Bandwidth Selection in Nonparametric Kernel Estimation

Dissertation

presented for the degree of Doctor rerum politicarum
at the Faculty of Economic Sciences
of the Georg-August-Universität Göttingen

by Anja Schindler from Annaberg-Buchholz, Germany

Göttingen, 2011

First Examiner: Prof. Dr. Stefan Sperlich
Second Examiner: Prof. Dr. Inmaculada Martínez-Zarzoso
Disputation: 29.09.2011

Acknowledgement

I want to thank my supervisor, Prof. Dr. Stefan Sperlich, Professor at the University of Geneva, for the ideas and his academic support. Furthermore I want to thank Prof. Dr. Inmaculada Martínez-Zarzoso and Prof. Dr. Thomas Kneib for their willingness to act as examiners. I gratefully acknowledge the work together with the co-authors of the first two parts.

Moreover I am grateful to Prof. Dr. Walter Zucchini for helpful suggestions and interesting discussions. I thank Dr. Oleg Nenadić for the good teaching work, for calming me down, when I was nervous or fret about something, he helped me with special details in the dissertation and above all I thank him for partly proof-reading.

Special thanks are due to my colleagues at the Institute of Statistics and Econometrics at the University of Göttingen, for the encouragement, it has been a pleasure to work there for the last years. Especially, I am very grateful to my colleague Daniel Adler to have always a sympathetic ear for me, for his straightforwardness, for constructive criticism, and for helping me with computer problems.

Finally I want to express my love to my family for their support, that they always believe in me, for giving me encouragement. Having such a strong family background is invaluable for life.

Contents

List of Figures

List of Tables

Notation

Symbols

X,Y,Z	random variables
x,y,z	scalars (realizations of X,Y,Z)
$X_1,\ldots,X_n$	random sample of size n
f	probability density function (pdf)
f_X	pdf of X
h	bandwidth
$\widehat{h}_\bullet$	ISE respectively ASE optimal bandwidth for a specific method
$h_\bullet$	(A)MISE optimal bandwidth
b	one sided bandwidth
$\widehat{b}_\bullet$	OISE optimal bandwidth for a specific method
$b_\bullet$	MOISE optimal bandwidth
$\widehat{f}_h$	estimated density function
$\widehat{f_{h,-i}}$	estimated density function when leaving out observation i
$\widehat{f}_{left,b}$	one-sided to the left kernel density estimator
K	symmetric kernel function
K_h	scaled kernel function, i.e. $K_h(u) = K(u/h)/h$
$\bar{K}$	local linear version of a one sided kernel
$\mu_l(K)$	lth moment of K, i.e. $\int u^l K(u)du$
$K \star K$	convolution of K, i.e. $K \star K(u) = \int K(u-v)K(v)dv$
$\Vert K\Vert_2^2$	squared L_2 norm of K, i.e. $\int [K(u)]^2 du$
$\xrightarrow{a.s.}$	almost sure convergence
$\overset{P}{\to}$	convergence in probability
α	weighting factor with $\alpha \in (0,1)$
$O_p(\bullet)$ / $O(\bullet)$ / $o(\bullet)$	Landau notation
ε_i	random variable with mean zero and unit variance
$\sigma^2(x)$	conditional variance, i.e. $Var(Y \vert X = x)$
$m(x)$	true regression function, i.e. $E(Y \mid X = x)$
β_j	parameter
$\hat{m}_h(x)$	local linear estimator of the regression function
$S_{h,j}$	weighted sums, i.e. $\sum_{i=1}^n K_h(x - X_i)(X_i - x)^j$

$W_{h,i}(x)$	weight for the local linear estimator
$w(X_j)$	trimming respectively weight function
$\Xi(.)$	penalizing function
$\hat{m}_{h,-i}$	estimated regression function when leaving out observation i
$\overset{\mathscr{L}}{\rightarrow}$	convergence in distribution
$\lfloor n \rfloor$	largest integer not greater than n
$\widehat{m}_{Q_j}$ / $\widehat{m}_{P_j}$	quartic / parabolic OLS estimator for block j
θ_{rs}	unknown functional, i.e. $\int_S m^{(r)}(x)m^{(s)}(x)f(x)dx$
$\hat{\theta}^Q_{rs}$ / $\hat{\theta}^P_{rs}$	blocked quartic / parabolic estimator of θ_{rs}
$\hat{\sigma}^2_Q$ / $\hat{\sigma}^2_P$	blocked quartic / parabolic estimator of σ^2

Abbreviations

cf.	compare (Latin: confer)
i.i.d.	independent and identically distributed
e.g.	for example (Latin: exempli gratia)
i. e.	that is (Latin: id est)
w.r.t.	with respect to
AMISE	asymptotic MISE
AMSE	asymptotic MSE
ARE	asymptotic relative efficiency
ASE	averaged squared error
IQR	interquartile range
ISE	integrated squared error
MASE	mean averaged squared error
MISE	mean integrated squared error
MOISE	one-sided MISE
MSE	mean squared error
OISE	one-sided ISE
OS	one sided
rot	rule-of-thumb
RSS	residual sum of squares

Important Methods

AIC	Akaikes information criterion
CV	cross-validation
DoV	double one sided cross-validation
DPI	direct PI
FGLS	feasible generalized least squares
FPE	finite prediction error

GCV	generalized CV
LSCV	least-squares CV
OLS	ordinary least squares
OSCV	one-sided CV
OSCV-l	OSCV to the left
OSCV-r	OSCV to the right
PI	plug-in
PIP	PI with a blocked parabolic parametric fit
PIQ	PI with a blocked quartic parametric fit
refPI	refined PI
Rice	Rice's T
SB	smoothed bootstrap
SBG	smoothed bootstrap with Gaussian kernel
Shib	Shibata's model selector
STAB	Stabilized bandwidth selection
WB	wild bootstrap
Mix	weighted mixture of two methods

Introduction

Kernel estimation is a common nonparametric method for data based estimation of densities or regression functions. Although one may consider nonparametric estimation as an estimation procedure without parameters, one has to estimate bandwidth parameters. The difference to parameter based estimation of e.g. density functions is that no specific form of the nonparametric density has to be assumed. This makes nonparametric estimation methods more flexible.
This thesis compromises three parts. The first part covers bandwidth selection in kernel density estimation, which is a common tool for empirical studies in many research areas. The discussion about finding the optimal bandwidth based on the data has been going on over three decades. The typical aim of empirical studies in the past was mostly to show that a new method outperforms existing ones. Review articles on comparing methods are very rare and were written a long time ago. Hence, chapter one of this thesis is an update review of existing methods comparing them on a set of different designs. The second part is on bandwidth selection in nonparametric kernel regression. The aim is similar to the first part: reviewing and comparing existing methods on a set of designs. In part one and two, smooth densities of a random variable X were assumed, therefore global bandwidth selection is adequate for the kernel estimation. In contrast to the first two parts we assume a density of X with a sharp peak and smooth areas in the third part. Usually local bandwidth selection is used in this case. However, we want to apply global bandwidth selection methods and hence, it is tested if good results can be obtained by a prior transformation. Therefore, part three covers a comparison between using a transformation and estimating the global bandwidth without a transformation. The main question is whether an improvement with respect to the typical error criteria in nonparametric regression can be made by using a prior transformation. Since the methods were extensively reviewed in the second part, only those who performed best were considered in this chapter. Then the estimation with and without prior transformation is compared, in order to evaluate the performance of the proposed transformation.

The thesis is written in LaTeX. The simulations are written in the programming languages Fortran, C and R. The evaluation and presentation of the results was written in R, since R provides the possibility to visualize the performance in nice graphics. The programming code is available from the author.

Chapter 1

Bandwidth Selection Methods for Kernel Density Estimation - A Review of Performance

Abstract

On the one hand, kernel density estimation is a common tool for empirical studies in any research area. This goes hand in hand with the fact that these estimators are provided by many software packages. On the other hand, since about three decades the discussion on bandwidth selection has been going on. Although a good part of the discussion is concerned about nonparametric regression, this issue is by no means less problematic for density estimation. This becomes obvious when reading empirical studies in which practitioners made use of kernel densities. New contributions typically provide simulations limited to show that the own invention outperforms existing methods. We review existing methods and compare them on a set of designs that exhibits features like few bumps and exponentially falling tails. We concentrate on small and moderate sample sizes. This essay is based on a joint work with my colleague Nils-Bastian Heidenreich and Prof. Dr. Stefan Sperlich. The main contribution of the author of this thesis is made in the evaluation of the plug-in and bootstrap methods and the evaluation of the estimation results.

1.1 Introduction

Suppose we have observed i.i.d. data $X_1, X_2, \ldots, X_n$ from a common distribution with density $f(\cdot)$, and we aim to estimate this density using the standard kernel (i.e. the Parzen-Rosenblatt) estimator

$$\widehat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \tag{1.1}$$

where K is a kernel and h the bandwidth parameter. The problem is to find a reliable data driven estimator of the optimal bandwidth. First one has to decide on a method of assessing the performance of $\widehat{f}_h$. The generally accepted performance measures are the integrated squared error

$$\mathrm{ISE}(h) \;=\; \mathrm{ISE}\{\widehat{f}_h(x)\} = \int \{\widehat{f}_h(x) - f(x)\}^2 \, dx \tag{1.2}$$

or alternatively, the mean integrated squared error, i.e.

$$\mathrm{MISE}(h) = \mathrm{MISE}\left[\hat{f}_h(x)\right] = \int \mathrm{MSE}\left[\hat{f}_h(x)\right] \, dx. \tag{1.3}$$

Let us denote the minimizers of these two criteria by $\widehat{h}_0$ and h_0 respectively. The main difference is that $ISE(h)$ is a stochastic process indexed by $h > 0$, while $MISE(h)$ is a deterministic function of h, see Cao (1993). Therefore we distinguish two classes of methods: the cross-validation methods trying to estimate $\widehat{h}_0$ and therefore looking at the ISE, and the plug-in methods which try to minimize the MISE to find h_0. It is evident that asymptotically these criteria coincide.

The main part of the nonparametric statistical community has accepted that there may not be a perfect procedure to select the optimal bandwidth. However, we should be able to say which is a reasonable bandwidth selector, at least for a particular problem. SiZer tries to show the practitioner what is a range of reasonable bandwidths, and is therefore quite attractive for data snooping, see Chaudhuri and Marron (1999) for an introduction, Godtliebsen, Marron and Chaudhuri (2002) for an extension to the bivariate case. Hanning and Marron (2006) made an improvement using extreme value theory. However, SiZer does not give back one specific data driven bandwidth as practitioners typically ask for.

Since until now the development of bandwidth selectors has been continuing, we believe it is helpful to review and compare the existing selectors to get an idea of the objective and performance of each selector. As we counted more than 30 bandwidth selectors - several of them being modifications for particular estimation problems - we decided to restrict this study in mainly two directions. Firstly, we considered independent observations. Secondly, we looked at smooth densities, namely we use four underlying distributions which are mixtures of at most three different normal and/or gamma distributions. This type of smoothness covers a broad range of problems in any research area; it is clearly rather different from estimating sharp peaks or highly oscillating functions. However, the latter problems should not be tackled with kernels anyway. Density problems with extreme tails are not included. It is well known that those problems should be transformed, see e.g. Wand, Marron and Ruppert (1991) or Yang and Marron (1999) for parametric, Ruppert and Cline (1994) for nonparametric transformations. After an appropriate transformation the remaining estimation problem falls into the here considered class, too. Note that the limitation to global bandwidths is not very restrictive neither, and quite common in density estimation. Actually, when X was transformed, and similar smoothness is assumed over the whole transformed support, then global bandwidths are most reasonable. Finally, we have restricted our study to already published methods.

The idea of cross validation methods goes back to Rudemo (1982) and Bowman (1984), but we should also mention the so-called pseudo-likelihood CV-methods invented by Habbema, Hermans and van den Broek (1974) and by Duin (1976). Due to the lack of stability of this method, see e.g. Wand and Jones (1995), different modifications have been proposed like the stabilized bandwidth selector recommended by Chiu (1991), smoothed CV proposed by Hall, Marron and Park (1992), the modified CV (MCV) by Stute (1992), or the version by Feluch and Koronacki (1992), and most recently the one-sided CV by Martínez-Miranda, Nielsen and Sperlich (2009) and the indirect CV by Savchuk, Hart and Sheather (2010). The biased CV (BCV) of Scott and Terrell (1987) is minimizing the asymptotic MISE like plug-in methods do but uses a jack-knife procedure (therefore called CV) to avoid the use of prior information. The recent kernel contrast method of Ahmad and Ran (2004) can be used for MISE minimization as well, but it is not really data adaptive (or fully automatic) and it performs particularly well rather for regression than for density estimation.

Compared to CV the so-called plug-in methods do not only minimize a different objective function, MISE instead of ISE, they are less volatile but not entirely data adaptive as they require pilot information. In contrast, CV allows to choose the bandwidth without making assumptions about the smoothness (or the like) to which the unknown density belongs. Certainly, if we have an appropriate pilot bandwidth the performance of plug-in methods is pretty good. Although, they have a faster convergence rate compared to CV, they can heavily depend on the choice of pilots. Among them, Silverman's (1986) rule of thumb is probably the most popular one. Various refinements were introduced like for example by Park and Marron (1990), Sheather and Jones (1991), or by Hall, Sheather, Jones and Marron (1991). Also the bootstrap methods of Taylor (1989) as well as all its modifications, see e.g. Cao (1993) or Chacón, Montanero and Nogales (2008), we count to plug-in methods as they aim to minimize the MISE.

There are already several papers dealing with a comparison of different automatic data driven bandwidth selection methods though, most of them are older than ten years. In the seventies and early eighties survey papers about density estimation were published by Wegman (1972), Tartar and Kronmal (1976), Fryer (1977), Wertz and Schneider (1979) as well as Bean and Tsokos (1980). A short introduction to various methods of smoothing parameter selection without a simulation study was released by Marron (1988a) and Park and Marron (1990). Then, extensive simulations studies have been published by Park and Turlach (1992), Marron and Wand (1992), Cao, Cuevas and González Manteiga (1994) and Chiu (1996). A brief survey is also announced by Jones, Marron and Sheather (1996a) and a more comprehensive one in the companion paper Jones, Marron and Sheather (1996b). A very comprehensive simulation study has been published by Devroye (1997). Furthermore, Loader (1999) has published a comparison paper. In recent years to our knowledge only Chacón, Montanero and Nogales (2008) has published a comparison paper on this topic. However, they concentrate on Bootstrap methods and only compare LSCV and the plug-in version of Sheather and Jones (1991). The general criticism against the two classes of selection methods can be summarized as follows: CV leads to undersmoothing and breaks down for large samples, whereas plug-in depends on prior information and often works bad for small data sets and much curvature.

For the statements about asymptotic theory, we make the following assumptions on kernel and density. For some methods we will modify them.

(A1) The kernel K is a compactly supported density function on $\mathbb{R}$, symmetric around zero with Hölder-continuous derivative, K'.

(A2) $\mu_2(K) < \infty$, where $\mu_l(K) = \int u^l K(u)du$.

(A3) The density, f, is bounded and twice differentiable, f' and f'' are bounded and integrable, and f'' is uniformly continuous.

In our simulation study we restrict on selection methods which consider no higher order kernels. The main motivation for the usage of higher order kernels is their theoretical advantage of faster asymptotic convergence rates. However, their substantial drawback is a loss in the practical interpretability as they involve negative weights and can even give back negative density estimates. A good illustration of the understanding of higher order kernels can be found in Marron (1994).

In the context of asymptotic theory we are aware of the trade-off between the classical plug-in method and standard cross-validation. The plug-in has always smaller asymptotic variance compared to cross-validation (Hall and Marron, 1987a). To our knowledge, no other bandwidth selection rule has outperformed the asymptotic properties of the plug-in method. Although Hall and Johnstone (1992) stated that such methods must theoretically exist, they couldn't give any practical example.

1.2 Cross-Validation methods in density estimation

Recall the used performance measure, i.e. the integrated squared error (ISE):

$$\mathrm{ISE}(h) = \int \widehat{f}_h^2(x)\,dx - 2E\{\widehat{f}_h(X)\} + \int f^2(x)\,dx. \tag{1.4}$$

Evidently, the first term can be calculated from the data, the second can be expressed as the expected value of $\widehat{f}_h(X)$, and the third term can be ignored since it does not depend on the bandwidth. Note that estimating $E\{\widehat{f}_h(X)\}$ by $\frac{1}{n}\sum_{i=1}^n \widehat{f}_h(X_i)$ is inadequate due to the implicit dependency ($\widehat{f}_h$ depends on X_i). So the different modifications of CV basically vary in the estimation of the problematic second part.

Ordinary least squares cross-validation

This is a straightforward approach by just dropping X_i when estimating $f(X_i)$, called jack-knife estimator and denoted by $\widehat{f}_{h,-i}(X_i)$. It yields the *least-squares CV criterion*

$$\underset{h}{min}\ \mathrm{CV}(h) = \int \widehat{f}_h^2(x)\,dx - 2\frac{1}{n}\sum_{i=1}^n \widehat{f}_{h,-i}(X_i). \tag{1.5}$$

Stone (1984) showed that under the assumptions (A1)-(A3), the minimizing argument, $\widehat{h}_{CV}$, fulfills
$\text{ISE}(\widehat{h}_{CV})\left\{\min_h \text{ISE}(h)\right\}^{-1} \xrightarrow{a.s.} 1$. However, Hall and Marron (1987a) stated that this happens at the slow rate of $O_p(n^{-1/10})$. Many practitioners use this CV method because of its intuitive definition and its practical flavor. But as mentioned above, it is not stable, tends to undersmooth and often breaks down for large samples.

