB(h) = \int_0^1 \left(E\left(\hat{r}_h(x)\right) - r(x)\right)^2\,dx = \int_0^{x_0-h}\left(E\left(\hat{r}_h(x)\right) - r(x)\right)^2\,dx + \int_{x_0-h}^{x_0+h}\left(E\left(\hat{r}_h(x)\right) - r(x)\right)^2\,dx + \int_{x_0+h}^1 \left(E\left(\hat{r}_h(x)\right) - r(x)\right)^2\,dx.$$

From the asymptotic expansion in the smooth case it follows that

$$\int_0^{x_0-h}\left(E\left(\hat{r}_h(x)\right) - r(x)\right)^2\,dx + \int_{x_0+h}^1 \left(E\left(\hat{r}_h(x)\right) - r(x)\right)^2\,dx = O\left(h^4\right) + O\left(\frac{h^2}{n}\right) + O\left(\frac{1}{n^2}\right). \quad \text{(C.9)}$$

Consider

$$\int_{x_0-h}^{x_0+h}\left(E\left(\hat{r}_h(x)\right) - r(x)\right)^2\,dx = h\int_{-1}^1 \left(E\left(\hat{r}_h(x_0-hz)\right) - r(x_0-hz)\right)^2\,dz,$$

with the last step following from the change of variables $z = \dfrac{x_0 - x}{h}$. Consider the bias

$$E\left(\hat{r}_h(x_0-hz)\right) - r(x_0-hz) = \frac{1}{h}\int_0^1 r(u)K\left(\frac{x_0-hz-u}{h}\right)\,du - r(x_0-hz) + O\left(\frac{1}{n^2h^2}\right),$$

which holds by Lemma IV.1. For $h < \min\left(\frac{x_0}{2}, \frac{1-x_0}{2}\right)$, the above bias is

$$\int_{-1}^{1} r(x_0 - h(z+v))K(v)\,dv - r(x_0 - hz) + O\left(\frac{1}{n^2h^2}\right) = \int_{-1}^{1} K(v)\left\{r(x_0 - h(z+v)) - r(x_0 - hz)\right\}\,dv + O\left(\frac{1}{n^2h^2}\right), \quad \text{(C.10)}$$

which follows from the change of variables $v = \frac{x_0 - hz - u}{h}$ and using the assumptions **(K4)** and **(K1)**. Notice that the last integral in (C.10) is $O(h)$, and it is a leading term due to the assumption (C.1).

We get the following:

$$\int_{x_0-h}^{x_0+h} \left(E\left(\hat{r}_h(x)\right) - r(x)\right)^2\,dx = h\int_{-1}^{1}\left\{\int_{-1}^{1} K(v)\left\{r(x_0 - h(z+v)) - r(x_0 - hz)\right\}\,dv\right\}^2 dz + O\left(\frac{1}{n^2}\right) + O\left(\frac{1}{n^4h^3}\right). \quad \text{(C.11)}$$

Combining (C.9) and (C.11), we get the following expression for the integrated squared bias:

$$ISB(h) = h\int_{-1}^{1}\left\{\int_{-1}^{1} K(v)\left\{r(x_0 - h(z+v)) - r(x_0 - hz)\right\}\,dv\right\}^2 dz + O\left(\frac{1}{n^2}\right) + O\left(\frac{1}{n^4h^3}\right) + O\left(\frac{h^2}{n}\right).$$

Taking into account the new constraint (C.1), we get:

$$ISB(h) = h\int_{-1}^{1}\left\{\int_{-1}^{1} K(v)\left\{r(x_0 - h(z+v)) - r(x_0 - hz)\right\}\,dv\right\}^2 dz + O\left(\frac{1}{n^2}\right) + O\left(\frac{h^2}{n}\right). \quad \text{(C.12)}$$

The main term in (C.12) can be written as

$$
\begin{aligned}
h\int_{-1}^{1}\left\{\int_{-1}^{1}K(v)\left\{r(x_0-h(z+v))-r(x_0-hz)\right\}\,dv\right\}^2 dz = \\
h\int_{-1}^{0}\left\{\int_{-1}^{1}K(v)\left\{r(x_0-h(z+v))-r(x_0-hz)\right\}\,dv\right\}^2 dz+ \\
h\int_{0}^{1}\left\{\int_{-1}^{1}K(v)\left\{r(x_0-h(z+v))-r(x_0-hz)\right\}\,dv\right\}^2 dz = \\
h\int_{-1}^{0}\left\{\int_{-1}^{-z}K(v)\left\{r(x_0-h(z+v))-r(x_0-hz)\right\}\,dv+\right. \\
\left.\int_{-z}^{1}K(v)\left\{r(x_0-h(z+v))-r(x_0-hz)\right\}\,dv\right\}^2 dz+ \\
h\int_{0}^{1}\left\{\int_{-1}^{-z}K(v)\left\{r(x_0-h(z+v))-r(x_0-hz)\right\}\,dv+\right. \\
\left.\int_{-z}^{1}K(v)\left\{r(x_0-h(z+v))-r(x_0-hz)\right\}\,dv\right\}^2 dz,
\end{aligned}
\tag{C.13}
$$

Next, we will consider each term in (C.13).

First, consider $\int_{-1}^{-z}K(v)\big\{r(x_0-h(z+v))-r(x_0-hz)\big\}\,dv$ when $<-1<z<0$. In this case $-h(z+v)>0$ and $-hz>0$. From the Taylor's expansion we get the following:

$$r(x_0-h(z+v))=r(x_0)-h(z+v)r'(x_0+)+O(h^2)$$

$$r(x_0-hz)=r(x_0)-hzr'(x_0+)+O(h^2).$$

Then the first integral corresponding to the case $<-1<z<0$ is found as

$$
\begin{aligned}
\int_{-1}^{-z}K(v)\big\{r(x_0-h(z+v))-r(x_0-hz)\big\}\,dv = \\
\int_{-1}^{-z}K(v)\big\{r(x_0)-h(z+v)r'(x_0+)-r(x_0)+hzr'(x_0+)\big\}\,dv+O(h^2) = \\
-hr'(x_0+)\int_{-1}^{-z}vK(v)\,dv+O(h^2)=-hr'(x_0+)G_K(-z)+O(h^2),
\end{aligned}
$$

where we define

$$G_K(z) = \int_{-1}^{z} uK(u)\,du. \tag{C.14}$$

Consider another integral corresponding to the case $-1 < z < 0$:

$$\begin{aligned}\int_{-z}^{1} K(v)\big\{r(x_0 - h(z+v)) - r(x_0 - hz)\big\}\,dv = \\ hz\big(r(x_0+) - r(x_0-)\big)\int_{-z}^{1} K(v) - hr'(x_0-)\int_{-z}^{1} vK(v)\,dv + O(h^2) = \\ hz\big(r(x_0+) - r(x_0-)\big)\big(1 - H_K(-z)\big) + hr'(x_0-)G_K(-z) + O(h^2),\end{aligned}$$

where we define

$$H_K(z) = \int_{-1}^{z} K(u)\,du. \tag{C.15}$$

Next, consider the case $0 < z < 1$. Compute

$$\begin{aligned}\int_{-1}^{-z} K(v)\big\{r(x_0 - h(z+v)) - r(x_0 - hz)\big\}\,dv = \\ - hz\big(r'(x_0+) - r'(x_0-)\big)\int_{-1}^{-z} K(v)\,dv - hr'(x_0+)\int_{-1}^{-z} vK(v)\,dv + O(h^2) = \\ - hz\big(r'(x_0+) - r'(x_0-)\big)H_K(-z) - hr'(x_0+)G_K(-z) + O(h^2).\end{aligned}$$

Finally, consider the second integral corresponding to the case $0 < z < 1$:

$$\begin{aligned}\int_{-z}^{1} K(v)\big\{r(x_0 - h(z+v)) - r(x_0 - hz)\big\}\,dv = \\ - hr'(x_0-)\int_{-z}^{1} vK(v)\,dv + O(h^2) = hr'(x_0-)G_K(-z) + O(h^2).\end{aligned}$$

Now we can find the first integral in (C.13):

$$\int_{-1}^{0}\left\{\int_{-1}^{1}K(v)\left\{r(x_0-h(z+v))-r(x_0-hz)\right\}dv\right\}^2 dz =$$
$$\int_{-1}^{0}\left\{h\big(r'(x_0+)-r'(x_0-)\big)\left(z\big(1-H_K(-z)\big)-G_K(-z)\right)+O(h^2)\right\}^2 dz =$$
$$h^2\big(r'(x_0+)-r'(x_0-)\big)^2\int_{-1}^{0}\left\{z\big(1-H_K(-z)\big)-G_K(-z)\right\}^2 dz+O(h^3)+O(h^4) =$$
$$h^2\big(r'(x_0+)-r'(x_0-)\big)^2\int_{0}^{1}\left\{z\big(1-H_K(z)\big)+G_K(z)\right\}^2 dz+O(h^3).$$

Similarly, the second integral in (C.13) is

$$\int_{0}^{1}\left\{\int_{-1}^{1}K(v)\left\{r(x_0-h(z+v))-r(x_0-hz)\right\}dv\right\}^2 dz =$$
$$h^2\big(r'(x_0+)-r'(x_0-)\big)^2\int_{0}^{1}\left\{zH_K(-z)+G_K(-z)\right\}^2 dz+O(h^3).$$

Putting steps all together, we get the following expression for the ISB:

$$ISB(h)=h^3\big(r'(x_0+)-r'(x_0-)\big)^2B_K+O(h^4)+O\left(\frac{1}{n^2}\right)+O\left(\frac{h^2}{n}\right),$$

where

$$B_K=\int_{0}^{1}\left\{z\big(1-H_K(z)\big)+G_K(z)\right\}^2 dz+\int_{0}^{1}\left\{zH_K(-z)+G_K(-z)\right\}^2 dz. \quad \text{(C.16)}$$

Below we summarize the properties of the functions $H_K(u)$, $G_K(u)$ and the constant B_K:

(P1) $1-H_K(z)=\int_z^1 K(u)\,du$;

(P2) $G_K(z)=-\int_z^1 uK(u)\,du$;

(P3) For a symmetric kernel K

$$H_K(-z) = 1 - H_K(z);$$

$$G_K(-z) = G_K(z);$$

$$B_K = 2\int_0^1 \left\{z\big(1 - H_K(z)\big) + G_K(z)\right\}^2 dz.$$

(P4) When K has support $(0, 1)$,

$$B_K = \int_0^1 \left\{z\big(1 - H_K(z)\big) + G_K(z)\right\}^2 dz.$$

Note that properties **(P1)** and **(P2)** follow from the assumptions **(K1)** and **(K2)** on the kernel K.