Modified cross-validation

Stute (1992) proposed a so-called modified CV (MCV). He approximated the problematic term by the aid of a Hajek projection. In fact, he showed that under some regularity assumptions given below, $2E[f_h(x)]$ is the projection of

$$\begin{aligned} S+\frac{1}{h}E\left[K\left(\frac{X_1-X_2}{h}\right)\right] &= S+\frac{1}{h}\int\int K\left(\frac{x-y}{h}\right) f(x)f(y)\,dx\,dy \\ &= S+\int f^2(y)dy+\frac{1}{2}h^2\int t^2K(t)dt\int f(y)f''(y)dy+O(h^3) \\ \text{for} \qquad S &:= \frac{1}{n(n-1)h}\sum_{i\neq j}K\left(\frac{X_i-X_j}{h}\right). \end{aligned}$$

This gives the criterion

$$\min_h \; MCV(h)=\int \widehat{f}_h^2(x)dx-S-\frac{\mu_2(K)}{2n(n-1)h}\sum_{i\neq j}K''\left(\frac{X_i-X_j}{h}\right). \qquad (1.6)$$

It can be shown then that under assumptions (A1),

(A2') K is three times differentiable, with $\int t^4|K(t)|\,dt<\infty$, $\int t^4|K''(t)|\,dt<\infty$, $\int t^4[K'(t)]^2\,dt<\infty$, and $\int t^2[K'''(t)]^2\,dt<\infty$,

(A3') f four times continuously differentiable, the derivatives being bounded and integrable,

you get the following consistency result:

$$\frac{\text{ISE}(\widehat{h}_0)}{\text{ISE}(\widehat{h}_{\text{MCV}})}\xrightarrow{P}1, \quad and \quad \frac{\widehat{h}_0}{\widehat{h}_{\text{MCV}}}\xrightarrow{P}1 \quad as \quad n\to\infty.$$

Stabilized bandwidth selection

Based on characteristic functions Chiu (1991) gave an expression for h_{CV} which reveals the source of variation. Note that the CV criterion is approximately equal to

$$\frac{1}{\pi}\int_0^\infty |\tilde{\phi}(\lambda)|^2\left\{w^2(h\lambda)-2w(h\lambda)\right\}d\lambda+2K(0)/(nh), \qquad (1.7)$$

with $\tilde{\phi}(\lambda) = \frac{1}{n}\sum_{j=1}^{n} e^{i\lambda X_j}$ and $w(\lambda) = \int e^{i\lambda u}K(u)du$. The noise in the CV estimate is mainly contributed by $|\tilde{\phi}(\lambda)|^2$ at high frequencies, which does not contain much information about f. To mitigate this problem, he looks at the difference of the CV criterion and the MISE. He defines Λ as the first λ fulfilling $|\tilde{\phi}(\lambda)|^2 \leq 3/n$ and replaces $|\tilde{\phi}(\lambda)|^2$ by $1/n$ for $\lambda > \Lambda$. This gives his criterion:

$$\begin{aligned} \min_h \; S_n(h) &= \int_0^{\Lambda} |\tilde{\phi}(\lambda)|^2 \left\{w^2(h\lambda) - 2w(h\lambda)\right\} d\lambda \\ &\quad + \frac{1}{n}\int_{\Lambda}^{\infty} \left\{w^2(h\lambda) - 2w(h\lambda)\right\} d\lambda + 2\pi K(0)/(nh), \qquad (1.8) \\ &= \frac{\pi}{nh}||K||_2^2 + \int_0^{\Lambda} \left\{|\tilde{\phi}(\lambda)|^2 - \frac{1}{n}\right\} \{w^2(h\lambda) - 2w(h\lambda)\} d\lambda, \qquad (1.9) \end{aligned}$$

with $||g||_2^2 = \int g^2(u)\,du$.
For the minimizer, $\widehat{h}_{ST}$, of this criterion, it can be shown that $\widehat{h}_{ST} \xrightarrow{a.s.} \widehat{h}_0$, and it converges to h_0 at the optimal $n^{-1/2}$-rate. In the calculation of Λ we came across with the computation of square roots of negative terms in our simulations. To avoid complex numbers we calculated the absolut value of the radicand. Note that in the literature this procedure is often counted among the plug-in methods as it minimizes the MISE.

One-sided cross-validation

Marron (1986) made the point that the harder the estimation problem the better CV works. Based on this idea Hart and Yi (1998) introduced an estimation procedure called one-sided cross-validation in the regression context. They concluded that OSCV in regression clearly outperforms the ordinary CV. Martínez-Miranda, Nielsen and Sperlich (2009) proposed to apply one-sided cross-validation in density estimation. They apply CV on a harder estimation problem and afterwards calculate the corresponding bandwidth for the underlying ”real” estimation problem. To make the estimation problem harder, they use a worse estimator, namely (1.1) but with a local linear version of a one sided kernel,

$$\bar{K}(u) = \frac{\mu_2(K) - u\left(2\int_{-\infty}^{0} tK(t)\,dt\right)}{\mu_2(K) - \left(2\int_{-\infty}^{0} tK(t)\,dt\right)^2} 2K(u)\mathbf{1}_{\{u<0\}}. \qquad (1.10)$$

Respectively to ISE and MISE they define the one-sided versions OISE and MOISE, with their minimizers $\widehat{b}_0$ and b_0. The one-sided CV criterion is

$$\min_b \; \text{OSCV}(b) = \int \widehat{f}_{left,b}^2(x)\,dx - \frac{2}{n}\sum_{i=1}^{n} \widehat{f}_{left,b}(X_i), \qquad (1.11)$$

where $\widehat{f}_{left,b}$ is the one-sided (to the left) kernel density estimator. Then they define the corresponding bandwidth for the ”real” estimation problem by

$$\widehat{h}_{OSCV} := C \cdot \widehat{b}_{OSCV} \quad \text{with} \quad C = h_0/b_0. \qquad (1.12)$$

Note that C is deterministic and depends only on kernel K since

$$h_0 = \left(\frac{||K||_2^2}{(\mu_2(K))^2 ||f''||_2^2 n} \right)^{1/5} \quad , \quad b_0 = \left(\frac{||\bar{K}||_2^2}{(\mu_2(\bar{K}))^2 ||f''||_2^2 n} \right)^{1/5} . \tag{1.13}$$

This gives, for example $C \approx 0.537$ for the Epanechnikov kernel. The theoretical justification for the improved convergence rate of one-sided CV is based on the result of Hall and Marron (1987a) that under the assumptions (A1) - (A3)

$$n^{3/10}(\widehat{h}_{CV} - \widehat{h}_0) \longrightarrow N(0, \sigma^2 c^{-2}). \tag{1.14}$$

with known terms σ and c depending only on f and K. From this we can calculate the variance reduction of OSCV compared to CV by $\{C\bar{\sigma}c/(\bar{c}\sigma)\}^2$ where $\bar{c}$, $\bar{\sigma}$ are just as c, σ but with $\bar{K}$ instead of K. The reduction of the variance for the Epanechnikov kernel is at least 35% and 50% for the Gaussian kernel. Note that $\bar{K}$ can also be constructed as a one sided kernel to the right.

Further cross-validation methods

Feluch and Koronacki (1992) proposed to cut out not only X_i when estimating $f(X_i)$ but rather dropping also the $m < n$ nearest neighbors with $m \rightarrow \infty$ such that $m/n \rightarrow 0$. They called this version *modified CV*. Unfortunately, it turned out that the quality of this method crucially depends on m. Therefore it cannot be considered as *automatic* or *data driven*, and will not be considered further. This idea is similar to the CV selection for time series data, see Härdle and Vieu (1992). Scott and Terrell (1987) introduced the Biased CV. They worried about unreliable small-sample results, i.e. the high variability while using the cross-validation criterion. However, they directly focused on minimizing the asymptotic MISE and estimated the unknown term $||f''(x)||_2^2$ via jack-knife methods. Already in their own paper they admitted a poor performance for small samples and mixtures of densities, see also Chiu (1996). In their simulation study, Jones, Marron and Sheather (1996b) underlined the deficient performance from 'quite good' to 'very poor'. The smoothed cross-validation (SCV) was evolved by Hall, Marron and Park (1992). The general idea is a kind of presmoothing of the data before applying the CV-criterion. This procedure of presmoothing results in smaller sample variability, but enlarges the bias. Therefore the resulting bandwidth is often oversmoothing and cuts off some features of the underlying density. With this method it is possible to achieve a relative order of convergence of $n^{-1/2}$ but only when using a kernel of order ≥ 6. In total, it seems to be appropriate - if at all - only for huge samples. Without using a higher order kernel Jones, Marron and Sheather (1996b) stated, that there exists an $n^{-1/10}$ convergent version of SCV that is identical to Taylor's bootstrap, see Taylor (1989). Additionally, with a special choice for g SCV results in an $n^{-5/14}$ version similar to a diagonal-in version of Park and Marron's plug-in, see Park and Marron (1990). Note that finally the SCV is closely related to the bootstrap method of Cao (1993). These three methods do not belong to the cross-validation methods, and hence, they are discussed later. In conclusion, we have not implemented these methods, because either it is very similar to other methods or it is necessary to use a higher order kernel.

The partitioned cross-validation (PCV) was suggested by Marron (1988b). He modified the CV-criterion by splitting the sample of size n into m subsamples. Then, the PCV is calculated by minimizing the average of the score functions of the CV-score for all subsamples. In a final step the resulting bandwidth needs to be rescaled. The number of subsamples affects the trade off between variance and bias. Hence the choice of m is the smoothing problem in this case. As Park and Marron (1990) noticed: ”this method ... is not quite fully objective”. Another drawback is the required separation of the subsamples.

The pseudo-likelihood (also called the Kullback-Leibler) cross-validation (invented by Habbema, Hermans and van den Broek (1974) and by Duin (1976)) aims to find the bandwidth maximizing a pseudo-likelihood criterion with leaving out the observation X_i. Due to the fact that lot of authors criticize this method being inappropriate for density estimation, we skipped also this method in our simulation study.

Wegkamp (1999) suggest a method being very much related to the cross-validation technique providing quasi-universal bandwidth selection for bounded densities. Nonetheless, his paper stays on a rather technical level but is not suitable for practitioners. Also Savchuk, Hart, and Sheather (2010) introduce a method with excellent theoretical properties, based on indirect cross-validation. For our implementation with Epanechnikov kernels it nevertheless worked well only for large samples.

Recently, Ahmad and Ran (2004) proposed a kernel contrast method for choosing bandwidths either minimizing ISE or alternatively the MISE. While it turned out to work quite well for regression, the results for density estimation were less promising. A major problem is that one needs two series of contrast coefficients which have a serious impact on the performance of the method. As we are not aware of an automatic data driven and well performing method to choose them, we will not consider this method further.

1.3 Plug-in methods in density estimation

Under (A1)-(A3) the MISE can be written for $n \to \infty$, $h \to 0$ as

$$\text{MISE}\left[\hat{f}_h(x)\right] = \frac{h^4}{4}\mu_2^2(K)||f''(x)||_2^2 + \frac{1}{nh}||K||_2^2 + o\left(\frac{1}{nh}\right) + o(h^4), \tag{1.15}$$

such that the asymptotically optimal bandwidth is

$$h_0 = ||K||_2^{2/5}\left(||f''||_2^2\,[\mu_2(K)]^2\,n\right)^{-1/5}, \tag{1.16}$$

where only $||f''||_2^2$ is unknown and has to be estimated. The most popular method is the ”rule-of-thumb” introduced by Silverman (1986). He uses the normal density as a prior for approximating $||f''||_2^2$. For the necessary estimation of the standard deviation of X he proposes a robust version making use of the interquartile range. If the true underlying density is unimodal, fairly symmetric and does not have fat tails, it works fine.

Park and Marron's Refined plug-in

Natural refinements consist of using nonparametric estimates for $||f''||_2^2$. Let us consider

$$\widehat{f}_g''(x) = \frac{1}{ng^3}\sum_{i=1}^{n} K''\left(\frac{x-X_i}{g}\right),$$

where g is a prior bandwidth. Hall and Marron (1987b) proposed several estimators for $||f''||_2^2$, all containing double sums over the sample. They pointed out that the diagonal elements give a non-stochastic term which does not depend on the sample and increases the bias. They therefore proposed the bias corrected estimator

$$\widehat{||f''||_2^2} = ||\widehat{f}_g''||_2^2 - \frac{1}{ng^5}||K''||_2^2. \tag{1.17}$$

The question which arises is how to obtain a proper prior bandwidth g. In Park and Marron (1990) g is the minimizer for the asymptotic mean squared error of $\widehat{||f''||_2^2}$. With (1.16) one gets a prior bandwidth in terms of h (using the notation in the original paper):

$$g = C_3(K)C_4(f)h^{10/13},$$

where $C_3(K)$ contains the fourth derivative and convolutions of K, and $C_4(f)$ the second and third derivatives of f. Substituting the normal with estimated variance for f gives

$$h = \left(\frac{||K||_2^2}{\widehat{||f''||_2^2}\mu_2^2(K)n}\right)^{1/5}. \tag{1.18}$$

The optimal bandwidth is then obtained by numerical solution of this equation. The relative rate of convergence to h_0 is of order $O_p(n^{-4/13})$, which is suboptimal compared to the optimal $n^{-1/2}$-rate, cf. Hall and Marron (1991).

Implemented Refined plug-in

For small samples and small (optimal) bandwidths, the above estimator $\widehat{||f''||_2^2}$ can easily fail in practice. Also, to find a numerical solution may become involved in practice. To avoid these problems and to offer a quick and easy solution, we propose to first take Silverman's rule-of-thumb bandwidth for Gaussian kernels, $h_S = 1.06\min\{1.34^{-1}IR, s_n\}n^{-1/5}$ with IR =interquartile range of X, and s_n the sample standard deviation, adjusted to Quartic kernels. This is done via the idea of canonical kernels and equivalence bandwidths, see Härdle, Müller, Sperlich, and Werwartz (2004). The Quartic which comes close to the Epanechnikov kernel but allows for second derivative estimation. Finally, we adjust for the slower optimal rate for second derivative estimation and obtain as a prior

$$g = h_S\frac{2.0362}{0.7764}n^{1/5-1/9} \tag{1.19}$$

for (1.17). This bandwidth leads to very reasonable estimates of the second derivative of f, and hence of $\widehat{||f''||_2^2}$. A further advantage is that this prior g is rather easily obtained. As the idea actually goes back to Park and Marron (1990) we will call the final bandwidth $\hat{h}_{PM}$.

Bootstrap methods

The idea of these methods is to select the bandwidth along bootstrap estimates of the ISE or the MISE. For a general description of this idea in nonparametric problems, see Hall (1990). Imagine, for a given pilot bandwidth g we have a Parzen-Rosenblatt estimate, $\widehat{f}_g$, from which we can draw bootstrap samples $(X_1^*, X_2^*, \ldots, X_n^*)$. Then, defining the bootstrap kernel density

$$\widehat{f}_h^*(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i^*}{h}\right), \tag{1.20}$$

the (mean) integrated squared error to be minimized can be approximated by

$$\begin{aligned} ISE^*(h) \quad &:= \quad \int \left(\widehat{f}_h^*(x) - \widehat{f}_g(x)\right)^2 dx\,, & (1.21)\\ MISE^*(h) \quad &:= \quad E_* \left[\int \left(\widehat{f}_h^*(x) - \widehat{f}_g(x)\right)^2 dx\right]\,. & (1.22) \end{aligned}$$

It can be shown that the expectation E_* and so the $MISE^*$ depends only on the original sample but not on the bootstrap samples. Consequently, there is actually no need to do resampling to obtain the $MISE^*$. More specific, using Fubini's theorem and decomposing the $MISE^* = V^* + SB^*$ into integrated variance

$$V^*(h) = \frac{1}{nh} \cdot ||K||_2^2 + \frac{1}{n} \cdot \int \left(\int K(u) \cdot \widehat{f}_g(x - hu)\, du\right)^2 dx \tag{1.23}$$

and squared bias

$$SB^*(h) = \int \left(\int K(u) \cdot (\widehat{f}_g(x - hu) - \widehat{f}_g(x))\, du\right)^2 dx \tag{1.24}$$

gives (where $\star$ denotes convolution)

$$V^*(h) = \frac{1}{nh}||K||_2^2 + \frac{1}{n^3} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[(K_h \star K_g) \star (K_h \star K_g)\right](X_i - X_j) \tag{1.25}$$

and

$$SB^*(h) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[(K_h \star K_g - K_g) \star (K_h \star K_g - K_g)\right](X_i - X_j). \tag{1.26}$$

In practice, it is hard to get explicit formulae for these integrals when kernels have bounded support. However, using the Gaussian kernel in the formulae (1.25) and (1.26) we can directly calculate the optimal bandwidth as the minimizer of

$$\begin{aligned} \text{MISE*}(h) \;=\; & \frac{1}{2nh\sqrt{\pi}} + \frac{1}{\sqrt{2\pi}} \left[\frac{\sum_{i,j}\left(\exp\left(-\frac{1}{2}\left(\frac{X_i - X_j}{g\sqrt{2}}\right)^2\right)\right)}{\sqrt{2g^2} \cdot n^2} \right. & (1.27)\\ & \left. - \frac{2 \cdot \sum_{i,j}\left(\exp\left(-\frac{1}{2}\left(\frac{X_i - X_j}{\sqrt{h^2 + 2g^2}}\right)^2\right)\right)}{\sqrt{h^2 + 2g^2} \cdot n^2} + \frac{(n+1)\sum_{i,j}\left(\exp\left(-\frac{1}{2}\left(\frac{X_i - X_j}{\sqrt{2(h^2 + g^2)}}\right)^2\right)\right)}{\sqrt{2(h^2 + g^2)} \cdot n^3} \right] \end{aligned}$$

The equivalent bandwidth for any other kernel can be obtained as described in Marron and Nolan (1988).

The bootstrap approach in kernel density estimation was first presented by Taylor (1989). However, many modified versions were published in the following, e.g. Faraway and Jhun (1990), Hall (1990) or Cao (1993). The crucial difference between these versions is the choice of the pilot bandwidth and the procedure to generate the resampling distribution.

Taylor (1989) suggested to take $g = h$ and used a Gaussian kernel. Several authors pointed out that this procedure has no finite minimum and hence chooses a local minimum or the upper limit of the bandwidths grid as its optimum. This leads to an inappropriate choice and a bias towards oversmoothing, see Marron (1992). Differing from this approach, Faraway and Jhun (1990) proposed a least-square cross-validation estimate to find g.

Hall (1990) recommended to use the empirical distribution to draw bootstrap samples of size $m < n$, proposed $m \simeq n^{1/2}$, $h = g(m/n)^{1/5}$, and minimized $MISE^*$ with respect to g. Cao, Cuevas and González-Manteiga (1994) demonstrated that the bootstrap version of Hall is quite unstable and shows a bad behavior especially for mixtures of normal distributions, which make up the biggest part of our simulation study. They found also that the methods of Faraway and Jhun (1990) as well as the one of Hall (1990) are outperformed by the method of Cao (1993), see below.

A bias corrected bootstrap estimate was developed by Grund and Polzehl (1997). They obtained an root-n convergent estimate which attained very good results for larger sample sizes, but only for few cases for moderate and small sample sizes. Moreover, to derive their asymptotic theory they used extraordinary strong assumptions, compared to other methods discussed here. In their simulation study Grund and Polzehl showed that the performance heavily depends on the choice of g. They stated that using their oversmoothing bandwidth, which provides a root-n convergence, seems to be far from optimal for smaller sample size. In contrast, using $g = h$ would achieve better performance in practical applications, but results in very slow convergence rate, namely of order $n^{-1/10}$. Summing up, they remarked that higher rates of convergence do not result in better practical performance, especially for small samples.

In the smoothed bootstrap version of Cao (1993) the pilot bandwidth g is estimated by asymptotic expressions of the minimizer of the dominant part of the mean squared error. For further details see Cao (1993). He noticed that in (1.26), for $i = j$ this terms will inflate the bias artificially. He therefore proposed a modified bootstrap integrated squared bias MB*

$$MB^*(h) = \frac{1}{n^2} \sum_{i \neq j} \left[(K_h \star K_g - K_g) \star (K_h \star K_g - K_g) \right] (X_i - X_j). \tag{1.28}$$

As to what concerns the convergence rates, he showed for his bandwidth h_0^*

$$\frac{\text{MISE}(h_0^*) - \text{MISE}(h_0)}{\text{MISE}(h_0)} = O_P(n^{-5/7}) \tag{1.29}$$

and

$$\frac{\text{MISE}(h_{0_M}^*) - \text{MISE}(h_0)}{\text{MISE}(h_0)} = O_P(n^{-8/13}) \tag{1.30}$$

The convergence rate for the original bootstrap version is slightly faster than that for his modified bootstrap version.

Recently, Chacón, Montanero and Nogales (2008) published a bootstrap version quite similar to Cao's (1993). They showed that the asymptotic expressions of his bandwidth estimates might be inadequate and defined an expression $g(h)$ for every fixed h. Their estimation procedure allows different kernels L and K for the bandwidths g and h. They calculated the optimal pilot bandwidth $g(h)$ using first the common way of reverting to a reference distribution, and afterwards via estimation. In their simulation study they stated that the former version outperforms the empirical approach, and is a good compromise between classical cross-validation and plug-in. However, it depends seriously on the reference density. On the contrary, the empirical version suffered from sample variability even more than classical CV. Exploring the asymptotics, they achieved root-n convergence under the use of higher-order kernels.

In sum, in our simulation study we concentrate on just one representative of the class of bootstrap estimates, going back to Cao (1993). He proved that the pilot bandwidth g as the minimizer of (1.22) coincides with the minimizer of the dominant part of the mean squared error. Concretely, it is given by

$$g = \left(\frac{||K||_2^2}{\widehat{||f'''||_2^2} \mu_2^2(K) n} \right)^{1/7} . \tag{1.31}$$

This formula is used for the pilot bandwidth g in the calculation of (1.27). In our simulations, we additionally ran the bootstrap for the Epanechnikov kernel calculating formulae (1.23) and (1.24) numerically. As this was much slower and gave uniformly worse results, we will neglect it for the rest of the paper.

Further Plug-in methods

Many other plug-in methods have been developed. Some of them show better asymptotic properties and others a better performance in particular small sample simulations. However, most of them have not become (widely) accepted or even known.

An often cited method is the so-called Sheather and Jones (1991) bandwidth, see also Jones and Sheather (1991). They used the same idea like Park and Marron (1990) but replaced the "diagonal-out" estimator of $||f''||_2^2$ by their "diagonal-in" version to avoid the problem that the estimator $\widehat{||f''||_2^2}$ (see (1.17)) may give negative results. They stated that the non-stochastic term in (1.17) is subducted because of its positive effect on the bias in estimating $||f''||_2^2$. The idea is to choose the prior bandwidth g such that the negative bias due to the smoothing compensates the impact of the diagonal-in term. As a result they estimated $||f''||_2^2$ by $||\widehat{f}_g''||_2^2$ which is always positive, and obtained

$$g = C(K,L) \left(\frac{||f''||_2^2}{||f'''||_2^2} \right)^{1/7} h^{5/7},$$

where $C(K,L)$ depends on L, a kernel introduced to estimate $||f''||_2^2$, and K, the kernel in the original estimation. Then, $||f''||_2^2$ and $||f'''||_2^2$ were estimated using $||\widehat{f}_a''||_2^2$ and $||\widehat{f}_b'''||_2^2$, where

a and b were estimated via the rule-of-thumb. Sheather and Jones (1991) showed that their optimal bandwidth has a relative order of convergence to h_0 of $O_p(n^{-5/14})$ which is only slightly better than that of Park and Marron (1990). Jones, Marron and Sheather (1996b) indicates the closeness of h_{PM} to h_{SJ} for practical purposes in their real data application. Hence, without beating h_{PM} in practical performance, having only a slightly better convergence rate but being computationally much more expensive, we favor h_{PM} to h_{SJ}.

Hall, Sheather, Jones and Marron (1991) introduced a plug-in method giving back a bandwidth $\widehat{h}_{HSJM}$ which achieves the optimal rate of convergence, i.e. $n^{-1/2}$. The problem with $\widehat{h}_{HSJM}$ is that they use higher order kernels to ensure the $n^{-1/2}$ convergence (actually a kernel of order 6). Marron and Wand (1992) showed that albeit their theoretical advantages, higher order kernels have a surprisingly bad performance in practice, at least for moderate samples. Furthermore, in the simulation study of Park and Turlach (1992) $\widehat{h}_{HSJM}$ behaved very bad for bi- and trimodal densities, i.e. those we plan to study.

Jones, Marron and Park (1991) developed a plug-in method based on the smooth CV idea. They used the prior bandwidth $g = C(f)n^p h^m$, where the normal is used as reference distribution to calculate the unknown $C(f)$. The advantage of this estimator is the $n^{-1/2}$ convergence rate if $m = -2$, $p = \frac{23}{45}$ even if the kernels are of order 2. However, in simulation studies Turlach (1994) and Chiu (1996) observed a small variance compared to the LSCV, but an unacceptable large bias.

Kim, Park and Marron (1994) also showed the existence of a $n^{-1/2}$ convergent method without using higher order kernels. The main idea of obtaining asymptotically best bandwidth selectors is based on an exact MISE expansion. But primarily the results of this paper are provided for "theoretical completeness" because the practical performance in simulation studies for moderate sample sizes is rather disappointing, which was already explicitly mentioned in their own paper and as well shown in the exhaustive simulation study of Jones, Marron and Sheather (1996b).

For the sake of completeness, we also refer to the "Double Kernel method" based on the L_2 loss function, see Jones (1998). He explore an modification of the L_1 based method proposed by Devroye (1989), see also Berlinet and Devroye (1994). This method has the advantage to be very universal. Under special assumptions it reduces to Taylor's bootstrap respectively biased CV. However, as already mentioned, these two methods have several disadvantages and again the Double Kernel method requires the use of higher order kernels. In Jones (1998) the performance of the Double Kernel method is assessed by comparing asymptotic convergence rates, but it does not provide the expected improvement in the estimation of h_0 (MISE optimal bandwidth), e.g. compared to SCV.

Finally, for further plug-in methods recall also Ahmad and Ran (2004) or the so-called biased CV, both already introduced in the section about cross validation methods.

1.4 Mixtures of methods

Recall that all authors criticize that the cross-validation criterion tends to undersmooth and suffers from high sample variability. At the same time, the plug-in estimates deliver a much more stable estimate but often oversmooths the density. We therefore also consider mixtures of classical cross-validation methods and plug-in estimates. Depending on the weighting factor $\alpha \in (0,1)$, the mixed methods are denoted by Mix(α), with $\alpha \cdot \widehat{h}_{CV} + (1-\alpha) \cdot \widehat{h}_{PM}$. We mix in three different proportions: Mix(1/2), Mix(1/3) and Mix(2/3). For the resulting mixed bandwidths we calculate the according ISE-value to assess the performance of the respective mix proportion.

We are aware of different approaches which combine various density estimators by using a mixture of their smoothing parameters. In the literature several papers address the problem of linear and/or convex aggregation, e.g. Rigollet and Tsybakov (2007), Samarov and Tsybakov (2007) as well as Yang (2000). However, as the main focus of this paper is not on the aggregation of different bandwidth estimators, we will not investigate this much in detail, but instead consider our mixtures as representatives.

1.5 Finite Sample Performance

The small sample performance of the different cross-validation methods, plug-in and bootstrap methods is compared, including Chiu (1991). For obvious reasons we limited the study to data adaptive methods without boundary correction. Although we tried many different designs we summarize here the results for four densities.

We have compared the performance by different measures based on the integrated squared error (ISE) of the resulting density estimate (not the bandwidth estimate), and on the distance to the real optimal bandwidth $\widehat{h}_0$ (of each simulation run, as it is sample-dependent). There are a lot of measures assessing the quality of the estimators. We will concentrate on the most meaningful ones, that are:

m_1: $mean\left[ISE(\hat{h})\right]$, the average (or expected) ISE

m_2: $std\left[(ISE(\hat{h})\right]$, the volatility of the ISE

m_3: $mean(\hat{h} - \widehat{h}_0)$, bias of the bandwidth selectors

m_4: $mean\left(\left[ISE(\hat{h}) - ISE(\widehat{h}_0)\right]^2\right)$, squared L_2 distance of the ISEs

m_5: $mean\left[\mid ISE(\hat{h}) - ISE(\widehat{h}_0)\mid\right]$, L_1-distance of the ISEs.

Further, we considered various densities for our simulation study, but for sake of presentation we give only the results for the following ones:

1. Simple normal distribution, $\mathscr{N}(0.5, 0.2^2)$ with only one mode
2. Mixture of $\mathscr{N}(0.35, 0.1^2)$ and $\mathscr{N}(0.65, 0.1^2)$ with two modes
3. Mixture of $\mathscr{N}(0.25, 0.075^2)$, $\mathscr{N}(0.5, 0.075^2)$, $\mathscr{N}(0.75, 0.075^2)$ with three modes

4. Mixture of three gamma, $Gamma(a_j, b_j)$, $a_j = b_j^2$, $b_1 = 1.5$, $b_2 = 3$ and $b_3 = 6$ applied on $8x$ giving two bumps and one plateau

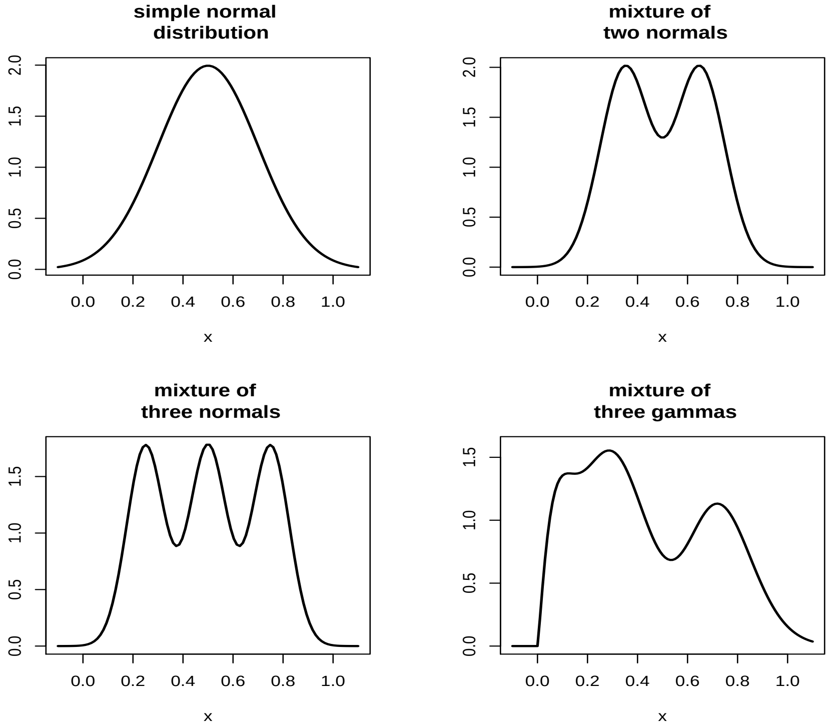


Figure 1.1: The data generating densities: design 1 to 4 from upper left to lower right.

As can be seen in Figure 1.1, all densities have the main mass in $[0,1]$ with exponentially decreasing tails. This way we can neglect possible boundary effects. Moreover, it is assumed that the empirical researcher has no knowledge on possible boundaries. We also simulated estimators with boundary corrections getting results very close to what we found in the present study.

We studied almost all selection methods, excluding the non-automatic ones and those having proved to perform uniformly worse than their competitors. In the presentation of the results we concentrate on the methods which delivered the best results at least for one density. Hence, some methods were dropped, e.g. the MCV sometimes provides multiple minima with a global minima far outside the range of reasonable bandwidths. In the range of bootstrap methods we concentrate on the presentation of the version (1.27) of the Smoothed Bootstrap which obtained the best results among all bootstrap methods. For our mixed version (CV with refined plug-in) we first concentrate on Mix(1/2) when comparing it to the other methods, and later sketch the results of all mixed versions.

To conclude, we present the following methods: CV (cross validation), OSCV-l (one-sided CV to the left), OSCV-r (one-sided CV to the right), STAB (stabilized), refPI (refined plug-in), SBG (smooth bootstrap with Gaussian kernel - the results refer to the equivalent bandwidth for the Epanechnikov kernel), Mix (mixed method for $\alpha = (1/2)$), and as a benchmark the ISE (infeasible ISE minimizing $\widehat{h}_0$).

1.5.1 Simulation results

In order to summarize the different methods of choosing the optimal bandwidth, we first consider the selected bandwidths and the corresponding biases for each method separately. Afterwards, we compare the methods by various measures. The shown results are based on 250 simulation runs.

Comparison of the bias for the different bandwidths

In Figure 2.7 we illustrate the Bias (m_3) of the different methods for the mixture of three normal distributions varying sample size and distribution.

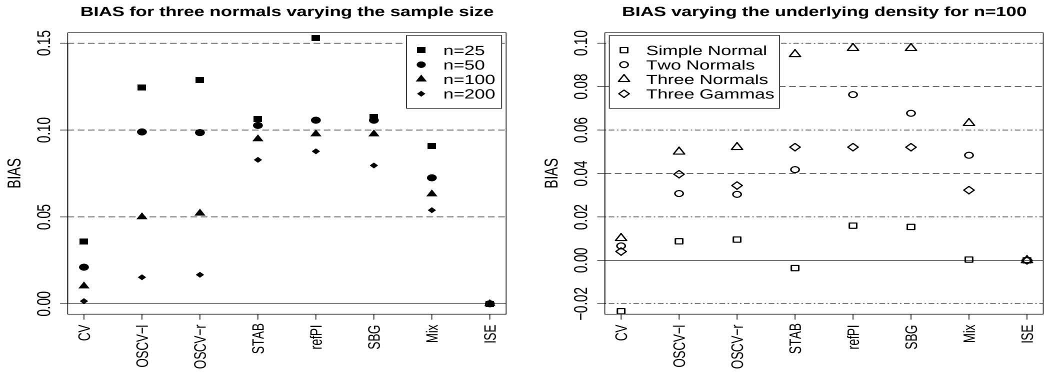


Figure 1.2: Comparison of the BIAS for different sample sizes (left) for a mixture of three normals (model 3), and different densities (right) for a sample size of n=100.

Let us consider the cross-validation method (CV). Many authors have mentioned the lack of stability of the CV-criterion and the tendency to undersmooth. In Figure 2.7 we see that CV has the smallest bias for all sample sizes and densities due to the fact that it chooses the smallest bandwidth. When the ISE optimal bandwidth is indeed very small, CV certainly does very well therefore. However, CV clearly undersmooths in the case of the simple normal distribution.

In contrast, the one-sided versions (OSCV) are more stable. Regarding the bias they are neither the best nor the worst in all sample sizes and models. As already stated by the authors, the OSCV tends to overestimate the bandwidth a little bit. While for $n = 25$ OSCV is outperformed by almost all other methods, this bias problem disappears rapidly for increasing n. In the left panel of Figure 2.7 we see that their biases are much smaller than for the other methods except CV, and STAB in the simple normal case. Moreover, their behavior is quite stable and they do not fail as dramatically as the other methods in one or more cases. This feature is an intuitive benefit of this method when in practice the underlying density is completely unknown. For the densities studied, the differences between the left-(OSCV-l) and the right-sided (OSCV-r) versions are negligible except for the gamma distributions because of the boundary effect that is present on the left side.

The stabilized procedure of Chiu (STAB) is excellent for the simple normal case but it falls short when estimating rougher densities: “hen the true density is not smooth enough, the stabilized procedure is more biased towards oversmoothing than CV” (Chiu (1991)). This fact can be seen in both panels of Figure 2.7 where STAB has increasing difficulties with an increasing number of bumps. Even though this method demonstrates here a reasonable performance, the results should be interpreted with care, since in the derivation of Λ one has to deal with complex numbers, a problem we solved in favor of this method for this simulation study such that all presented results are slightly biased in favor of STAB.

The refined plug-in (refPI) and the smoothed bootstrap SBG show a similar behavior as the stabilized procedure for $n = 100$, though the bias is much worse for refPI in small samples. Not surprisingly, in general, the bias for these MISE minimizing methods is larger than for all others. This partly results from the fact that we assume for the prior bandwidth that the second or third derivative comes from a simple normal distribution. Note that the bias of the SBG bandwidth is not as big as for the refPI.

The mixture of CV and plug-in is a compromise with biases lying between the ISE and the MISE minimizing methods. It will be interesting whether this leads also to a more stable performance (see below). Note that there is only a slight difference between the three versions of mixtures (not shown). Clearly, the larger the share of the respective method, the bigger their impact on the resulting estimate.

Comparison of the ISE-values

Next we compare the ISE-values of the density estimates based on the selected bandwidths. The results are given in form of boxplots plus the mean (linked filled squares) displaying the distribution of the ISEs after 250 simulation runs such that we get an idea of measures m_1 and m_2, m_4, and m_5 in one figure. In Figure 1.3 we consider the mixture of three normal distributions (model 3) and compare different sample sizes, whereas in figure 1.4 the sample size is fixed to $n = 100$ while the distribution varies.

Certainly, for all methods the ISE values increase with the complexity of the estimation problem. As expected, the classical CV-criterion shows a high variation for all cases (upper extreme values not shown for the sake of presentation), doing somewhat better for more complex densities. The one-sided and the mixed versions do considerably better, though the least variation is achieved by the MISE minimizing methods (STAB, refPI and SBG). The drawback of these three methods becomes obvious when looking at the size of its ISE-values; they are clearly smaller for the CV-based methods for $n \geq 25$. Moreover, for increasing sample size their ISE values decrease very slowly whereas for the CV-methods these values come close to the optimal achievable ISE-values. Note that in the sense of minimizing the ISE, the one-sided and the Mix(1/2) versions show the best performance. They do not vary as much as the classical CV-criterion and their mean value is almost always smaller than for the other methods, see Figure 1.4.

The stabilized procedure of Chiu (STAB) delivers - as the name suggests - a very stable estimate for the bandwidth. But in the end it is hardly more stable than the one-sided CV methods but much worse in the mean and median. We else see confirmed what we already discussed in the context of biases above. The mixture of CV and plug-in lowers the negative impacts of both versions and does surprisingly well; they deliver a more stable estimate, and gives good density estimates (looking at the ISE).

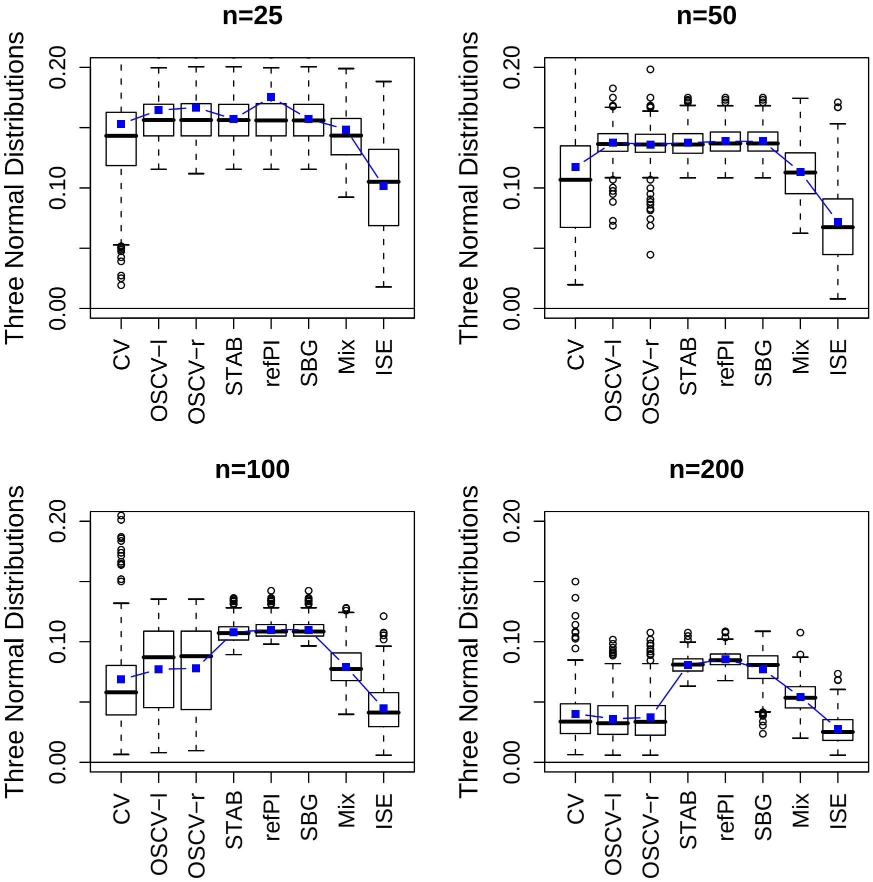


Figure 1.3: Box-plots and means (■) of the ISE-values for the mixture of three normal densities with different sample sizes

Comparison of the L1- and L2-distance of the ISE

To get an even better idea of the distance between the achieved ISE values of the selection methods and the ISE optimal (i.e. achievable) values, we have a closer look at m_5 and m_4, i.e. the L_1 and L_2 distances. In our opinion, these measures should be the most interesting for practitioners. Figure 1.5 and 1.6 show the L1-distance, and the L2-distance, respectively, for different sample sizes and models.