Finally, we get the following asymptotic expansion of MISE:

$$MISE(h) = \frac{R(K)\sigma^2}{nh} + h^3\big(r'(x_0+) - r'(x_0-)\big)^2 B_K + o\left(\frac{1}{nh}\right) + o(h^4). \qquad \text{(C.17)}$$

The asymptotic minimizer of (C.17) has the following form:

$$h_n^* = \left(\frac{\sigma^2}{3\big(r'(x_0+) - r'(x_0-)\big)^2}\right)^{1/4} \left(\frac{R(K)}{B_K}\right)^{1/4} n^{-1/4}. \qquad \text{(C.18)}$$

Notice that the order of the MISE minimizer (C.18) is such that the assumption (C.1) is satisfied.

Extensions to other settings.

We can extend the results for the asymptotic MISE expansion (C.17) and its minimizer (C.18) to the following cases.

- By a similar argument it can be shown that the derived results (C.17)

and (C.18) hold for the Gasser-Müller and Priestley-Chao estimators.

- **(k cusps.)** The results (C.17) and (C.18) are given for the case of a single cusp located at the point x_0, but they can be extended to the case of k cusps occurring at the points $\{u_t\}$, $t = 1, \ldots, k$. In the latter case, the quantity $\big(r'(x_0+) - r'(x_0-)\big)^2$ in (C.17) and (C.18) should be replaced with $\sum_{t=1}^{k}\big(r'(u_t+) - r'(u_t-)\big)^2$. This extension follows from the linearity property of an integral and does not require any additional proof.

- **(Heteroscedastic errors.)** To extend (C.17) and (C.18) for the case of heteroscedastic errors, σ^2 should be replaced with $\int_0^1 v(x)\,dx$, where $v(x)$ is the variance function.

- **(Irregular design.)** The results of our work can be extended to the case of an irregular design, when the design points are such that

  (D1) $x_i = Q\left(\frac{i-1/2}{n}\right)$, $i = 1, \ldots, n$, where Q is the inverse of a cdf having density f that satisfies

  (D2) f has support $(0, 1)$,

  (D3) f is Lipschitz continuous on $[0, 1]$, and

  (D4) $f(x) > 0$ for each $x \in [0, 1]$.

  In the case of an irregular design we should expand the weighted MISE function (3.5).

- **Kernels K with infinite support.** Suppose K is a function which satisfies all the conditions imposed on the kernel except for **(K4)**. The results (C.17) and (C.18) can be extended to this case if certain additional constraints are imposed on the tails of K.

- **Continuous kernels with cusps**. The results (C.17) and (C.18) also hold for the case when K is continuous and *piecewise* twice differentiable, which is the case for the Epanechnikov and Laplace kernels.

APPENDIX D

MORE SIMULATION RESULTS FOR THE ROBUST OSCV METHOD

Simulation results for the regression function r_2 in the case of the fixed, evenly spaced design are given in Table XV and Figure 38. Simulation results for the functions r_1, r_2, and r_3 in the case of the Uniform(0,1) design are given in Tables XVI, XVII, XVIII and in Figures 39, 40, and 41.

Table XV. Simulation results for r_2. Design: fixed, evenly spaced.

n	σ	**R OSCV**	**OSCV**	**PI**	**CV**	**ASE**
$\hat{\mathrm{E}}(\hat{h})$						
100	1/250	0.04451301	0.05254226	0.04788073	0.05063080	0.05138712
	1/500	0.03195355	0.03764036	0.03621886	0.03598396	0.03638195
	1/1000	0.02310231	0.02713216	0.02706505	0.02567717	0.02568816
300	1/250	0.03418282	0.04029476	0.03854830	0.03846096	0.03877243
	1/500	0.02457805	0.02889493	0.02852991	0.02723184	0.02750377
	1/1000	0.01761207	0.02064199	0.02171727	0.01916771	0.01948894
1000	1/250	0.02565697	0.03017704	0.02953395	0.02854604	0.02871046
	1/500	0.01835511	0.02152430	0.02241550	0.02021740	0.02029310
	1/1000	0.01310097	0.01532149	0.01740454	0.01425469	0.01435227
$\hat{\mathrm{SD}}(\hat{h}) \cdot 10^3$						
100	1/250	4.17377531	4.96611050	8.21838032	11.96528882	9.86158664
	1/500	2.59102501	3.09897203	5.00691750	7.75322072	5.86682730
	1/1000	1.67807085	2.00591976	2.31176304	4.84963938	3.42402975
300	1/250	2.51738203	3.00467688	5.60409309	7.54206592	6.10303447
	1/500	1.51334467	1.80580804	2.93284324	4.68929460	3.72324244
	1/1000	0.91486089	1.08549757	1.31674733	2.90937100	2.22693317
1000	1/250	1.48168544	1.76741883	3.28673157	4.64034477	3.92229115
	1/500	0.88175399	1.04749095	1.36337438	2.74885972	2.30834537
	1/1000	0.53811967	0.63312130	0.79473548	1.64543362	1.38023335
$\hat{\mathrm{E}}\big(\mathrm{ASE}(\hat{h})/\mathrm{ASE}(\hat{h}_0)\big)$						
100	1/250	1.13919783	1.10365144	1.16973490	1.25638651	
	1/500	1.09782258	1.07219720	1.09824981	1.18350311	
	1/1000	1.06694086	1.05427403	1.05477526	1.13481368	
300	1/250	1.09724629	1.07061871	1.10486133	1.16357049	
	1/500	1.07090100	1.05365733	1.06557090	1.11926407	
	1/1000	1.04970132	1.03965545	1.05377369	1.08207093	
1000	1/250	1.06518737	1.04997135	1.06509046	1.10877299	
	1/500	1.04753036	1.03696557	1.04963212	1.07443487	
	1/1000	1.03659350	1.02985680	1.08758391	1.05045221	
$\widehat{\mathrm{Median}}\big(\mathrm{ASE}(\hat{h})/\mathrm{ASE}(\hat{h}_0)\big)$						
100	1/250	1.04421790	1.04617793	1.06084278	1.08014415	
	1/500	1.03418149	1.03219198	1.04466608	1.06619746	
	1/1000	1.02404147	1.02670809	1.02408392	1.04856178	
300	1/250	1.03615976	1.03371349	1.05116435	1.05986069	
	1/500	1.02867076	1.02536189	1.03160595	1.04406362	
	1/1000	1.01914650	1.02032292	1.03029875	1.02922946	
1000	1/250	1.02153413	1.02735898	1.03031578	1.04128633	
	1/500	1.01729040	1.02047193	1.03206717	1.02957047	
	1/1000	1.01327938	1.01549158	1.07345813	1.01892252	