The pictures show that for CV, the m_5 are really big if the underlying density is not wiggly. This obviously is due to the the high variability of the selected bandwidths. Here, it does especially apply for small sample sizes (the value for $n = 25$ is even out of the range of the pictures); but for large samples like $n = 500$ the classical CV does not work at all (not shown). However, for the mixture of three normals the CV delivers almost the smallest m_5.

While both OSCV have problems with particularly small sample sizes, they easily compete with all other selectors. One may say that again, for the normal densities the OSCV methods are neither the best nor the worst methods, but always close to the best method. This corroborates our statement from above that the OSCV-criteria could be used if we do not know anything about

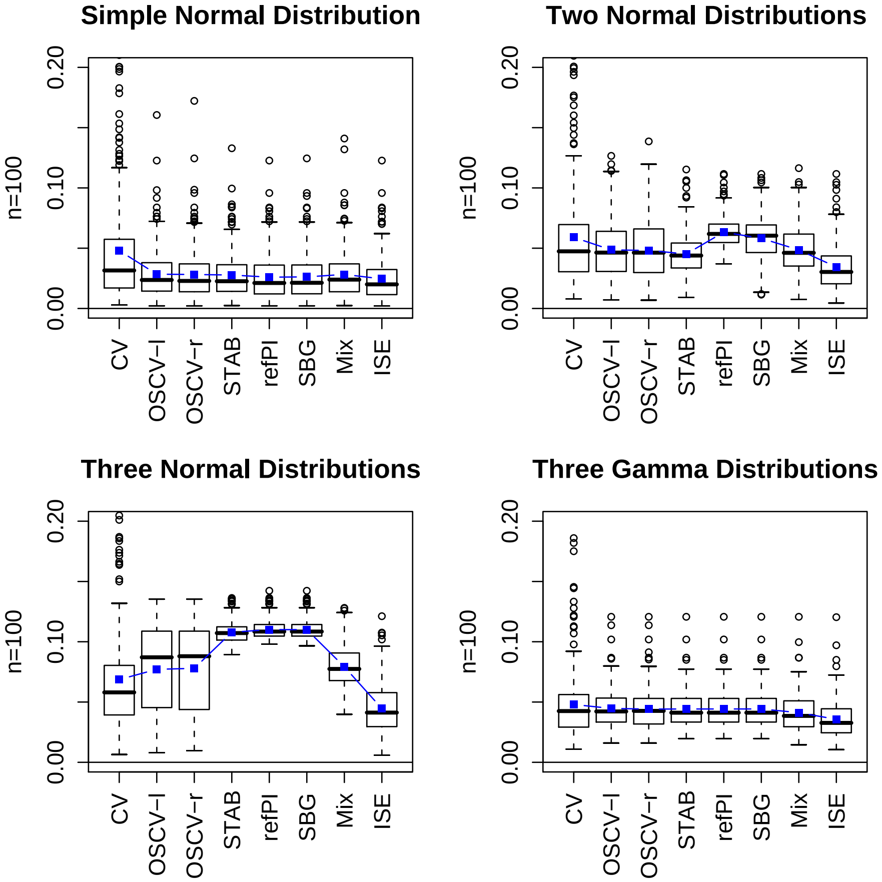


Figure 1.4: Box-plots and means (■) of the ISE-values for different distributions with sample size 100

the underlying density. Another conspicuous finding in Figure 1.5 is the difference between the two one-sided versions for the gamma distributions. Because of missing boundary correction on the left, the OSCV-l behaves very badly especially for a small sample size of $n = 25$ (out of the range) and $n = 50$. We get a similar result for $n = 25$ when looking at the L2-distances (out of the displayed range in Figure 1.6).

The three MISE minimizing methods do very well for the simple normal distribution, but else we observe a behavior for L1 and L2 which can be traced back to the fact of the prior problem described above. Even for bigger sample sizes all three methods deliver a relative big L1-distance for the mixture models. They further do not benefit as much from increasing n as other methods do. Within this MISE minimizing group, the STAB shows a better L1-distance for more complex densities. Actually, for the mixture of the three Gamma distributions we can see that the L1-distances are always very small, except for the refPI with $n = 25$ (out of the plotted range).

The mixture of CV and refined plug-in reflects the negative attributes of the CV, but for larger samples it is often in the range of the best methods. A further advantage of the mixed version is that it is much more stable than the CV or refPI when varying the sample size.

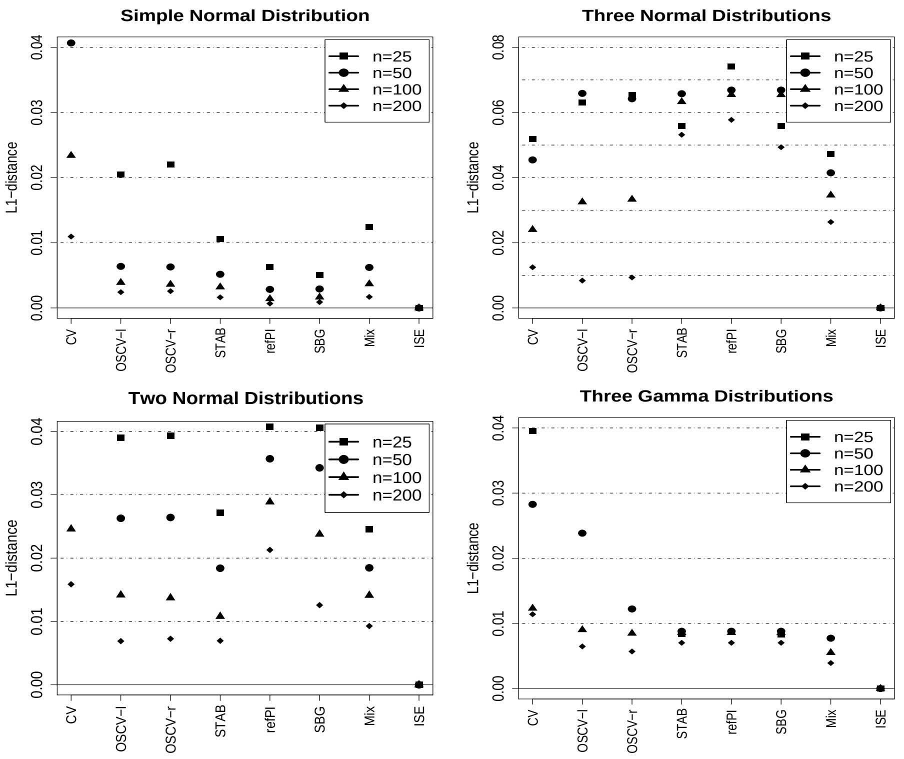


Figure 1.5: L1-distance for different sample sizes of the four underlying densities

We obtain not the same but similar results for the L2-distance given Figure 1.6. We skipped the values for $n = 25$ because they were too big for most of the methods. CV obtains very large values for small sample sizes, so that they fall out of the range of the pictures in many cases. The one-sided versions show an important improvement. The three MISE minimizing methods are excellent for the simple normal (not surprisingly) and the mixture of gammas. Among them, the STAB shows the smallest L2 distance. For sample sizes $n > 50$ the one sided CV versions outperform the others - for simple normal and gamma mixtures giving the same results as STAB but else having much smaller L2 distances. A huge difference between the left and the right one-sided occurs because of the boundary problem.

A comparison of the L1- and the L2-distance for $n = 100$ varying the distributions is shown in Figure 1.7. As can be seen in both pictures, the performance of all measures (without CV) for the simple normal distribution and the mixture of the three gamma distributions is pretty good. Also for the mixture of two normals most of the methods deliver good results, only the values for CV, refPI and the SBG become larger. For more complex densities like the mixtures of three normals, the pictures show that the MISE minimizing measures deliver worse results, because of the large biases. The most stable versions are the OSCV and the Mix(1/2). For smaller sample sizes (not shown) the pictures are quite similar, but the tendencies are strengthened and only the Mix(1/2) version delivers stable results for all distributions.

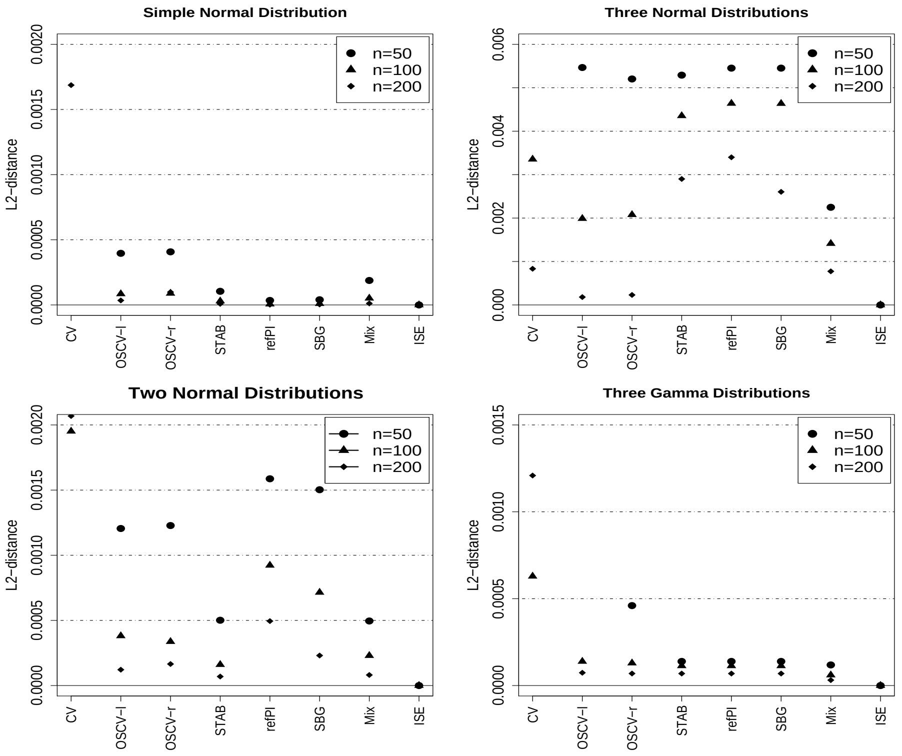


Figure 1.6: L2-distance for different sample sizes of the four underlying densities

Comparison of the mixed methods

Finally we have a closer look to the quite promising results obtained by mixing CV with refPI. We did this in different proportions as described above. In Table 1.1 and Table 1.2 we have tabulated different measures looking at the ISE, the bias of the chosen bandwidth as well as the L1- and L2-distances for the four densities. We also give the values for the infeasible ISE-minimizing bandwidths.

For design 1 (simple normal density) in Table 1.1 the Mix(1/3) is the best. This is an expected result because we know from above that the refPI works very well for this distribution. The only measure where this mix is not the best is the bias (m_3). The reason is that CV gives the smallest bias here. For design 2 (mixture of two normal densities) in Table 1.1 the Mix(2/3) wins except for the standard deviation of the ISE values (m_2) where Mix(1/3) is superior. This is explained by the very large sample variation typical for CV.

For design 3 (trimodal distribution) in Table 1.2, Mix(2/3) does best except for the standard deviation of the ISE-values (m_2). This is not surprising because above we have always stated that for more complex distributions the CV works very well while the refPI performs poorly. For the mixture of the three gammas (design 4) we observe that the values of the different measures are nearly the same, especially for the L2-distance. The main differences occur for small sample sizes. The best is Mix(2/3). As we can see from the results, sometimes Mix(2/3) is the best and sometimes Mix(1/3). The Mix(1/2) lies in between. Consequently, the main conclusion is that the mix yields very stable results and is an attractive competitor to the other

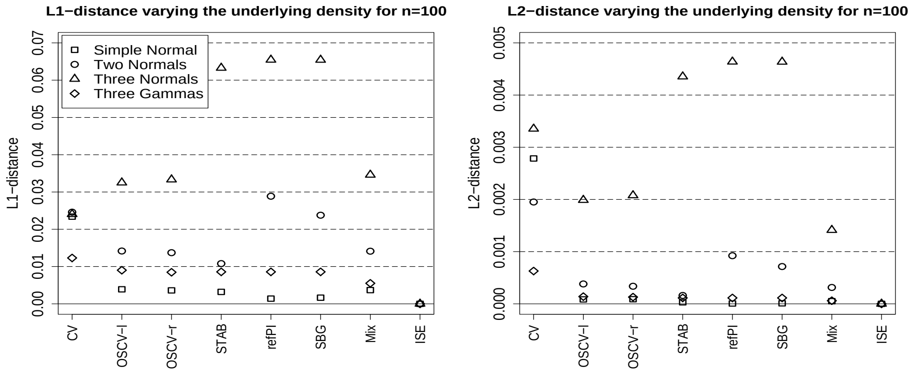


Figure 1.7: L1- and L2-distances for different underlying densities when $n = 100$

bandwidth selection methods.

1.6 Conclusions

A first finding is that it definitely makes a difference which bandwidth selector is chosen; not only in numerical terms but also for the quality of density estimation. We can identify clear differences in quality, and we can say in which situation what kind of selector is preferable.

As well known, the CV leads to a small bias but large variance. It works well for rather wiggly densities and moderate sample size. However, it neither behaves well for rather small nor for rather large samples. The quality is unfortunately dominated by its variability. A fully automatic alternative is the one sided version. In contrast to the classical CV, the OSCV methods show a behavior which is very stable. Moreover, they are maybe not uniformly the best but quite often, and never the worst.

The refPI and the SBG show a similarly stable behavior due to the fact that they are minimizing the MISE, and depend on prior information. The need of prior knowledge is the main disadvantage of these methods, and – as explained above – typically require a smooth underlying density. The worst case for these methods is when trying to estimate a trimodal normal density.

Also the STAB method is quite stable as suggested by its name. Although the full name refers to cross validation, it actually minimizes the MISE like refPI and SBG do. Consequently, it performs particularly well for the estimation of rather smooth densities but else not. It again shows the worst behavior for trimodal densities, indeed.

While the mix-methods (combining CV and plug-in) show an excellent - maybe the best - behavior, one can certainly not identify a “best mix” in advance. A further evident computational disadvantage is that we first have to apply two other methods (CV and refPI) to achieve good results.

Our conclusion is therefore that among all existing (automatic) methods for kernel density estimation, to the best of our knowledge the OSCVs seem to outperform all competitors when no (or almost no) prior knowledge is available – maybe except the one about possible boundary problems. Depending on the boundary, one would apply left- or right-hand OSCV. For moderate sample sizes however, the mixture of CV and refPI seems to be an attractive alternative until

		Design 1				Design 2			
n	Crit.	ISE	MIX(2/3)	MIX(1/3)	MIX(1/2)	ISE	MIX(2/3)	MIX(1/3)	MIX(1/2)
25	m_1	.0605	.0802	.0699	.0730	.0810	.1017	.1104	.1055
	m_2	.0571	.0646	.0604	.0610	.0426	.0447	.0342	.0390
	m_3	0	-.0112	.0048	-.0018	0	.0459	.0663	.0576
	m_4	0	.0014	4e-04	6e-04	0	.001	.0013	.0011
	m_5	0	.0197	.0094	.0124	0	.0207	.0294	.0246
50	m_1	.0374	.0471	.0420	.0436	.0561	.0706	.0793	.0745
	m_2	.0298	.0365	.0320	.0333	.0325	.0338	.0265	.0303
	m_3	0	-.0050	.0083	.0029	0	.0385	.0613	.0519
	m_4	0	4e-04	1e-04	2e-04	0	5e-04	8e-04	6e-04
	m_5	0	.0097	.0046	.0062	0	.0146	.0233	.0185
100	m_1	.0246	.0307	.0271	.0282	.0344	.0452	.0525	.0485
	m_2	.0184	.0226	.0193	.0203	.0197	.0217	.018	.0199
	m_3	0	-.0070	.0056	3e-04	0	.0359	.0578	.0484
	m_4	0	1e-04	2e-05	5e-05	0	2e-04	4e-04	3e-04
	m_5	0	.0061	.0025	.0037	0	.0108	.0181	.0141
200	m_1	.0146	.0173	.0158	.0163	.0225	.0289	.0349	.0318
	m_2	.0106	.0127	.0113	.0117	.0135	.0148	.0135	.0143
	m_3	0	-.0028	.0055	.0021	0	.0283	.0491	.0404
	m_4	0	3e-05	5e-06	1e-05	0	1e-04	2e-04	1e-04
	m_5	0	.0027	.0012	.0017	0	.0064	.0124	.0093

Table 1.1: Simple normal distribution and mixture of two normal distributions

n becomes large and CV fails completely.

References

Ahmad, I.A. and Ran, I.S. (2004). Data based bandwidth selection in kernel density estimation with parametric start via kernel contrasts, *Journal of Nonparametric Statistics* **16**(6): 841-877.

Bean, S.J. and Tsokos, C.P. (1980). Developments in nonparametric density estimation, *International Statitstical Review* **48**: 267-287.

Berlinet, A. and Devroye, L. (1994). A comparison of kernel density estimates, *Publications de l'Institut de Statistique de l'Université de Paris* **38(3)**: 3-59.

Bowman, A. (1984). An alternative method of cross-validation for the smoothing of density estimates, *Biometrika* **71**: 353-360.

Cao, R. (1993). Bootstrapping the Mean Integrated Squared Error, *Journal of multivariate analysis* **45**: 137-160.

		Design 3				Design 4			
n	Crit.	ISE	MIX(2/3)	MIX(1/3)	MIX(1/2)	ISE	MIX(2/3)	MIX(1/3)	MIX(1/2)
25	m_1	.1013	.1355	.1543	.1486	.0777	.088	.1326	.1331
	m_2	.0407	.0302	.0490	.0506	.0397	.0417	.1116	.1118
	m_3	0	.0666	.1003	.0909	0	.0163	.1987	.1922
	m_4	0	.0021	.0063	.0056	0	3e-04	.0141	.0141
	m_5	0	.0342	.0530	.0472	0	.0103	.0549	.0554
50	m_1	.0718	.1010	.1230	.1133	.0527	.0618	.0602	.0604
	m_2	.0342	.0309	.0165	.0224	.0245	.0266	.0233	.0243
	m_3	0	.0573	.0839	.0725	0	.0227	.0384	.0319
	m_4	0	.0014	.0032	.0022	0	2e-04	1e-04	1e-04
	m_5	0	.0292	.0512	.0415	0	.0091	.0075	.0077
100	m_1	.0446	.0667	.0898	.0792	.0357	.0409	.0419	.0412
	m_2	.0217	.0226	.0132	.0177	.0152	.0159	.015	.0153
	m_3	0	.0472	.075	.0632	0	.0234	.0387	.0323
	m_4	0	7e-04	.0023	.0014	0	6e-05	6e-05	6e-05
	m_5	0	.0221	.0452	.0346	0	.0052	.0062	.0055
200	m_1	.0278	.0432	.0642	.0542	.0245	.0280	.0291	.0284
	m_2	.0126	.0154	.0112	.0135	.0095	.0095	.0092	.0093
	m_3	0	.0387	.0654	.0539	0	.0204	.0365	.0297
	m_4	0	3e-04	.0014	8e-04	0	2e-05	3e-05	3e-05
	m_5	0	.0154	.0364	.0264	0	.0035	.0046	.0039

Table 1.2: Mixture of three normal respective three gamma distributions

Cao. R.; Cuevas, A. and González Manteiga, W. (1994). A comparative study of several smoothing methods in density estimation, *Computational Statistics and Data Analysis* **17**: 153-176.

Chacón, J.E.; Montanero, J. and Nogales, A.G. (2008). Bootstrap bandwidth selection using an h-dependent pilot bandwidth, *Scandinavian Journal of Statistics* **35**: 139-157.

Chaudhuri, P. and Marron, J.S. (1999). SiZer for Exploration of Structures in Curves, *Journal of the American Statistical Association* **94**(447): 807-823.

Chiu, S.T. (1991). Bandwidth Selection for Kernel Density Estimation, *The Annals of Statistics* **19**: 1883-1905.

Chiu, S.T. (1996). A comparative Review of Bandwidth Selection for Kernel Density Estimation, *Statistica Sinica* **6**: 129-145.

Devroye, L. (1989). The double kernel method in density estimation, *Ann. Inst. Henri Poincar* **25**: 533-580.

DEVROYE, L. (1997). Universal smoothing factor selection in density estimation: theory and practice (with discussion), *Test* **6**: 223-320.

DUIN, R.P.W. (1976). On the choice of smoothing parameters of Parzen estimators of probability density functions, *IEEE Transactions on Computers* **25**: 1175-1179.

FARAWAY, J.J. AND JHUN, M. (1990). Bootstrap Choice of Bandwidth for Density Estimation, *Journal of the American Statistical Association* **85/412**: 1119-1122.

FELUCH, W. AND KORONACKI, J. (1992). A note on modified cross-validation in density estimation, *Computational Statistics & Data Analysis* **13**: 143-151.

FRYER, M.J. (1977). A review of some non-parametric methods of density estimation, *Journal of Applied Mathematics* **20**(3): 335-354.

GODTLIEBSEN, F.; MARRON, J.S. AND CHAUDHURI, P. (2002). Significance in Scale Space for Bivariate Density Estimation, *Journal of Computational and Graphical Statistics* **11**: 1-21.

GRUND, B. AND POLZEHL, J. (1997). Bias corrected bootstrap bandwidth selection, *Journal of nonparametric statistics* **8**: 97-126.

HABBEMA, J.D.F.; HERMANS, J. AND VAN DEN BROEK, K. (1974). A stepwise discrimination analysis program using density estimation, in: BRUCKMAN, G. (Ed.), *COMPSTAT '74. Proceedings in Computational Statistics,* Physica, Vienna: 101-110.

HALL, P. (1990). Using the bootstrap to estimate mean square error and select smoothing parameters in nonparametric problems, *Journal of Multivariate Analysis* **32**: 177-203.

HALL, P. AND JOHNSTONE, I. (1992). Empirical Functionals and Efficient Smoothing Parameter Selection, *Journal of the Royal Statistical Socienty B* **54** (2): 475-530.

HALL, P. AND MARRON, J.S. (1987a). Extent to which Least-Squares Cross-Validation Minimises Integrated Square Error in Nonparametric Density Estimation, *Probability Theory and Related Fields* **74**: 567-581.

HALL, P. AND MARRON, J.S. (1987b). Estimation of integrated squared density derivatives, *Statistics & Probability Letters* **6**: 109-115.

HALL, P. AND MARRON, J.S. (1991). Lower bounds for bandwidth selection in density estimation, *Probability Theory and Related Fields* **90**: 149-173.

HALL, P.; MARRON, J.S. AND PARK, B.U. (1992). Smoothed cross-validation, *Probability Theory and Related Fields* **92**: 1-20.

HALL, P.; SHEATER, S.J.; JONES, M.C. AND MARRON, J.S. (1991). On optimal databased bandwidth selection in kernel density estimation, *Biometrika* **78**: 263-269.

HANNING, J. AND MARRON, J.S. (2006). Advanced Distribution Theory for SiZer, *Journal of the American Statistical Association* **101**: 484-499.

HÄRDLE, W.; MÜLLER, M.; SPERLICH, S. AND WERWATZ, A. (2004). Nonparametric and Semiparametric Models, *Springer Series in Statistics*, Berlin.

HÄRDLE, W. AND VIEU, P. (1992). Kernel regression smoothing of time series, *Journal of Time Series Analysis* **13**: 209-232.

HART, J.D. AND YI, S. (1998). One-sided cross-validation, *Journal of the American Statistical Association* **93**: 620-631.

JONES, M.C. (1991). The roles of ISE and MISE in density estimation, *Statistics & Probability Letters* **12**: 51-56.

JONES, M.C. (1998). On some kernel density estimation bandwidth selectors related to the double kernel method, *Sankhyā Ser. A* **60**: 249-264.

JONES, M.C., MARRON, J.S. AND PARK, B.U. (1991). Asimple root *n* bandwidth selector, *The annals of statistics* **19**(4): 1919-1932.

JONES, M.C., MARRON, J.S. AND SHEATHER, S.J. (1996a). A brief survey of bandwidth selection for density estimation, *Journal of the American Statistical Association* **91**: 401-407.

JONES, M.C., MARRON, J.S. AND SHEATHER, S.J. (1996b). Progress in data-based bandwidth selection for kernel density estimation, *Computational Statistics* **11**: 337-381.

JONES, M.C. AND SHEATHER, S.J. (1991). Using non-stochastic terms to advantage in kernel-based estimation of integrated squared density derivatives, *Statistics & Probability Letters* **11**: 511-514.

KIM, W.C., PARK, B.U. AND MARRON, J.S. (1994). Asymptotically best bandwidth selectors in kernel density estimation, *Statistics & Probability Letters* **19**: 119-127.

LOADER, C.R. (1999). Bandwidth selection: Classical or Plug-In?, *The annals of statistics* **27**(2): 415-438.

MARRON, J.S. (1986). Convergence properties of an empirical error criterion for multivariate density estimation, *Journal of Multivariate Analysis* **19**: 1-13.

MARRON, J.S. (1988a). Automatic Smoothing parameter selection: A survey, *Empirical Economics* **13**: 187-208.

MARRON, J.S. (1988b). Partitioned cross-validation, *Economic Reviews* **6**: 271-283.

MARRON, J.S. (1992). Bootstrap Bandwidth Selection, in: *Exploring the limits of bootstrap, eds. R. LePage and L. Billard, Wiley, New York*: 249-262.

MARRON, J.S. (1994). Visual understanding of higher order kernels, *Journal of Computational and Graphical Statistics* **3**: 447-458.

MARRON, J.S. AND NOLAN, D. (1988). Canonical kernels for density estimation, *Statistics and Probability Letters* **7(3)**: 195-199.

MARRON, J.S. AND WAND, M.P. (1992). Exact mean integrated squared errors, *Annals of statistics* **20**: 712-736.

MARTÍNEZ-MIRANDA, M.D.; NIELSEN, J. AND SPERLICH, S. (2009). *One sided Cross Validation in density estimation*, In “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.

PARK, B.U. AND MARRON, J.S. (1990). Comparison of Data-Driven Bandwidth Selectors, *Journal of the American Statistical Association* **85**: 66-72.

PARK, B.U. AND TURLACH, B.A. (1992). Practical performance of several data driven bandwidth selectors, *CORE Discussion Paper* **9205**.

RIGOLLET, P. AND TSYBAKOV, A. (2007). Linear and convex aggregation of density estimators, *Mathematical Methods of Statistics* **16**: 260-280.

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

RUPPERT, D. AND B.H. CLINE, B.H. (1994). Bias Reduction in Kernel Density Estimation by Smoothed Empirical Transformations, *The Annals of Statistics* **22**: 185-210.

SAMAROV, A. AND TSYBAKOV, A. (2007). Aggregation of density estimators and dimension reduction, In:*Advances in Statistical Modeling and Inference: essays in honor of Kjell A. Doksum*, ed. V. Nair,233-251.

SAVCHUK, O.J., HART, J.D. AND SHEATHER, S.J. (2010). Indirect Cross-Validation for Density Estimation, *Journal of the American Statistical Association* **105(489)**: 415-423.

SILVERMAN, B.W. (1986). Density estimation for Statistics and Data Analysis, Vol. 26 of *Monographs on Statistics and Applied Probability*, Chapman and Hall, London.

SCOTT, D.W. AND TERRELL, G. R. (1987). Biased and unbiased cross-validation in density estimation, *Journal of the American Statistical Association* **82**(400): 1131-1146.

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

STONE, C.J. (1984). An asymptotically optimal window selection rule for kernel density estimates, *Annals of Statistics* **12**(4): 1285-1297.

STUTE, W. (1992). Modified cross validation in density estimation, *Journal of statistical planning and Inferenz* **30**: 293-305.

TARTAR, M.E. AND KRONMAL, R.A. (1976). An introduction to the implementation and theory of nonparametric density estimation, *The American Statistician* **30**: 105-112.

TAYLOR, C.C. (1989). Bootstrap choice of the smoothing parameter in kernel density estimation,
Biometrika **76**: 705-712.

TURLACH, B.A. (1994). Bandwidth selection in Kernel density estimation: A Review, *Working Paper*.

WAND, M.P. AND JONES, M.C. (1995). Kernel Smoothing, Vol. 60 of *Monographs on Statistics and Applied Probability*, Chapman and Hall, London.

WAND, M.P.; MARRON, J.S. AND RUPPERT, D. (1991). Transformations in Density Estimation, *Journal of the American Statistical Association* **Vol. 86**, No. 414: 343-353.

WEGKAMP, M.H. (1999). Quasi universal bandwidth selection for kernel density estimators, *Canadian Journal of Statistics* **27**: 409-420.

WEGMAN, E.J.(1972). Nonparametric probability density estimation: I. A summary of available methods, *Technometrics* **14**: 533-546.

WERTZ, W. AND SCHNEIDER, B. (1979). Statistical density estimation: a bibliography, *International Statistical Review* **47**: 155-175.

YANG, Y. (2000). Mixing strategies for density estimation, *Annals of Statistics* **28(1)**: 75-87.

YANG, L. AND MARRON, S. (1999). Iterated Transformation-Kernel Density Estimation. *Journal of the American Statistical Association*, **94(446)**: 580-589.

Chapter 2

A Review and Comparison of Bandwidth Selection Methods for Kernel Regression

Abstract

Over the last four decades, several methods for selecting the smoothing parameter, generally called the bandwidth, have been introduced in kernel regression. They differ quite a bit, and although there already exist more selection methods than for any other regression smoother we can still see coming up new ones. Given the need of automatic data-driven bandwidth selectors for applied statistics, this review is intended to explain and compare these methods. This essay is based on a joint work with my colleague Max Köhler and Prof. Dr. Stefan Sperlich. The main contribution of the author of this thesis is made in the evaluation of the plug-in, bootstrap and mixed methods and the presentation and evaluation of the estimation results.

2.1 Introduction

Today, kernel regression is a common tool for empirical studies in many research areas. This is partly a consequence of the fact that nowadays kernel regression curve estimators are provided by many software packages. Even though for explorative nonparametric regression the most popular and distributed methods are based on P-spline smoothing, kernel smoothing methods are still common in econometric standard methods, for example for estimation of the scedasticity function, estimation of robust standard errors in time series and panel regression models. Still quite recently, kernel regression has experienced a kind of revival in the econometric literature on treatment effect estimation and impact evaluation, respectively. Nevertheless, until today the discussion about bandwidth selection has been going on - or at least not be closed with a clear device or suggestion for practitioners. Typically, software implementations apply some defaults which in many cases are questionable, and new contributions provide simulations limited to show that the own invention outperforms existing methods in particularly designed cases. An explicit review or comparison article can be found only about bandwidth selection for density estimation, see Heidenreich, Schindler and Sperlich (2010) and references therein.

There are many, quite different approaches dealing with the problem of bandwidth selection for kernel regression. One family of selection methods is based on the corrected ASE criterion and uses ideas from model selection to choose an optimal bandwidth. To the best of our knowledge this was first introduced by Rice (1984). A second family has become quite popular under the name of cross-validation (CV) going back to Clark (1975). A disadvantage of the CV approach is that it can easily lead to highly variable bandwidths, see Härdle, Hall and Marron (1988). A recently studied way to improve it is the one-sided cross-validation (OSCV) method proposed by Hart and Yi (1998). Alternatives to the ASE minimizing and CV approaches are the so-called plug-in methods. They look rather at the asymptotic mean integrated squared error where the unknown quantities, depending on the density of the covariate, $f(x)$, the regression function $m(x)$, and the variance (function) of the conditional response, are replaced by pre-estimates or priors, cf. for example Ruppert, Sheather and Wand (1995). Finally, there exist various bootstrap approaches but mainly focusing on the local optimal bandwidth for which reason they a fair comparison is hardly possible. Cao-Abad and González-Manteiga (1993) proposed a smoothed bootstrap, and González-Manteiga, Martínez Miranda and Pérez González (2004) a wild bootstrap procedure, both requiring a pilot bandwidth to be plugged in. As it is the case for the aforementioned plug-in methods, if we have an appropriate pilot or pre-estimator, then the performance of these methods is typically excellent, else not. Asymptotics including the rate of convergence of these methods was first studied by Hall, Marron and Park (1992).

We review a big set of existing selection methods for regression and compare them on a set of different data for which we vary the variances of the residuals, the sparseness of the design and the smoothness of the underlying curve. For different reasons we concentrate on small and moderate samples and restrict to global bandwidths. Due to the complexity of the problem we have had to be rather restrictive and decided to concentrate on designs and models which we believe are interesting (with regard to their smoothness and statistical properties rather than the specific functional form) for social and economic sciences. We are aware that neither the set of methods nor the comparison study can be comprehensive but hope it nevertheless may serve as a fair guide for applied researchers. Note that most of them cannot be found in any software package. We are probably the first who implemented all the here reviewed selection methods.

Suppose we have random pairs $(X_1,Y_1),\dots,(X_n,Y_n)$, $n \in \mathbb{N}$, where the X_i's are explanatory variables drawn from a continuous distribution with density function f. Without loss of generality, we assume $X_1 < X_2 < \dots < X_n$. The Y_i's are response variables generated by the following model:

$$Y_i = m(X_i) + \sigma(X_i)\varepsilon_i, \qquad i = 1,\dots,n, \tag{2.1}$$

with i.i.d. random variables ε_i with mean zero and unit variance. Further, $\sigma^2(x) = Var(Y|X = x)$ is finite, and the ε_i are independent of all X_j. Assume one aims to estimate $m(x) = E(Y \mid X = x)$ for an arbitrary point $x \in \mathbb{R}$.

Let $K : \mathbb{R} \to \mathbb{R}$ be a kernel function that fulfills $\int_{-\infty}^{\infty} K(u)\,\mathrm{d}u = 1$, $\int_{-\infty}^{\infty} uK(u)\,\mathrm{d}u = 0$ and $\int_{-\infty}^{\infty} u^2K(u)\,\mathrm{d}u =: \mu_2(K) < \infty$. Furthermore, denote $K_h(u) := \frac{1}{h}K(u/h)$, where $h \in \mathbb{R}^+$ is our bandwidth and or smoothing parameter. When speaking of kernel regression, there exist slightly different approaches for estimating $m(x)$. The maybe most popular ones are the Nadaraya-Watson estimator proposed by Nadaraya (1964) and Watson (1964) and the local linear estimator. Thinking of least squares estimation, the first one approximates $m(x)$ locally by a constant, whereas the latter one approximates $m(x)$ locally by a linear function. Before the local linear or more generally, the local polynomial smoother became popular, a well known alternative to the Nadaraya-Watson estimator was the so-called Gasser-Müller estimator, see Gasser and Müller (1979), which is an improved version of the kernel estimator proposed by Priestly and Chao (1972). Fan (1992) presents a list of the biases and variances of each estimator, see that paper also for more details. It is easy to see that the bias of the Nadaraya-Watson estimator is large when $|f'(x)/f(x)|$ is large, e.g. for clustered data, or when $|m'(x)|$ is large. The bias of the Gasser-Müller estimator looks simpler, does not have these drawbacks and is design-independent so that the function estimation in regions of sparse observations is improved compared to the Nadaraya-Watson estimator. On the other hand, the variance of the Gasser-Müller estimator is 1.5 times larger than that of the Nadaraya-Watson estimator. The local linear estimator has got the same variance as the Nadaraya-Watson estimator and the same bias as the Gasser-Müller estimator. When approximating $m(x)$ with higher order polynomials, a further reduction of the bias is possible but these methods require mode assumptions - and in practice also larger samples. For implementation, these methods are less attractive when facing multivariate regression, and several considered bandwidth selection methods are not made for these extensions. Most of these arguments hold also for higher order kernels. When comparing the local linear with the Gasser-Müller and the Nadaraya-Watson estimator, both theoretical approaches and simulation studies show that the local linear estimator in most cases corrects best for boundary effects, see also Fan and Gijbels (1992) or Cheng, Fan and Marron (1997). Moreover, in econometrics it is preferred to use models that nest the linear model without bias and directly provides the marginal impact and elasticities, i.e. the first derivatives. All this is provided automatically by the local linear but unfortunately not by the Nadaraya-Watson estimator. Consequently, we will concentrate in the following on the local linear estimator. More precisely, consider

$$\min_{\beta_0,\,\beta_1\in\mathbb{R}} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1(X_i - x))^2 K_h(x - X_i) \tag{2.2}$$

where the minimizer can be expressed as a weighted sum of the Y_i, i.e. $1/n\sum_{i=n}^{n} W_{h,i}(x)Y_i$. Denote $S_{h,j} = \sum_{i=1}^{n} K_h(x-X_i)(X_i-x)^j$ and consider the following two cases:

- If

$$\det\begin{pmatrix} S_{h,0}(x) & S_{h,1}(x) \\ S_{h,1}(x) & S_{h,2}(x) \end{pmatrix} = S_{h,0}(x)S_{h,2}(x) - (S_{h,1}(x))^2 \neq 0 \tag{2.3}$$

  the minimizer of (2.2) is unique and given below.

- If $S_{h,0}(x)S_{h,2}(x) - (S_{h,1}(x))^2 = 0$ we distinguish between

  ◇ $x = X_k$ for a $k \in \{1,\dots,n\}$ but X_k does not have its neighbors close to it such that $K_h(X_k - X_i) = 0$ for all $i \neq k$ such that $S_{h,1}(x_k) = S_{h,2}(x_k) = 0$. In this case, the minimizing problem (2.2) is solved by $\beta_0 = Y_k$, and β_1 can be chosen arbitrarily.