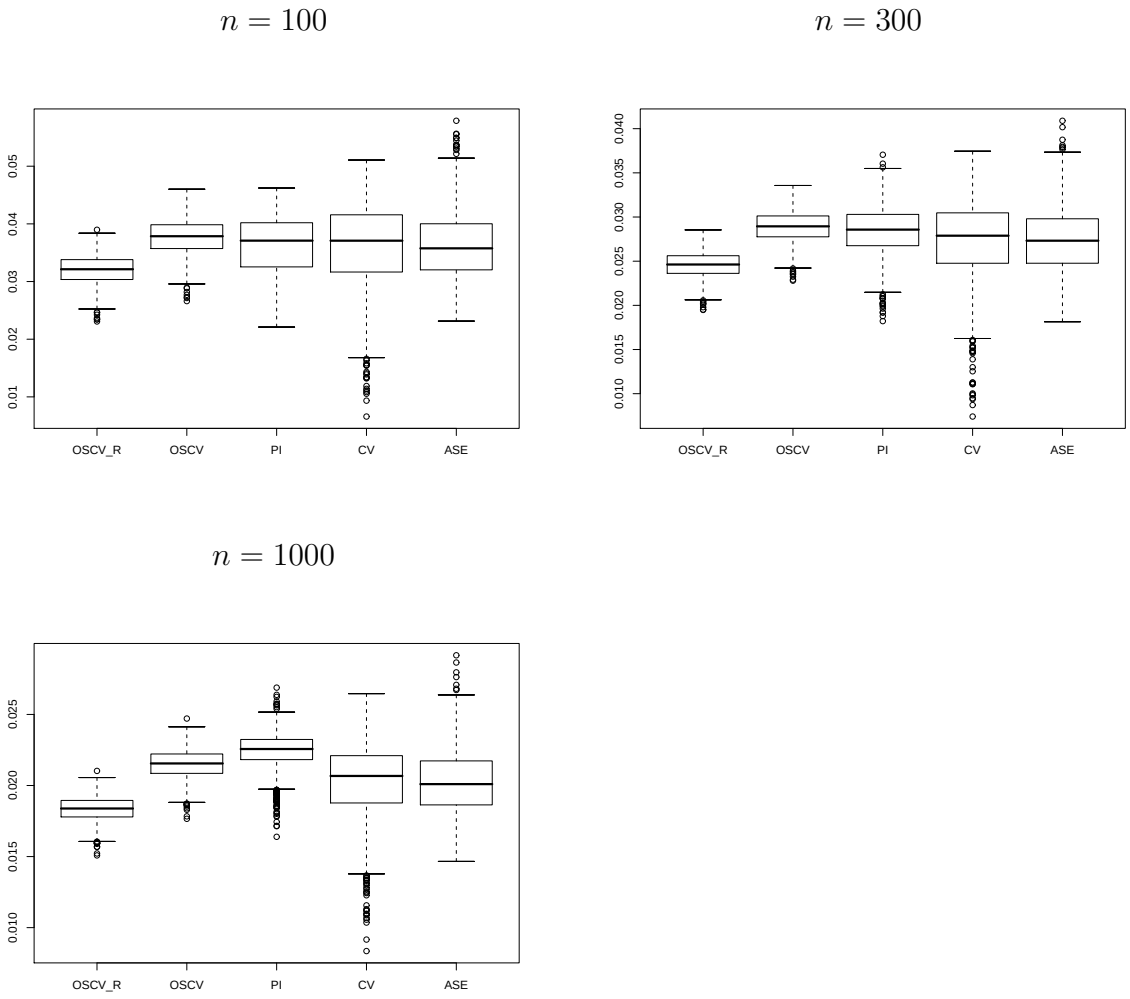


Fig. 38. Boxplots for the data-driven bandwidths in the case of regression function r_2. The standard deviation of the added noise is $\sigma = 1/500$; the design is fixed, evenly spaced.