  ◇ $x \neq X_k$ for all $k \in \{1,\dots,n\}$. Then the local linear estimator is simply not defined as there are no observations close to x.

Summarizing, for our purpose we define the local linear estimator by

$$\hat{m}_h(x) = \frac{1}{n}\sum_{i=1}^{n} W_{h,i}(x)Y_i \tag{2.4}$$

with weights

$$W_{h,i}(x) = \begin{cases} \frac{nS_{h,2}(x)K_h(x-X_i) - nS_{h,1}(x)K_h(x-X_i)(X_i-x)}{S_{h,0}(x)S_{h,2}(x)-S_{h,1}(x)^2} & \text{, if } S_{h,0}(x)S_{h,2}(x) \neq S_{h,1}(x)^2, \\ n & \text{, if } S_{h,0}(x)S_{h,2}(x) = S_{h,1}(x)^2,\ x = x_i \\ 0 & \text{, else} \end{cases}$$

if $W_{h,i}(x) > 0$ for at least one i. If $W_{h,i}(x) = 0\ \forall\ i$ the local linear estimator is not defined. Note that the matrix with entrances $\{W_{h,i}(X_j)\}_{i,j}$ gives the so-called hat-matrix in kernel regression. Thanks to the very limited set of assumptions, such a nonparametric regressor is most appropriate for explorative data analysis but also for further statistical inference when model specification is crucial for the question of interest, simply because model misspecification can be reduced here to a minimum. The main drawback is, however, that if the empirical researcher has no specific idea about the smoothness of $m(x)$ but - which is commonly the case - he does not know how to choose bandwidth h. Indeed, one could say that therefore the selection of smoothing parameters is one of the fundamental model selection problems of nonparametric statistics. For practitioners this bandwidth choice is probably the main reason for not using nonparametric estimation.

To the best of our knowledge there are hardly - and no recent - reviews available comparing either theoretically or numerically the different existing bandwidth selection methods for regression. Some older studies to be mentioned are Rice (1984), Hurvich, Simonoff and Tsai

(1998), or Hart and Yi (1998). Yang and Tschernig (1999) compared two plug-in methods for multivariate regression, and more recently, González-Manteiga, Martínez Miranda and Pérez González (2004) compared a new wild bootstrap and cross validation but with a focus on local bandwidths. None of these studies compared several global bandwidth selectors for random designs. The aim was typically to introduce a new methods and compare it with a standard method.

In the next section we briefly discuss three risk measures (or say objective functions) on which bandwidth selection could and should be based on. In Section 2.3 and Section 2.4 we introduce and discuss the various selection methods we could find in the literature, separately for the three different risk measures. In Section 2.5 we present in detail extensive simulation studies to compare all here discussed selection methods. Section 2.6 concludes.

2.2 Typically used Risk Measures

We now address the problem of which bandwidth h is optimal, beginning with the question what means 'optimal'. In order to do so let us consider the well known density weighted integrated squared error (dwISE) and the mean integrated squared error (MISE), i.e. the expectation of the dwISE, of the local linear estimator:

$$\begin{aligned} MISE(\hat{m}_h(x) \mid X_1,\ldots,X_n) &= E[\, dwISE \,] = E\left[\int \{\hat{m}_h(x) - m(x)\}^2 f(x)\, dx\right] \\ &= \frac{1}{nh}||K||_2^2 \int_S \sigma^2(x) dx \\ &+ \frac{h^4}{4}\mu_2^2(K) \int_S (m''(x))^2 f(x) dx + o_P\left(\frac{1}{nh} + h^4\right), \end{aligned}$$

where $f(x)$ indicates the density of X, $||K||_2^2 = \int K(u)^2 du$, $\mu_l(K) = \int u^l K(u) du$, and f the unknown density of the explanatory variable X with the compact support $S = [a,b] \subset \mathbb{R}$. Hence, assuming homoscedasticity, the AMISE (asymptotic MISE) is given by:

$$AMISE(\hat{m}_h(x) \mid X_1,\ldots,X_n) = \frac{1}{nh}||K||_2^2 \sigma^2 (b-a) + \frac{h^4}{4}\mu_2^2(K) \int_S (m''(x))^2 f(x) dx, \tag{2.5}$$

where the first summand is the mean integrated asymptotic variance, and the second summand the asymptotic mean integrated squared bias; cf. Ruppert, Sheather, and Wand (1995). That is, we integrated squared bias and variance over the density of X, i.e. we weight the squared error by the design. Finding a reasonable bandwidth means to balance the variance and the bias part of (2.5). An obvious choice of for defining an optimal bandwidth is to say choose h such that (2.5) is minimized. Clearly, the AMISE consists mainly of unknown functions and parameters. Consequently, the selection methods' main challenge is to find appropriate substitutes or estimates. This will lead us either to the so-called plug-in methods or to bootstrap estimates of the AMISE.

For estimating a reasonable bandwidth from the data we have to find an error criterion that can be estimated in practice. Focusing on practical issues rises not only the question of how to get appropriate substitutes for the unknown functions and parameters of (2.5) but also the question of why we should look at the mean integrated squared error, i.e. a population oriented risk measure, when we just need a bandwidth for our particular sample at hand. If one does not take the expectation over the sample, i.e. considers the dwISE, one finds in the literature the so-called *ASE* (for average squared error) replacing the integration over the density of x by averaging over the sample. So this risk measure is a discrete approximation of the (density-weighted) integration of the squared deviation of our estimate from the true function. We define our *ASE* by

$$ASE(h) = \frac{1}{n}\sum_{j=1}^{n}\left(\hat{m}_h(X_j) - m(X_j)\right)^2 w(X_j), \tag{2.6}$$

where we introduced an additional trimming or weight function w to eliminate summands $(\hat{m}_h(X_j) - m(X_j))^2$ where X_j is near to the boundary. Having the explanatory variables ordered, we can simply set $w(X_j) = 1_{[X_{l+1}, X_{n-l}]}$ for a given l.By this means, we can reduce seriously the variability of the *ASE* score function, see Gasser and Müller (1979). Denote the minimizer of *ASE* by $\hat{h}_0$. Note that the *ASE* differs from the MISE in two points; first we do not integrate but average over the design, and second we do not take the expectation with respect to the estimator. If one wants to do the latter, one speaks of the *MASE* with optimal bandwidth h_0. A visual impression of what this function looks like is given in Figure 2.1. For the sake of illustration we have to anticipate here some definitions given in detail at the beginning of our simulation Section 2.5. When we refer here and in the following illustrations of this section to certain models, for details please consult Section 2.5.

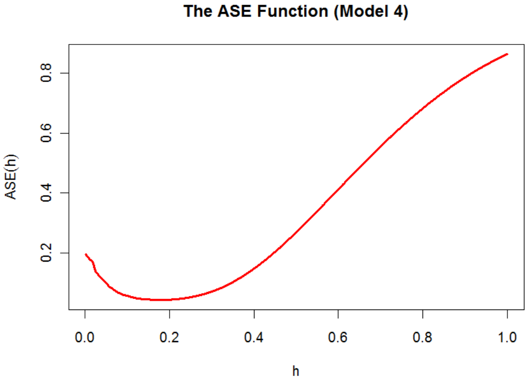


Figure 2.1: *ASE* with $w(X_j) = 1_{[X_6, X_{144}]}$ for $n = 150$ simulated data following Model 3

For now we denote a minimizer of any other score function by $\hat{h}$. Following Shibata (1981), the bandwidth selection rule is called asymptotically optimal with respect to the *ASE* risk measure, if and only if

$$\lim_{n\to\infty}\frac{ASE(\hat{h})}{ASE(\hat{h}_0)} = 1 \tag{2.7}$$

almost surely. If (2.7) is fulfilled, it follows easily that

$$\frac{ASE(\hat{h})}{ASE(\hat{h}_0)} \xrightarrow{P} 1 \tag{2.8}$$

or nearly equivalently

$$\frac{\hat{h}}{\hat{h}_0} \xrightarrow{P} 1, \tag{2.9}$$

where $\xrightarrow{P}$ stands for convergence in probability. Note that optimality can also be defined with respect to the other risk measures like $MISE$ or $MASE$.

Before we start we should emphasize that we consider the ASE risk measure as our benchmark that should be minimized. All alternative criteria are typically motivated by the fact that asymptotically they are all the same. We believe that in explorative nonparametric fitting the practitioner is interested in finding the bandwidth that minimizes the (density weighted) integrated squared error for the given data, she/he is not interested in a bandwidth that minimizes the squared error for other samples or in average over all possible samples.

2.3 Choosing the smoothing parameter based on ASE

Having said that, it is intuitively obvious that one suggests to use ASE estimates for obtaining a good estimate of the 'optimal' bandwidth h. Therefore, all score functions introduced in this section are approaches to estimate the ASE function in practice when the true function m is not known. An obvious and easy approach for estimating the ASE function is plugging into (2.6) response Y_j for $m(X_j)$. This yields the substitution estimate

$$p(h) = \frac{1}{n}\sum_{j=1}^{n}(\hat{m}_h(X_j) - Y_j)^2 w(X_j). \tag{2.10}$$

It can easily be shown, that this is a biased estimator of $ASE(h)$, see for example Härdle (1992), chapter 5. One can accept a bias that is independent of h as in this case the minimizer of (2.10) is the same as that of (2.6). Unfortunately this is not the case for $p(h)$.

We present two approaches to correct for the bias. First the corrected ASE methods that penalizes each summand of (2.10) when choosing h too small, and second the cross validation (CV) method that applies the leave one out estimator. Furthermore, we introduce the most recent one-sided cross validation (OSCV) method which is a remarkable enhancement of the classic CV.

2.3.1 The Corrected ASE

It is clear that $h \downarrow 0$ leads to interpolation, i.e. $\hat{m}_h(X_j) \to Y_j$, so that the function to be minimized, namely $p(h)$, could become arbitrarily small. On the other hand, this would surely cause a very large variance of $\hat{m}_h$ what indicates that such a criterion function would not balance bias and variance. Consequently, the corrected ASE penalizes when choosing h too small in an (at least asymptotically) reasonable sense. We define

$$G(h) = \frac{1}{n}\sum_{j=1}^{n}(Y_j - \hat{m}_h(X_j))^2\, \Xi\left(\frac{1}{n}W_{h,j}(X_j)\right) w(X_j), \tag{2.11}$$

where we use $w(X_j) = 1_{[X_{l+1},X_{n-l}]}$ to trim near the boundary. $\Xi(.)$ is a penalizing function with first-order Taylor expansion

$$\Xi(u) = 1 + 2u + O(u^2)\ ,\ u \to 0. \tag{2.12}$$

The smaller we choose bandwidth h the larger gets $W_{h,j}(X_j)$ and the penalizing factor $\Xi\left(\frac{1}{n}W_{h,j}(X_j)\right)$ increases. By conducting a first-order Taylor expansion of G and disregarding lower order terms it is easy to show that $G(h)$ is roughly equal to $ASE(h)$ up to a shift that is independent of h. The following list presents a number of proposed penalizing functions that satisfy the expansion $\Xi(u) = 1 + 2u + O(u^2)\ ,\ u \to 0$:

- Shibata's model selector $\hat{h}_S = \underset{h\in\mathbb{R}^+}{\operatorname{argmin}}\, G_S(h)$, see Shibata (1981)
$$\text{with} \qquad \Xi_S(u) = 1 + 2u\ . \tag{2.13}$$

- Generalized cross validation (GCV) $\hat{h}_{GCV} = \underset{h\in\mathbb{R}^+}{\operatorname{argmin}}\, G_{GCV}(h)$, see Craven and Wahba (1979)
$$\text{with} \qquad \Xi_{GCV}(u) = (1-u)^{-2}\ . \tag{2.14}$$

- Akaikes information criterion (AIC) $\hat{h}_{AIC} = \underset{h\in\mathbb{R}^+}{\operatorname{argmin}}\, G_{AIC}(h)$, see Akaike (1974)
$$\text{with} \qquad \Xi_{AIC}(u) = \exp(2u)\ . \tag{2.15}$$

- The finite prediction error (FPE) $\hat{h}_{FPE} = \underset{h\in\mathbb{R}^+}{\operatorname{argmin}}\, G_{FPE}(h)$, see Akaike (1970)
$$\text{with} \qquad \Xi_{FPE}(u) = \frac{1+u}{1-u}\ . \tag{2.16}$$

- Rice's T (T) $\hat{h}_T = \underset{h\in\mathbb{R}^+}{\operatorname{argmin}}\, G_T(h)$, see Rice (1984)
$$\text{with} \qquad \Xi_T(u) = (1-2u)^{-1}\ . \tag{2.17}$$

All these corrected ASE bandwidth selection rules are consistent for $n \to \infty$ and $nh \to \infty$ as $h \downarrow 0$. In practice they certainly exhibit some deficiencies. To mitigate the problems that may occur for too small bandwidths, we will fix a data-adaptive lower bound for $\hat{h}$. Notice that for $h \leq h_{min,j} := \min\left\{X_j - X_{j-1}, X_{j+1} - X_j\right\}$ (recall that the explanatory variables are ordered for the sake of presentation), we get $\frac{1}{n}W_{h,j}(X_j) = 1$ and $\frac{1}{n}W_{h,i}(X_j) = 0$ for all $i \neq j$. In this case the j'th summand of (2.11) is not defined if we choose $\Xi(.) = \Xi_{GCV}(.)$ or $\Xi(.) = \Xi_{FPE}(.)$ but is $\Xi(1)$ finite for all other penalizing functions such that the j'th summand of (2.11) gets zero. This shows that for sufficient small bandwidths h the score function $G(h)$ is either not defined or can be arbitrarily small. This does surely not solve the problem of balancing bias and variance of the local linear estimator. Therefore, we first calculate the infimum of the set of all bandwidths for which (2.11) can be evaluated,

$$h_{min,G} = \max\left\{h_{min,l+1}, \ldots, h_{min,n-l}\right\}. \tag{2.18}$$

When minimizing $G(h)$ for any of the above listed criteria, we used only the bandwidths h that fulfill $h > h_{min,G}$, all taken from the grid in (2.18).

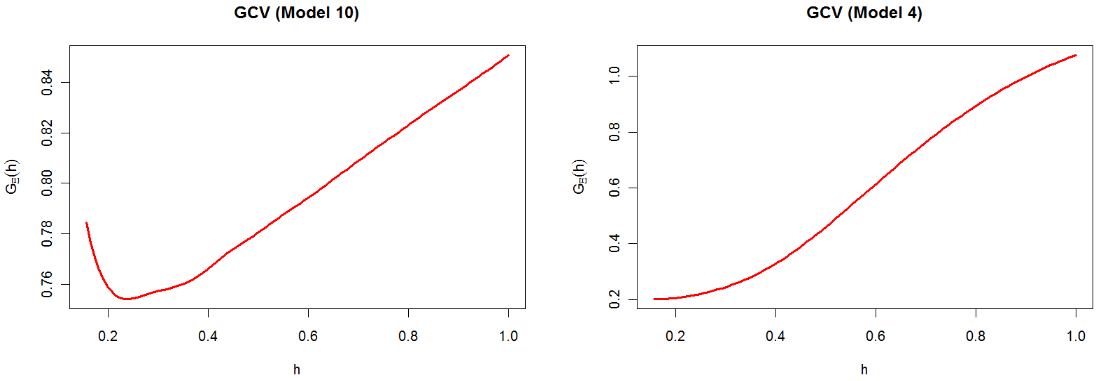


Figure 2.2: The Corrected ASE Functions for $n = 150$ independent data following Model 4 and Model 10, respectively.

Figure (2.2) shows a plot of the corrected ASE score functions when using the Rice's T penalizing function. Not surprisingly, the optimal bandwidth that is related to the simulated smooth model 10 shows a clear optimum whereas the corrected ASE function corresponding to the rather wiggly regression $m(x)$ in model 4 takes it smallest value at the fixed (see above) minimum. However, even the smooth model might cause problems depending on how the minimum is ascertained: often one has at least two local minimums. These are typical problems of the corrected ASE bandwidth selection rules that we observed for almost all penalizing function. Recall that the models used for these calculations are specified in Section 2.5.

2.3.2 The Cross-Validation

In the following we present the CV method introduced by Clark (1977). To the best of our knowledge he was the first who proposed the score function

$$CV(h) = \frac{1}{n}\sum_{j=1}^{n}(Y_j - \hat{m}_{h,-j}(X_j))^2 w(X_j)\,, \tag{2.19}$$

where $\hat{m}_{h,-j}(X_j)$ is the leave one out estimator which is simply the local linear estimator based on the data $(X_1,Y_1),\dots(X_{j-1},Y_{j-1})$, $(X_{j+1},Y_{j+1}),\dots,(X_n,Y_n)$. In analogy to the ASE function, the weights $w(\cdot)$ are used to reduce the variability of $CV(h)$. We again apply the trimming $w(X_j) = 1_{[X_{l+1},X_{n-l}]}$ to get rid of boundary effects. It can easily be shown that this score function is a biased estimator of $ASE(h)$ but the bias is independent of h. This motivates the until today most popular data-driven bandwidth selection rule:

$$\hat{h}_{CV} = \underset{h\in\mathbb{R}^+}{\operatorname{argmin}}\, CV(h)\,. \tag{2.20}$$

As for the corrected ASE bandwidth selection rules, the CV bandwidth selection rule is consistent but in practice, curiously has especially serious problems as $n\to\infty$. The reason is that this criterion hardly stabilizes for increasing n and the variance of the resulting bandwidth estimate $\hat{h}$ is often huge. Clearly, for $h < h_{min,j} := \min\{X_j - X_{j-1}, X_{j+1} - X_j\}$ we have similar problems as for the corrected ASE methods as then the local linear estimator $\hat{m}_h(X_j)$ is not defined. Therefore, (2.19) is only defined if we fix $h > h_{min,CV}$ with

$$h_{min,CV} := \max\{h_{min,l+1},\dots,h_{min,n-l}\}\,. \tag{2.21}$$

Although this mitigates the problems at the lower bound of the bandwidth scale (i.e. for bandwidth approaching zero), Figure 2.3 exhibits similar problems for the CV as we saw them for the corrected ASE criteria. Figure 2.3 shows the CV score functions when data followed model 10 and model 4. Again, for the wiggly model 4 we simply take the smallest possible bandwidth whereas for the smooth model 10 we seem to have a clear global minimum.

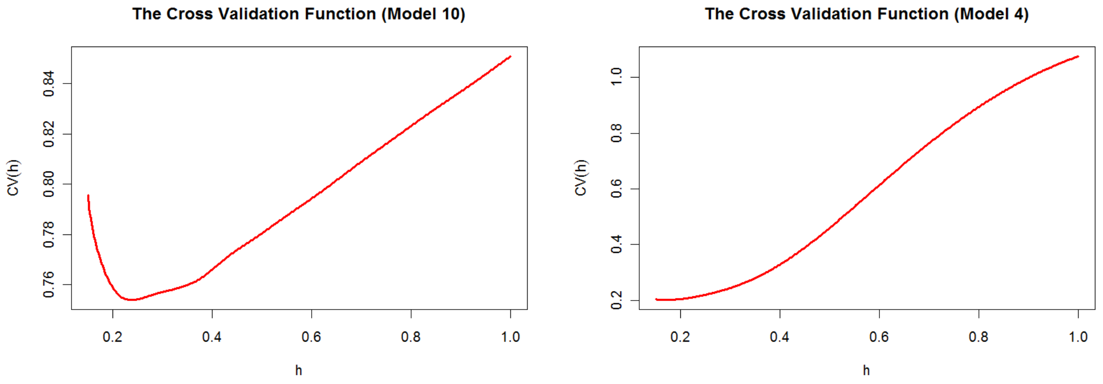


Figure 2.3: The CV functions for $n = 150$ simulated data following Model 4 and Model 10, respectively.