Table XVI. Simulation results for r_1. Design: Uniform$(0, 1)$.

n	σ	**R OSCV**	**OSCV**	**PI**	**CV**	**ASE**
$\hat{\mathrm{E}}(\hat{h})$						
100	1/250	0.03232221	0.03806921	0.03545243	0.03743289	0.03792869
	1/500	0.02490912	0.02922465	0.02828651	0.02842373	0.02864920
	1/1000	0.01913825	0.02218616	0.02124425	0.02145490	0.02155682
300	1/250	0.02515952	0.02956033	0.02942132	0.02865374	0.02904550
	1/500	0.01881788	0.02197205	0.02175634	0.02155956	0.02176808
	1/1000	0.01425331	0.01652614	0.01631989	0.01620413	0.01625569
1000	1/250	0.01945687	0.02274063	0.02295855	0.02231343	0.02266633
	1/500	0.01425573	0.01660195	0.01696484	0.01622938	0.01659012
	1/1000	0.01056406	0.01227973	0.01274135	0.01199043	0.01221388
$\hat{\mathrm{SD}}(\hat{h}) \cdot 10^3$						
100	1/250	3.24267172	3.89309730	3.85335072	6.74814328	5.84837881
	1/500	1.77457491	2.15113418	2.08928267	4.60298419	3.56255079
	1/1000	1.57119458	1.91538584	0.99034115	3.25746964	2.50837823
300	1/250	1.24116731	1.48894501	2.07674822	4.23284836	3.85143218
	1/500	0.73860867	0.89238240	0.72208219	2.87095794	2.64524401
	1/1000	0.54205733	0.64240864	0.39625917	1.80064350	1.70619790
1000	1/250	0.69221563	0.82391663	0.83573111	2.72908882	2.75237702
	1/500	0.50027594	0.57498926	0.31859809	1.76589744	1.73124358
	1/1000	0.31614607	0.35936873	0.16182203	1.09410883	1.09240274
$\hat{\mathrm{E}}\big(\mathrm{ASE}(\hat{h})/\mathrm{ASE}(\hat{h}_0)\big)$						
100	1/250	1.12891644	1.08293352	1.09723519	1.17180538	
	1/500	1.10040517	1.05234968	1.05587649	1.12481472	
	1/1000	1.06676509	1.04526470	1.03571893	1.10001978	
300	1/250	1.08849336	1.05145015	1.05655636	1.11414223	
	1/500	1.08143529	1.04118719	1.03651587	1.09295312	
	1/1000	1.06492368	1.03190902	1.02610560	1.06636760	
1000	1/250	1.08261639	1.03979774	1.03600888	1.08261912	
	1/500	1.06849752	1.02919136	1.02341043	1.06224536	
	1/1000	1.05659024	1.02109855	1.01937600	1.04392716	
$\widehat{\mathrm{Median}}\big(\mathrm{ASE}(\hat{h})/\mathrm{ASE}(\hat{h}_0)\big)$						
100	1/250	1.04903265	1.03658481	1.03665227	1.06256481	
	1/500	1.03845981	1.02260015	1.02123268	1.05049224	
	1/1000	1.02370471	1.02119828	1.01431556	1.04004291	
300	1/250	1.03643733	1.02471755	1.02527598	1.04018390	
	1/500	1.03501438	1.01942064	1.01600506	1.03752091	
	1/1000	1.03230476	1.01481048	1.01280845	1.02665804	
1000	1/250	1.03461158	1.01688693	1.01701970	1.02709110	
	1/500	1.03365213	1.01295630	1.01091911	1.02343976	
	1/1000	1.02943637	1.00956386	1.00958079	1.01643634	