2.3.3 The One-Sided Cross Validation

As mentioned above the main problem of CV is the lack of stability resulting in large variances of its estimated bandwidths. As has been already noted by Marron (1986), the harder the estimation problem the better CV works. Based on this idea, Hart and Yi (1998) developed a new modification of CV.

Consider the estimator $\hat{m}_{\hat{h}_{CV}}$ with kernel K with support $[-1,1]$ that uses the CV bandwidth $\hat{h}_{CV}$. Furthermore, we consider a second estimator $\tilde{m}_b$ with smoothing parameter b based on a (selection) kernel L with support $[0,1]$. Then define

$$OSCV(b) = \frac{1}{n-2l} \sum_{i=l+1}^{n-l} (\tilde{m}_b^{-i}(X_i) - Y_i)^2, \tag{2.22}$$

where $\tilde{m}_b^{-i}(X_i)$ is the leave-one-out estimator based on kernel L. Note that l must be at least 2. This ensures that in each summand of (2.22) at least $l-1$ data points can be used. Denote the minimizer of (2.22) by $\hat{b}$. The OSCV method makes use of the fact that a transformation $h : \mathbb{R}^+ \to \mathbb{R}^+$ exists, such that $E(h(\hat{b})) \approx E(\hat{h}_{CV})$ and $Var(h(\hat{b})) < Var(\hat{h}_{CV})$. More precisely, (2.22) is an unbiased estimator of

$$\sigma^2 + E\left[\frac{1}{n-2l} \sum_{i=l+1}^{n-l} (\tilde{m}_b(X_i) - m(X_i))^2\right].$$

Therefore, minimizing (2.22) is approximately the same as minimizing

$$E\left[\frac{1}{n-2l} \sum_{i=l+1}^{n-l} (\tilde{m}_b(X_i) - m(X_i))^2\right]. \tag{2.23}$$

In almost the same manner it can be argued that minimizing $MASE(h)$ is approximately the same as minimizing $CV(h)$. We denote the minimizer of (2.23) by b_n and the $MASE(h)$ minimizer by h_n. Using the results in Fan (1992) for minimizing the MASE-expressions, dividing the minimizers and taking limits yields

$$\frac{h_n}{b_n} \to \left[\frac{||K||_2^2}{(\mu_2^2(K))^2} * \frac{(\mu_2^2(L))^2}{||L||_2^2}\right]^{1/5} =: C,$$

see Yi (2001). Note that the constant C only depends on known expressions of kernels K and L. One can therefore define the data driven bandwidth selector

$$\hat{h}_{OSCV} = C \cdot \hat{b}. \tag{2.24}$$

According to which selection kernel is used one gets different OSCV-values. A list of recommended and well studied selection kernels is given in Table 2.1, see also Figure 2.4. The transforming constants C of L_1 to L_4 are given together with the values $\mu_2^2(L_i)$ and $||L_i||_2^2$ in Table 2.2.

As for the corrected ASE and CV bandwidth selection rules, the OSCV bandwidth selection rule is consistent. Now consider the i'th summand of (2.22). Analogously to prior discussions, (2.22) is only defined if $b > b_{min,lOSCV} = \max\{X_{l+1} - X_l, \dots, X_{n-l} - X_{n-l-1}\}$, so that for

Table 2.1: Selection kernels for left OSCV.

Kernel	Formulae
One Sided Quartic	$L_1(x) = 15/8(1-x^2)^2 1_{[0,1]}$
Local Linear Epanechnikov	$L_2(x) = 12/19(8-15x)(1-x^2)1_{[0,1]}$
Local Linear Quartic	$L_3(x) = 10/27(16-35x)(1-x^2)^2 1_{[0,1]}$
opt. Kernel from Hart and Yi (1998)	$L_4(x) = (1-x^2)(6.92-23.08x+16.15x^2)1_{[0,1]}$

Table 2.2: Selection kernels for left OSCV.

Kernel	$\mu_2^2(L)$	$\|\|L\|\|_2^2$	C
L_1	0.148571	1.428571	0.8843141
L_2	-0.1157895	4.497982	0.6363232
L_3	-0.08862434	5.11357	0.5573012
L_4	-0.07692307	5.486053	0.5192593

minimizing (2.22) we consider only bandwidths $b > h_{min,CV}$. Because of

$$\begin{aligned}
h_{min,G} &= h_{min,CV} \\
&= \max\left\{h_{min,l+1}, \ldots, h_{min,m-l}\right\} \\
&= \max\left\{\min\left\{X_{l+1}-X_l, X_{l+2}-X_{l-1}\right\}, \ldots, \min\left\{X_{n-l}-X_{n-l-1}, X_{n-l+1}-X_{n-l}\right\}\right\} \\
&\geq \max\left\{X_{l+1}-X_l, \ldots, X_{n-l}-X_{n-l-1}\right\} \\
&= b_{min,lOSCV} \\
&= 1/C * h_{min,lOSCV} \\
&\geq h_{min,lOSCV}
\end{aligned}$$

this problem is much less serious for the OSCV than for the other methods. Due to the fact that $\tilde{m}_b(x)$ uses only data that are smaller than the regression point x, the variance of $\tilde{m}_b(x)$ reacts much more sensitive when decreasing b. This makes it more likely that the true minimum of (2.22) is larger than $b_{min,lOSCV}$. And indeed, in our simulations the problem of not finding the true minimum did not occur. Clearly, the OSCV score functions show a wiggly behavior when choosing b small due to a lack of data when using data only from one side. Moreover, this selection rule overweights the variance reduction. Figure (2.5) demonstrates the problem: while for Model 4 we observe a clear minimum, for Model 10 we observe that the OSCV score function does not seem to visualize a punishment when b is chosen disproportionately large. In what follows we will deal with this problem and introduce modified OS kernels.

Note that the regression estimator used at the bandwidth selection stage, namely $\tilde{m}_b(x)$ in (2.24), uses only the data X_i that are smaller than the regression point x. This explains the notion left OSCV. For implementing the right OSCV, we use the kernel $R(u) := L(-u)$. Note that this kernel has support $[-1,0]$ and therefore $\tilde{m}_b(x)$ uses only data at the right side of x. The transforming constant C in (2.24) does not change. There is evidence that the difference of

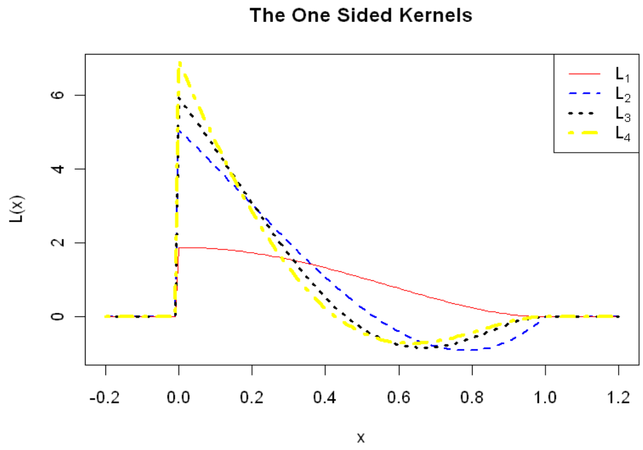


Figure 2.4: The One Sided Selection Kernels used for left OSCV.

left and right sided OSCV is negligible. Hart and Yi (1998) considered the kernel estimator proposed by Priestley and Chao (1972) in an equidistant fixed and circular design setting and argued that the OSCV score function using any left sided kernel L is the same as the OSCV score function, when using its right sided version with kernel $L(-u)$. Furthermore, they conducted simulations with a fixed design setting using the local linear estimator and argued that in all simulations they had done, a correlation of the minimizers of the left and the right OSCV score function of larger than 0.9 was observed. Thus, in the theoretical considerations we only concentrate on the left sided OSCV and assume that the corresponding right sided OSCV has the same behavior.

When implementing the OSCV method one has to choose the one sided kernel L. Hart and Yi (1998) calculated the asymptotic relative efficiency, i.e.

$$ARE(K,L) = \lim_{n\to\infty} \frac{E((\hat{h}_{OSCV} - \hat{h}_0)^2)}{E((\hat{h}_{CV} - \hat{h}_0)^2)} \tag{2.25}$$

for different kernels for L. The setting was a fixed design using the kernel estimator for estimating m. They observed an almost twenty-fold reduction in variance compared to the CV method, when simply using the right kind of kernel L. They introduced two optimal kernels. One of them is the one sided local linear kernel based on Epanechnikov that is originally used for boundary correction in density estimation (see Nielsen (1999)). For finding the optimal kernel in our case we conducted a simulation study, where we simulated 30 times the data $(X_1,Y_1),\ldots,(X_n,Y_n)$ for different data sets and different n. We compared the left OSCV methods, when using the kernels listed up in Table 2.1.

We calculated the bandwidths $(\hat{h}_0)_i$, $(\hat{h}_{CV})_i$ and $(\hat{h}_{OSCV})_i$ $(i = 1,\ldots,30)$ and then estimated $ARE(K,L)$ by

$$\widehat{ARE}(K,L) = \frac{\sum_{i=1}^{30}((\hat{h}_{OSCV})_i - (\hat{h}_0)_i)^2}{\sum_{i=1}^{30}((\hat{h}_{CV})_i - (\hat{h}_0)_i)^2}. \tag{2.26}$$

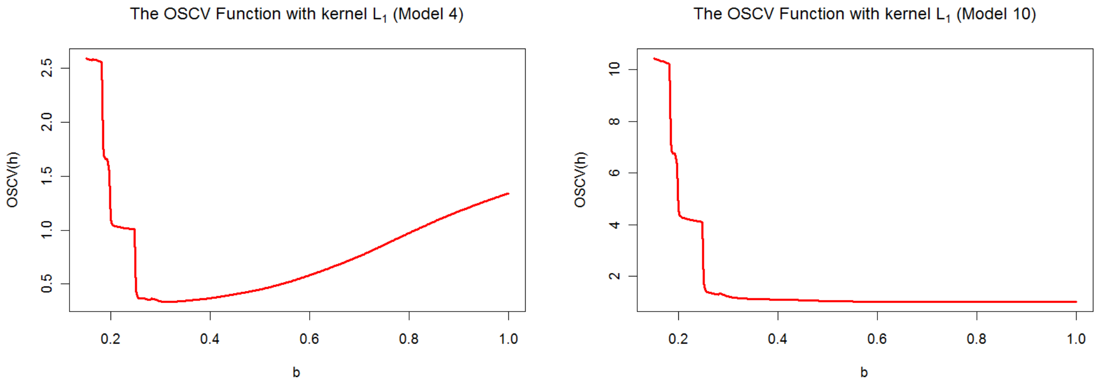


Figure 2.5: The OSCV Functions based on 150 independent data (X_i, Y_i).

The results in the case of $n = 150$ are given in Table 2.3. We observed that in seven out of the twelve different cases using the kernel L_4 is best, in only three cases L_3 is best and kernel L_1 is only best in one case. When conducting the same simulation study with $n = 50$, $n = 100$ and $n = 200$ we observed very similar results. Therefore, we decided to use kernel L_4 in the following simulation studies.

Table 2.3: The estimated $ARE(K, L_i)$ $i = 1, \ldots 4$ and $n = 150$.

Model	$\widehat{ARE}(K,L_1)$	$\widehat{ARE}(K,L_2)$	$\widehat{ARE}(K,L_3)$	$\widehat{ARE}(K,L_4)$	Best
1	5.828767	0.801370	0.915525	1.061644	L_2
2	96.290685	1.152327	19.722925	1.170663	L_2
3	6.928571	1.103896	1.032468	0.714286	L_4
4	2.051266	1.014796	1.013574	0.071266	L_4
5	1.541477	0.427530	0.427530	0.413856	L_4
6	2.025299	2.015951	1.000943	1.013723	L_3
7	2.674820	0.424460	0.250360	0.283453	L_3
8	1.519437	1.002538	0.998917	0.997350	L_4
9	3.474171	2.652201	2.651982	2.927879	L_3
10	3.945909	1.010591	1.000613	0.999650	L_4
11	47.943458	45.635282	38.257424	30.616100	L_4
12	1.484678	0.998468	0.524996	0.997636	L_3

A plot of the left OSCV Function, when using kernel L_4 is given in Figure 2.6. We observe that the OSCV functions are very wiggly when we use the kernel L_4 compared to using kernel L_1. The same wiggliness can be observed by using kernels L_2 and L_3. This behavior can also be observed when plotting the OSCV functions based on other data sets.

Even though one-sided cross validation from the left or from the right should not differ (from a theoretical point of view), in practice they do. To stabilize the behavior, Mammen, Martinez-

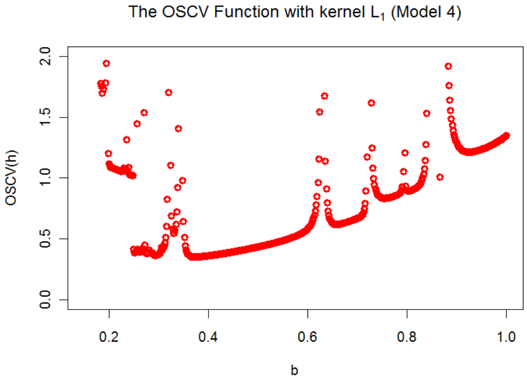


Figure 2.6: The left OSCV function using kernel L_4.

Miranda, Nielsen, and Sperlich (2011) proposed to merge them to a so-called double one-sided or simply do-validation (half from the left-sided, half from the right-sided OSCV bandwidth) for kernel density estimation and obtained amazingly good results with that procedure.

2.3.4 Notes on the Asymptotic Behavior

During the last two decades, a lot of asymptotic results for the corrected ASE methods and the CV method have been derived. Unfortunately, these asymptotic results are often only derived in the fixed and equidistant design case, when a kernel estimator or the Nadaraya-Watson estimator is considered. However, it is not hard to see that the results discussed in the following carry over to the local linear estimator which asymptotically can be considered as a Nadaraya-Watson estimator with higher order kernels.
Rice (1984) considered the kernel estimator

$$\hat{m}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) Y_i \tag{2.27}$$

proposed by Priestley and Chao (1972) in an equidistant and fixed design setting. Using Fourier-analysis, he analyzed the unbiased risk estimator of $p(h)$ introduced by Mallows (1976), and proved that its minimizer fulfills condition (2.9). He made some smoothness assumptions on K and m and considered bandwidths in the range of $H_n = \left[an^{-1/5}, bn^{-1/5}\right]$ for given a, b. Furthermore, he argued that this bandwidth selection rule is asymptotically equivalent to the corrected ASE and the CV selection rules and therefore, the minimizers of the corrected ASE functions also fulfill condition (2.9).
Härdle and Marron (1985) considered the Nadaraya-Watson estimator in a multivariate random design setting. They proved the optimality condition (2.7) for the minimizer of the CV score function with respect to the ASE, ISE and MASE risk measures for the CV method. They made the assumption of h belonging to a range of possible bandwidths that is wider

than $\left[an^{-1/5}, bn^{1/5}\right]$ so that the user of CV does not need to worry about the roughness of the underlying curve m. Further assumptions are the existence of the moments $E(Y^k|X=x)$, a Hölder continuous kernel K, i.e. $|K(u)-K(v)| \leq L||u-v||^{\xi}$ for a $\xi \in (0,1)$ and an $L>0$, $\int ||u||^{\xi}|K(u)|\,\mathrm{d}u < \infty$, the Hölder continuity of f and m and that the density f is bounded from below and compactly supported.

If conditions (2.8) and (2.9) are fulfilled for the bandwidth selection rules based on the CV and the corrected ASE score functions the question of the speed of convergence arises. Härdle and Marron (1988) considered the fixed and equidistant design case. They assumed i.i.d. errors ε_i for which all moments exist, a compactly supported kernel with Hölder continuous derivative and that the regression function has uniformly continuous integrable second derivative. Let $\hat{h}$ be any minimizer of a corrected ASE or the CV score function. Then, as $n \to \infty$,

$$n^{3/10}(\hat{h}-\hat{h}_0) \overset{\mathscr{L}}{\to} N(0,\sigma^2) \tag{2.28}$$

and

$$n^{3/10}(ASE(\hat{h})-ASE(\hat{h}_0)) \overset{\mathscr{L}}{\to} C\chi_1^2 \tag{2.29}$$

hold, where σ and C are constants depending on the kernel, the regression function and the observation error. It is interesting to observe that σ is independent of the particular penalizing function $\Xi()$ used. Taking the asymptotic rates of h's and ASE's into account, one finds that condition (2.28) is of order $n^{1/10}$ and condition (2.29) is of order $n^{1/5}$. They also show that the differences $\hat{h}_0 - h_0$ and $ASE(\hat{h}_0) - ASE(h_0)$ have the same small rates of convergence. The authors conjecture that the slow rate of convergence of $\hat{h}$ and $\hat{h}_0$ is the best possible in the minimax sense.

Chiu (1990) considered the unbiased risk minimizer using the kernel estimator in an equidistant, fixed design setting with periodic regression function (so-called circular design). He made the assumptions of independent errors ε_i for which all moments exist, some smoothness assumptions on the symmetric kernel K and m completed by technical conditions for the circular design. He only considered bandwidths belonging to a range that is slightly smaller than H_n. He pointed out that the normal distribution is not a good approximation for $\hat{h}$ because of its slow rate of convergence. Having finite samples in mind, he reasoned that

$$n^{3/10}(\hat{h}-h_0) \approx \sum_{j=1}^{\lfloor n/2 \rfloor}(V_j-2)w_K(j), \tag{2.30}$$

where $V_1, \ldots, V_{\lfloor n/2 \rfloor}$ are i.i.d. χ_2^2-distributed random variables with weights $w_K(j)$ that only depend on the kernel K. This approximation has got interesting implications. Having in mind that the $MASE$ minimizer is asymptotically the same as the ASE minimizer and that the unbiased risk minimizer is asymptotically the same as the minimizer of the corrected ASE's and the CV score functions, it follows for example

$$n^{3/10}(\hat{h}_{CV}-h_0) \approx \sum_{j=1}^{\lfloor n/2 \rfloor}(V_j-2)w_K(j). \tag{2.31}$$

When Hart and Yi (1998) computed the first twenty weights $w_K(j)$ $(j = 1, 2, \dots, 20)$ and for the quartic kernel K and $n = 100$, they observed that $w_K(1)$ and $w_K(2)$ are large and negative but $w_K(3), \dots, w_K(20)$ much smaller and mostly positive. This confirms that the distribution of $\hat{h}_{CV}$ is skewed to the left.
Assuming some further smoothness assumptions on the one sided selection kernel L and some technical conditions on L to be able to work with a circular design, they derived a similar result to (2.31) for OSCV, namely

$$n^{3/10}(\hat{h}_{OSCV} - h_0) \approx \sum_{j=1}^{\lfloor n/2 \rfloor} (V_j - 2) w_L(j). \tag{2.32}$$

When they calculated the weights $w_L(j)$ $(j = 1, 2, \dots, 20)$ in (2.32) for L_4 and $n = 100$, they observed that these were now smaller in magnitude and almost symmetric around zero, indicating a symmetric distribution of $\hat{h}_{OSCV}$ with small(er) variance.
Yi (2001) proved the asymptotic stability of the OSCV selection rule. More precisely, let b_0 be the *MASE* optimal bandwidth using selection kernel L and $\hat{b}$ be the minimizer of the unbiased risk estimator. This is asymptotically the same as the minimizer of the OSCV score function, namely $\hat{b}_{CV}$. Then, for $Cb_0 - h_0 = o_P(\hat{b} - b_0)$ with constant C,

$$\lim_{n \to \infty} E((n^{3/10}(\hat{h}_{OSCV} - h_0))^2) = C^2 V(L), \tag{2.33}$$

where $V(L)$ is a constant that only depends on the selection kernel L. As before, he considered only an equidistant fixed design case, assumed normally distributed i.i.d. errors, some smoothness for m, K and L with symmetric and compactly supported kernel K, and further technical conditions on m to be able to work with a circular design. Note that, when taking the rates of convergence of $\hat{h}_{OSCV}$ and h_0 into account, one finds, that his limit theorem (2.33) is of order $n^{1/5}$.