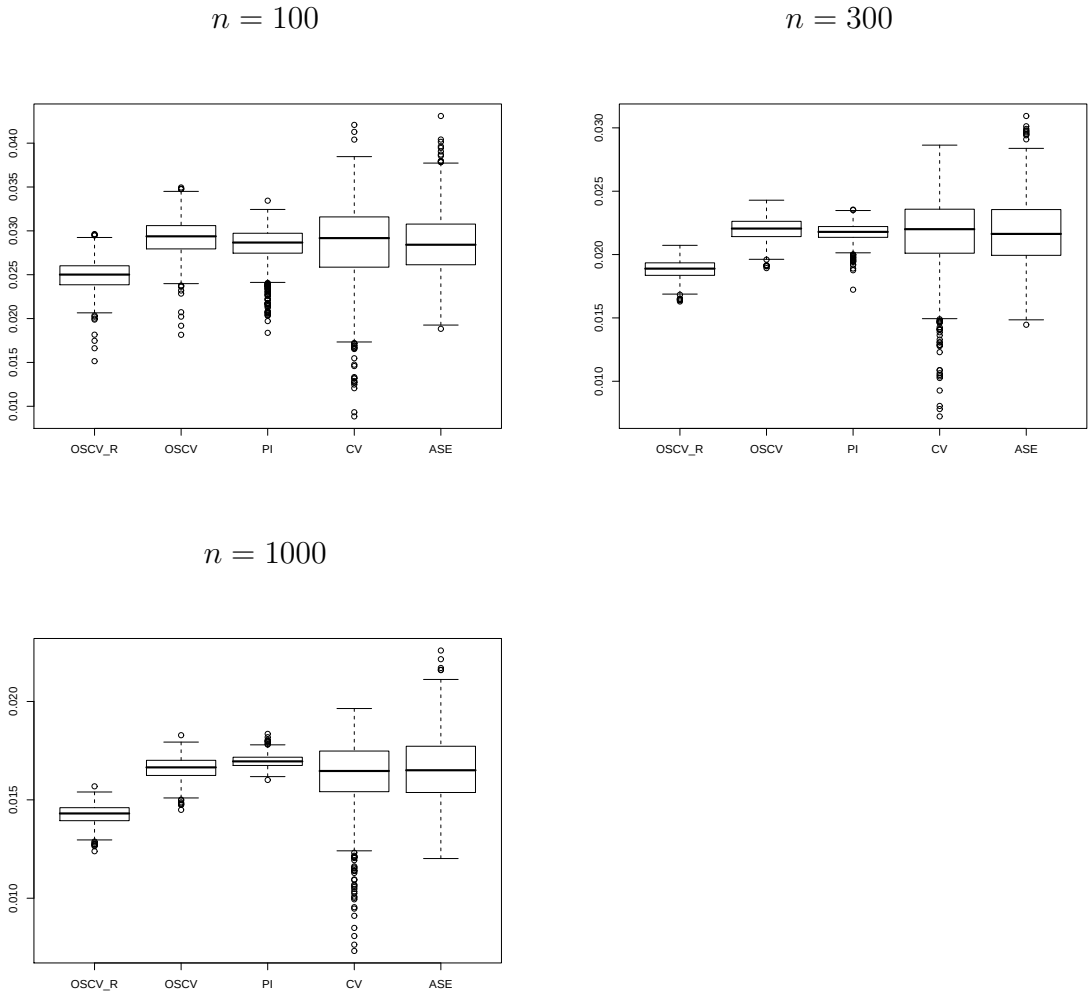


Fig. 39. Boxplots for the bandwidths in the case of regression function r_1. The standard deviation of the added noise is $\sigma = 1/500$; the design is Uniform$(0, 1)$.

Table XVII. Simulation results for r_2. Design: Uniform$(0, 1)$.

n	σ	**R OSCV**	**OSCV**	**PI**	**CV**	**ASE**
$\hat{\mathrm{E}}(\hat{h})$						
100	1/250	0.04528986	0.05348694	0.04941739	0.05234910	0.05293601
	1/500	0.03297073	0.03883176	0.03564665	0.03807817	0.03842472
	1/1000	0.02401140	0.02817743	0.02498657	0.02700382	0.02683877
300	1/250	0.03436888	0.04051179	0.03915591	0.03851786	0.03905650
	1/500	0.02440595	0.02866698	0.02713008	0.02775206	0.02801996
	1/1000	0.01767143	0.02070383	0.02042552	0.01998088	0.01985346
1000	1/250	0.02567259	0.03019127	0.02917635	0.02902262	0.02951896
	1/500	0.01832578	0.02148972	0.02150440	0.02033297	0.02073369
	1/1000	0.01309426	0.01531247	0.01610439	0.01443667	0.01472242
$\hat{\mathrm{SD}}(\hat{h}) \cdot 10^3$						
100	1/250	4.75485756	5.65157153	8.87231491	11.35875404	9.67949808
	1/500	2.80026425	3.34091320	6.08978153	7.58816792	6.27390415
	1/1000	2.03259526	2.42330504	3.24644235	4.99546198	3.77239645
300	1/250	2.61022619	3.11156080	5.85828959	7.29501318	6.26113564
	1/500	1.49753413	1.78474525	2.74912085	4.65548337	3.62015264
	1/1000	0.87795393	1.04562822	1.05238406	3.23608148	2.56629731
1000	1/250	1.49508186	1.78410444	3.18019279	4.75837906	4.12465480
	1/500	0.89774062	1.06608870	1.34997302	2.83552671	2.43844642
	1/1000	0.53631847	0.63544076	0.85113877	1.81480355	1.51756062
$\hat{\mathrm{E}}\big(\mathrm{ASE}(\hat{h})/\mathrm{ASE}(\hat{h}_0)\big)$						
100	1/250	1.15068885	1.10391892	1.18257420	1.20478084	
	1/500	1.10266342	1.06531141	1.12698597	1.16235228	
	1/1000	1.06455803	1.05229360	1.07236641	1.11109367	
300	1/250	1.10792801	1.08001730	1.12595063	1.16400796	
	1/500	1.07860594	1.04679639	1.06169803	1.11187922	
	1/1000	1.05340408	1.03562955	1.03423719	1.08646924	
1000	1/250	1.08354116	1.05229939	1.06672520	1.11570054	
	1/500	1.06048549	1.03783929	1.03902651	1.08125756	
	1/1000	1.04455745	1.02708377	1.03769955	1.05746756	
$\widehat{\mathrm{Median}}\big(\mathrm{ASE}(\hat{h})/\mathrm{ASE}(\hat{h}_0)\big)$						
100	1/250	1.04931640	1.05098907	1.06239284	1.08623321	
	1/500	1.04008791	1.02715976	1.04409101	1.05494761	
	1/1000	1.02525002	1.02592786	1.03127843	1.04567497	
300	1/250	1.03690213	1.03752322	1.05514394	1.05694722	
	1/500	1.03272763	1.02561648	1.02979449	1.04409069	
	1/1000	1.02334481	1.01763904	1.01655295	1.03123326	
1000	1/250	1.03180304	1.02448325	1.02972084	1.04469758	
	1/500	1.02550633	1.01927112	1.01940507	1.02975975	
	1/1000	1.01930478	1.01486523	1.02161334	1.02191551	

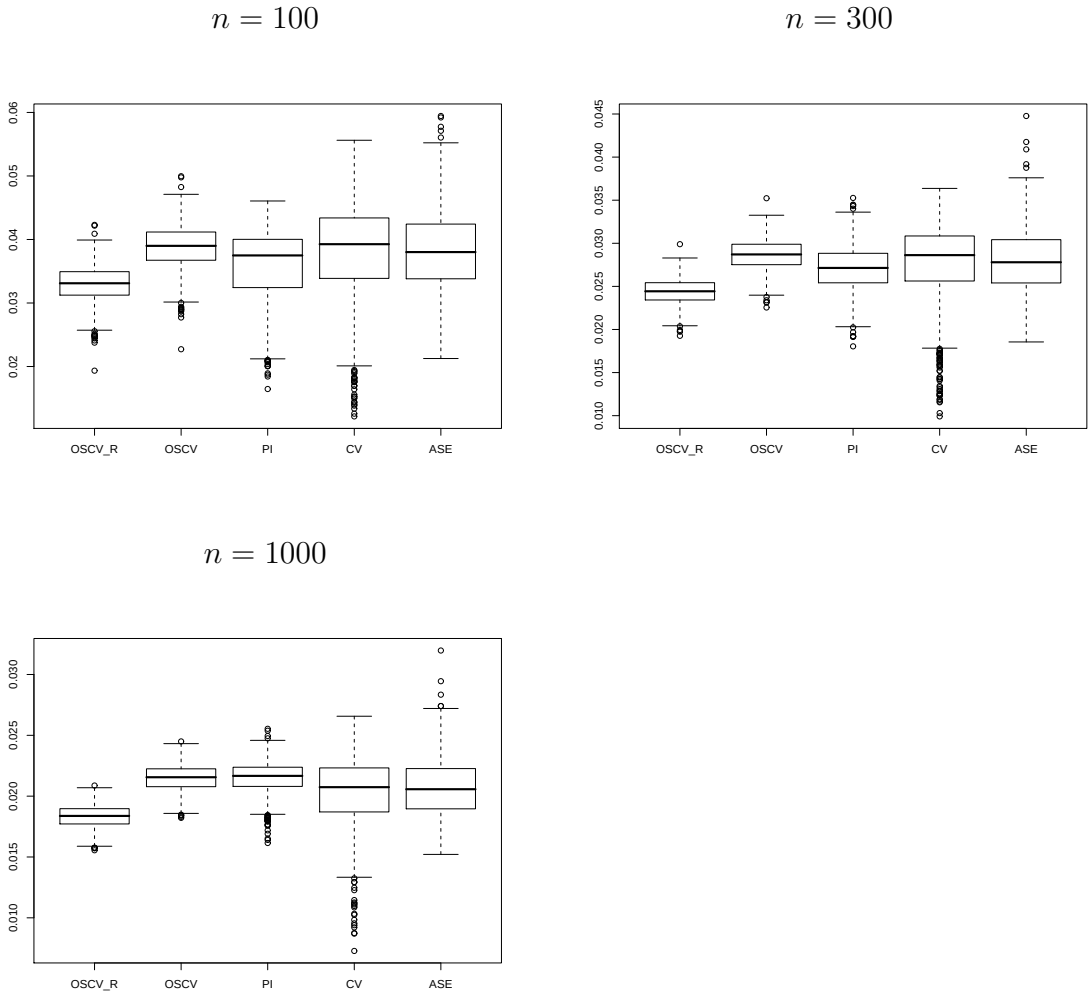


Fig. 40. Boxplots for the bandwidths in the case of regression function r_2. The standard deviation of the added noise is $\sigma = 1/500$; the design is Uniform$(0, 1)$.

Table XVIII. Simulation results for r_3. Design: Uniform$(0, 1)$.

n	σ	**R OSCV**	**OSCV**	**PI**	**CV**	**ASE**
$\hat{\mathrm{E}}(\hat{h})$						
100	1/250	0.02764573	0.03243142	0.02982051	0.02926722	0.02898387
	1/500	0.01958961	0.02264601	0.02101072	0.02316956	0.02128753
	1/1000	0.01498380	0.01651557	0.014973887	0.01726621	0.01542466
300	1/250	0.01959382	0.02301398	0.02282689	0.02111411	0.02116256
	1/500	0.01401467	0.01645239	0.01718795	0.01435196	0.01447137
1000	1/250	0.01428014	0.01678820	0.01773797	0.01560092	0.01558614
	1/500	0.01016672	0.01193513	0.01299631	0.01060943	0.01071737
	1/1000	0.00698158	0.00818222	0.00948285	0.00713901	0.00720262
$\hat{\mathrm{SD}}(\hat{h}) \cdot 10^3$						
100	1/250	3.44346365	4.12972396	4.05586827	5.76602406	3.75966637
	1/500	2.08744573	2.61453099	3.09840220	4.37903938	2.31610740
	1/1000	1.58595198	2.00319852	1.51616408	3.27097202	1.60810832
300	1/250	1.33502298	1.58817027	2.12753048	3.37297535	2.54993038
	1/500	0.87176679	1.03745076	1.06468140	1.95469604	1.42009273
1000	1/250	0.72956718	0.86639625	1.06879559	1.94715815	1.51174985
	1/500	0.45193187	0.53663231	0.41109787	1.25010387	0.94623753
	1/1000	0.28115319	0.33089202	0.17871324	0.66383006	0.49246551
$\hat{\mathrm{E}}\big(\mathrm{ASE}(\hat{h})/\mathrm{ASE}(\hat{h}_0)\big)$						
100	1/250	1.07914593	1.09861497	1.08495306	1.13358548	
	1/500	1.04996210	1.05507657	1.06971981	1.12052416	
	1/1000	1.03384982	1.05176590	1.03869050	1.12137235	
300	1/250	1.04871542	1.05080314	1.05539121	1.08565309	
	1/500	1.03102802	1.05570791	1.07514590	1.06087083	
1000	1/250	1.03865684	1.03521224	1.05315654	1.05447328	
	1/500	1.02171882	1.03373305	1.06709853	1.03857692	
	1/1000	1.01445837	1.03725795	1.13168982	1.02571757	
$\widehat{\mathrm{Median}}\big(\mathrm{ASE}(\hat{h})/\mathrm{ASE}(\hat{h}_0)\big)$						
100	1/250	1.03198808	1.05250521	1.04124859	1.06704661	
	1/500	1.02102021	1.02669730	1.03275517	1.05712287	
	1/1000	1.01522171	1.02800707	1.01775410	1.05875204	
300	1/250	1.01940512	1.02504385	1.02814879	1.03691132	
	1/500	1.01212335	1.02989500	1.05011906	1.02476130	
1000	1/250	1.01596532	1.01709755	1.03103668	1.02334128	
	1/500	1.00928082	1.01831604	1.05441287	1.01679691	
	1/1000	1.00567991	1.02609596	1.12520355	1.01143505	

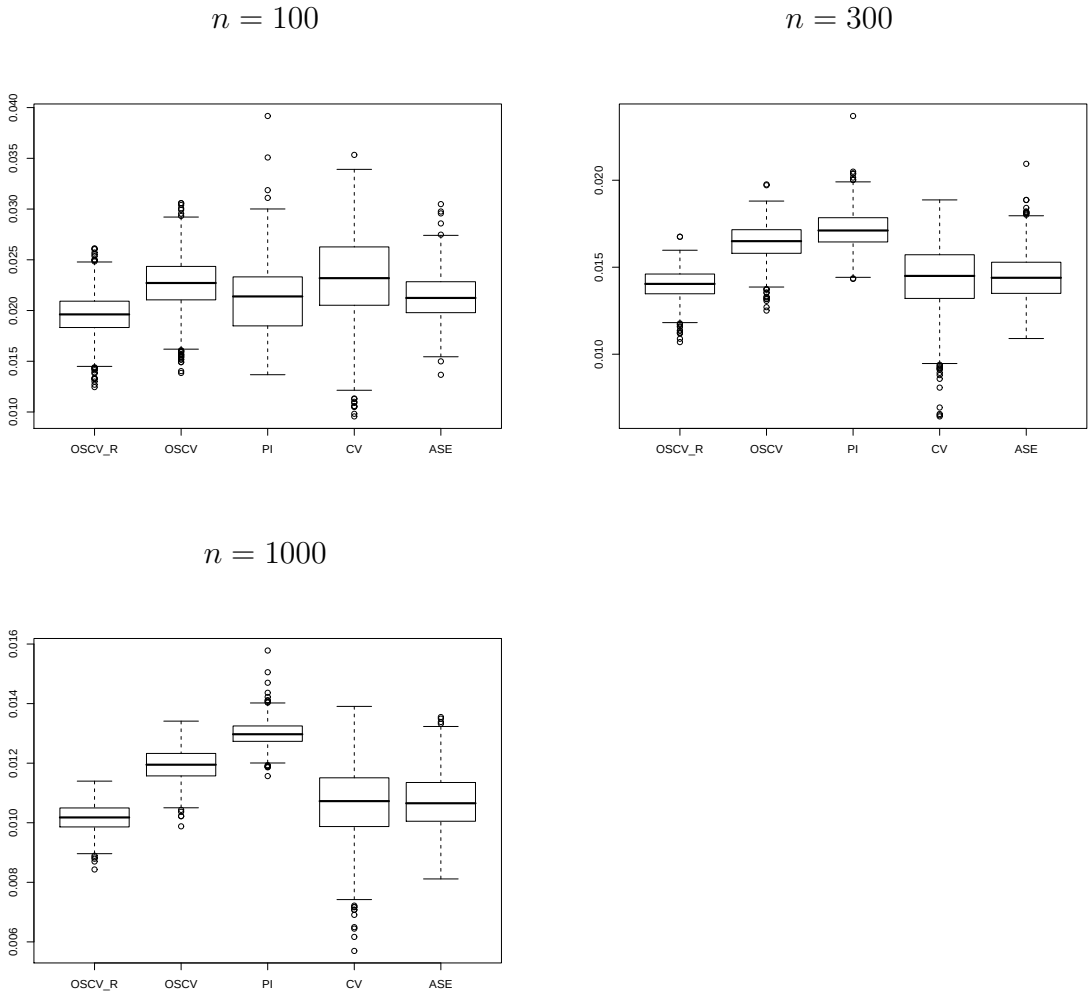


Fig. 41. Boxplots for the bandwidths in the case of regression function r_3. The standard deviation of the added noise is $\sigma = 1/500$; the design is Uniform$(0, 1)$.

VITA

Olga Savchuk was born in Kyiv, Ukraine. She received her B.S. and M.S. in electrical engineering in June 2000 and June 2002, respectively, from National Technical University of Ukraine. She received her second M.S. and Ph.D. in statistics from Texas A&M University in May 2006 and August 2009, respectively. Her current research interests lie in nonparametric function estimation. Her permanent address is: Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX, 77843-3143. Email: olgasavchuk@tamu.edu

The typist for this dissertation was Olga Savchuk.