2.4 Choosing the smoothing parameter based on (A)MISE

In contrast to the cross-validation and corrected-ASE methods, the plug-in methods try to minimize the MISE or the AMISE. The conditional weighted AMISE of the local linear estimator $\hat{m}_h(x)$ was already given in (2.5). Minimizing w.r.t. h, leads to the AMISE-optimal bandwidth (h_{AMISE}), given by:

$$h_{AMISE} = \left(\frac{||K||_2^2 \cdot \int_S \sigma^2(x)\, dx}{\mu_2^2(K) \cdot \int_S (m''(x))^2 f(x) dx \cdot n} \right)^{1/5}, \tag{2.34}$$

where $S = [a, b] \subset \mathbb{R}$ is the support of the sample X of size n. One has the two unknown quantities, $\int_S \sigma^2(x) dx$ and $\int_S (m''(x))^2 f(x) dx$, that have to be replaced by appropriate estimates. Under homoscedasticity and using the quartic kernel, the h_{AMISE} reduces to:

$$h_{AMISE} = \left(\frac{35 \cdot \sigma^2 (b-a)}{\theta_{22} \cdot n} \right)^{1/5}, \qquad \theta_{rs} = \int_S m^{(r)}(x) m^{(s)}(x) f(x) dx, \tag{2.35}$$

where $m^{(l)}$ denotes the lth derivative of m.
The plug-in idea is to replace the unknown quantities by mainly three different strategies:

1. Rule-of-thumb bandwidth selector h_{rot}:
   The unknown quantities are replaced by parametric OLS estimators.

2. Direct-plug-in bandwidth selector h_{DPI}:
   Replace the unknown quantities by nonparametric estimates, where we need to choose 'prior (or pilot) bandwidths' for the two nonparametric estimators. In the second stage we use a parametric estimate for the calculation of these bandwidths.

3. Bootstrap based bandwidth selection h_{SB} and h_{WB}:
   The unknown expression are estimated by bootstrap methods. In case of the smoothed bootstrap (giving h_{SB}), again the unknown expressions in (2.35) are estimated, while the wild bootstrap method (h_{WB}) directly estimates the MISE of $\hat{m}_h$ and the minimizes with respect to h. Both methods require a 'prior bandwidth'.

There exist also a bandwidth selector which does not require prior bandwidths but tries to solve numerically implicit equations. This procedure follows the solve-the-equation approach in kernel density estimation, see Park and Marron (1990) or Sheather and Jones (1991). However, the results of this bandwidth selector are not uniformly better than those of the direct-plug-in approach (see Ruppert, Sheather and Wand (1995)) but require a much bigger computational effort, and are therefore quite unattractive in practice.
For the first two strategies a parametric pre-estimate in some stage is required. We have opted here for a piece-wise polynomial regression. For the sake of presentation assume the sample to be sorted in ascending order. The parametric OLS-fit is a blocked quartic fit, i.e. the sample of size n is divided in N blocks $\chi_j = \left(X_{(j-1)n/N+1}, \ldots, X_{jn/N}\right)$, $(j = 1, \ldots, N)$. For each of these blocks we fit the model:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4 + \varepsilon_i \quad i = (j-1)n/N + 1, \ldots, jn/N,$$

giving

$$\widehat{m}_{Q_j}(x) = \hat{\beta}_{0j} + \hat{\beta}_{1j} x_i + \hat{\beta}_{2j} x_i^2 + \hat{\beta}_{3j} x_i^3 + \hat{\beta}_{4j} x_i^4.$$

Then, the formula for the blocked quartic parametric estimator $\hat{\theta}_{rs}$, with $max(r,s) \leq 4$, is given by:

$$\hat{\theta}_{rs}^{Q}(N) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{N} \widehat{m}_{Q_j}^{(r)}(X_i) \widehat{m}_{Q_j}^{(s)}(X_i) \mathbf{1}_{\{X_i \in \chi_j\}}.$$

Similarly, the blocked quartic estimator for σ^2 is

$$\hat{\sigma}_{Q}^{2}(N) = \frac{1}{n-5N} \sum_{i=1}^{n} \sum_{j=1}^{N} (Y_i - \widehat{m}_{Q_j}(X_i))^2 \mathbf{1}_{\{X_i \in \chi_j\}}.$$

To choose N we follow Ruppert, Sheather, and Wand (1995), respectively Mallows (1973): take the $\widehat{N}$ from $(1,2,\dots,N_{max})$ that minimizes

$$C_p(N) = \frac{RSS(N)\cdot(n-5N_{max})}{RSS(N_{max})} - (n-10N),$$

where $RSS(N)$ is the residual sum of squares of a blocked quartic N-block-OLS, and

$$N_{max} = max\left[min(\lfloor n/20 \rfloor, N^*), 1\right],$$

with $N^* = 5$ in our simulations. Another approach to the blocked parametric fit is to use nonparametric estimators for the unknown quantities in (2.35), see Subsection 2.4.2.

2.4.1 Rule-of-thumb plug-in bandwidth selection

The idea of the rule-of-thumb bandwidth selector is to replace the unknown quantities in (2.35) directly by parametric estimates, i.e. for θ_{22} use

$$\begin{aligned}\hat{\theta}_{22}^{Q}(N) &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{N}\widehat{m}_{Q_j}^{(2)}(X_i)\widehat{m}_{Q_j}^{(2)}(X_i)\mathbf{1}_{\{X_i\in\chi_j\}} \\ &= \frac{1}{n}\sum_{i=(j-1)n/N+1}^{jn/N}\sum_{j=1}^{N}\left(2\hat{\beta}_{2j}+6\hat{\beta}_{3j}x_i+12\hat{\beta}_{4j}x_i^2\right)^2,\end{aligned}$$

and the estimator for σ^2

$$\begin{aligned}\hat{\sigma}_Q^2(N) &= \frac{1}{n-5N}\sum_{i=1}^{n}\sum_{j=1}^{N}(Y_i-\widehat{m}_{Q_j}(X_i))^2\mathbf{1}_{\{X_i\in\chi_j\}} \\ &= \frac{1}{n-5N}\sum_{i=(j-1)n/N+1}^{jn/N}\sum_{j=1}^{N}\left(y_i-\hat{\beta}_{0j}-\hat{\beta}_{1j}x_i-\hat{\beta}_{2j}x_i^2-\hat{\beta}_{3j}x_i^3-\hat{\beta}_{4j}x_i^4\right)^2 \qquad (2.36)\end{aligned}$$

The resulting rule-of-thumb bandwidth selector h_{rot} is given by

$$h_{rot} = \left(\frac{35\cdot\hat{\sigma}_Q^2(N)(b-a)}{\hat{\theta}_{22}^{Q}(N)\cdot n}\right)^{1/5},$$

which now is completely specified and feasible due to the various pre-estimates.

2.4.2 Direct plug-in bandwidth selection

In this approach the unknown quantities in (2.35) are first replaced by nonparametric estimates. Then, for the nonparametric estimator of θ_{22} a bandwidth g is needed. An obvious candidate is the bandwidth g_{AMSE} that minimizes the AMSE (asymptotic mean squared error) of the nonparametric estimator of θ_{22}. Furthermore, a prior bandwidth λ_{AMSE} has to be determined for the nonparametric estimator of σ^2. These prior bandwidths are calculated with a parametric OLS-block-fit.

A nonparametric estimator $\hat{\theta}_{22}(g_{AMSE})$ can be defined by

$$\hat{\theta}_{22}(g) = n^{-1}\sum_{i=1}^{n}\left[\hat{m}_g^{(2)}(X_i)\right]^2, \tag{2.37}$$

where we use local polynomials of order ≥ 2. As local polynomial estimates of higher derivatives can be extremely variable near the boundaries, see Gasser et al. (1991), we apply some trimming, i.e.

$$\hat{\theta}_{22}^{\alpha}(g_{AMSE}) = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{m}^{(2)}_{g_{AMSE}}(X_i)\right]^2\mathbf{1}_{\{(1-\alpha)a+\alpha b<X_i<\alpha a+(1-\alpha)b\}}, \tag{2.38}$$

here the data are truncated within $100\cdot\alpha\%$ of the boundaries of support $S=[a,b]$, for some small $\alpha\in(0,1)$. The reason for this truncation is that local polynomial kernel estimates of higher derivatives can be extremely variable near the boundaries, also recommended by Gasser et al. (1991). Since for increasing α increases the bias, α must not be too large. In our simulations we follow the proposition $\alpha = 0.05$ of Ruppert et al. (1995).

The prior bandwidth g_{AMSE}, i.e. the minimizer of the conditional asymptotic mean squared error of $\hat{\theta}_{22}(g)$ is given by

$$g_{AMSE} = \left[C_2(K)\frac{\sigma^2\cdot(b-a)}{|\theta_{24}|n}\right]^{1/7} \tag{2.39}$$

where the kernel dependent constant $C_2(K)$ for the quartic kernel is

$$C_2(K) = \begin{cases} \frac{8505}{13} & \text{if } \theta_{24}<0 \\ \frac{42525}{26} & \text{if } \theta_{24}>0 \end{cases}$$

The two unknown quantities are replaced by (block-wise) quartic parametric fits. For the prior estimation of σ^2 one uses the same as for the rule-of thumb bandwidth selector (see (2.36)). For θ_{24} we use:

$$\begin{aligned}\hat{\theta}_{24}^{Q}(\widehat{N}) &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{N}\widehat{m}^{(2)}_{Q_j}(X_i)\widehat{m}^{(4)}_{Q_j}(X_i)\mathbf{1}_{\{X_i\in\chi_j\}} \\ &= \frac{1}{n}\sum_{i=(j-1)n/N+1}^{jn/N}\sum_{j=1}^{N}\left(2\hat{\beta}_{2j}+6\hat{\beta}_{3j}x_i+12\hat{\beta}_{4j}x_i^2\right)\cdot 24\hat{\beta}_{4j}.\end{aligned}$$

This gives first an estimate for the g_{AMSE}, and afterward for θ_{22}^{α}.

The nonparametric estimator for σ^2 is:

$$\hat{\sigma}^2 = \nu^{-1}\sum_{i=1}^{n}\left[Y_i - \hat{m}_{\lambda_{AMSE}}(X_i)\right]^2, \tag{2.40}$$

where $\nu = n - 2\sum_i w_{ii} + \sum_i \sum_j w_{ij}^2$ with $\{w_{ij}\}_{i,j=1}^n$ is the hat-matrix of $\hat{m}_{\lambda_{AMSE}}$. The prior bandwidth λ_{AMSE} is calculated as the minimizer of the conditional AMSE of $\hat{\sigma}_1^2$, see Ruppert et al. (1995). Hence, λ_{AMSE} is given by

$$\hat{\lambda}_{AMSE} = \left[C_3(K) \frac{\hat{\sigma}_Q^4(\widehat{N})(b-a)}{\left(\hat{\theta}_{22}^{.05}(\hat{g}_{AMSE})\right)^2 n^2} \right]^{1/9}$$

with the kernel dependent constant $C_3(K) = \frac{146735}{14339}$.
Now, the direct-plug-in bandwidth h_{dpi} is given by:

$$h_{DPI} = \left[35 \frac{\hat{\sigma}^2(\hat{\lambda}_{AMSE})(b-a)}{\hat{\theta}_{22}^{.05}(\hat{g}_{AMSE})n} \right]^{1/5}.$$

2.4.3 Using smoothed bootstrap

The idea of is to apply bootstrap to estimate the MISE of $\hat{m}_h$ or some specific parameters of the regression or its derivatives. For a general description of this idea in nonparametric problems, see Hall (1990) or Härdle and Bowman (1988), though they only consider fixed designs. Cao-Abad and González-Manteiga (1993) discussed and theoretically analyzed several bootstrap methods for nonparametric kernel regression. They proposed the smoothed bootstrap as an alternative to wild bootstrap because the wild bootstrap mimics the model when the design is fixed. If one refers to the random design, i.e. not the ISE or ASE but MISE or MASE are of interest, hence the following resampling method is proposed: Draw bootstrap samples $(X_1^*, Y_1^*), (X_2^*, Y_2^*), \ldots, (X_n^*, Y_n^*)$ from the two-dimensional distribution estimate

$$\hat{F}_n(x,y) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\{Y_i \le y\}} \int_{-\infty}^x \mathbf{K}_g(t - X_i)\, dt,$$

where g is a prior bandwidth asymptotically larger than h, see below. Cao-Abad and González-Manteiga (1993) state that, as the marginal density of X^* is the kernel density estimate of X given the original data and bandwidth g, and the marginal distribution of Y^* is the empirical distribution function of $\{y_i\}_{i=1}^n$, one has $E^*(Y^* \mid X^* = x) = \hat{m}_g(x)$, and a natural estimator for $Var(Y|x)$ is

$$\hat{\sigma}_g^2(x) = \frac{1}{n} \sum_{i=1}^n W_{gi} Y_i^2 - [\hat{m}_g(x)]^2 = Var^*(Y^* \mid X^* = x). \tag{2.41}$$

For the estimation of $\hat{\sigma}$ assuming homoscedasticity, we average (2.41) over $x = X_i^*$. Additionally, a nonparametric estimator for θ_{22} is calculated as in formula (2.37) using cubic splines on our bootstrap sample and with the same pilot bandwidth g. With an estimate of σ^2 and θ_2^2 at hand we can use formula (2.35) to calculate a smoothed bootstrap bandwidth $\hat{h}_{SB}$ which is certainly still a function of the pilot bandwidth.

2.4.4 Using Wild Bootstrap

For early papers about the resampling plan of the wild bootstrap, see Cao-Abad (1991) or Härdle and Marron (1991). For its special use in bandwidth selection, see González-Manteiga, Martínez-Miranda and Pérez-González (2004). We will use their estimation procedure of the MSE. As we are not interested in obtaining bootstrap samples but in obtaining bootstrap estimates of the MASE, there is no need to introduce the creating of bootstrap samples. The squared bootstrap bias and the bootstrap variance can be calculated as

$$Bias^*_{h,g}(x) = \sum_{i=1}^{n} W_{hi}(x)\hat{m}_g(X_i) - \hat{m}_g(x)$$

and

$$Var^*_{h,g}(x) = \sum_{i=1}^{n} (W_{hi}(x))^2 (Y_i - \hat{m}_g(X_i))^2,$$

where g is again a pilot bandwidth that has to be chosen. For the selection of bandwidth h we are interested in the MISE or the MASE, an error criterion independent from x. For simplicity we opted for the

$$MASE(g,h) = \frac{1}{n}\sum_{i=1}^{n} MSE^*_{h,g}(X_i) \qquad (2.42)$$

with $MSE^*_{h,g}(x) = \left[Bias^*_{h,g}(x)\right]^2 + Var^*_{h,g}(x)$. To get consistent estimators, for both the wild and the smooth backfitting, the pilot bandwidth g has to be larger (in sample-size-dependent rates) than bandwidth h. Having chosen g, the MASE only depends on h so that minimizing (2.42) gives finally the optimal wild bootstrap bandwidth $\hat{h}_{WB}$. It can be easily seen, however, that the necessity of choosing a pilot (or also called prior) bandwidth, is the main disadvantage of the bootstrap methods.

2.4.5 Notes on the Asymptotic Behavior

It is clear that consistency can only be stated for the case where proper priors were used. Consequently, the rule-of-thumb estimator has no consistency properties itself, because of possible inconsistency of the there applied estimator for θ_{22}. We therefore will concentrate on results for the relative error of $\hat{h}_{DPI}$. Ruppert, Sheather, and Wand (1995) stated for the asymptotic behavior of $\hat{h}_{PDI}$

$$\frac{\hat{h}_{DPI} - h_{MISE}}{h_{MISE}} \xrightarrow{P} D, \qquad (2.43)$$

and that the method used to estimate $\hat{h}_{DPI}$, is of order $O_P(n^{-2/7})$.
Here, D is the error $\theta_{22}^{-1}\left[\frac{1}{2}\mu_4(K_{2,3})\theta_{24}G^2 + \sigma^2(b-a)||K_{2,3}||_2^2 G^{-5}\right]$ with $g = Gn^{-1/7}$ the prior bandwidth and $G > 0$ its constant. This consistency statement is based on (2.39), (2.40) with

$$\hat{\sigma}^2(\hat{\lambda}_{AMSE}) - \sigma^2 = O_P(n^{-1/2}) \ ,$$
$$\hat{\theta}_{22}(g)^{-1/5} - \theta_{22}^{-1/5} \simeq -\frac{1}{5}\theta_{22}^{-6/5}\left[\hat{\theta}_{22}(g) - \theta_{22}\right]$$

conditional on $X_1,\dots,X_n$. Both together gives

$$\frac{\hat{h}_{DPI} - h_{MISE}}{h_{MISE}} \simeq -\frac{1}{5}\theta_{22}^{-1}\left[\hat{\theta}_{22}(g) - \theta_{22}\right]$$

leading to our (2.43), see Ruppert, Sheather, and Wand (1995) for details. We know already from results of Fan (1992) and Ruppert and Wand (1994) that

$$h_{MISE} = h_{AMISE} + O_P(n^{-3/5})$$

so that one can conclude from (2.43) to consistency with respect to h_{AMISE}. The theoretical optimal prior bandwidth g is obtained by choosing G such that D equals zero – asymptotically not achievable, see Ruppert, Sheather, and Wand (1995) for further discussion.

Cao-Abad and González-Manteiga (1993) studied in detail the statistical behavior of smoothed bootstrap. For early consistency results of the wild bootsrap, see Cao-Abad (1991). The consistency of MSE estimation via wild bootstrap has been proved in González-Manteiga, Martínez-Miranda and Pérez-González (2004). The optimal prior bandwidth for the both, the smoothed and the wild bootstrap is of order $n^{-2/9}$, see for example Härdle and Marron (1991). The specific expressions however, see for example Cao-Abad and González-Manteiga (1993) or González-Manteiga, Martínez-Miranda and Pérez-González (2004), depend again on various unknown expressions so that we face similar problems as for h_{rot} and h_{PDI}.

2.4.6 A Mixture of methods

As already has been found by others, while some methods tend to over-smooth others under-smooth. In kernel density estimation it is even clear that the plug-in bandwidth and cross-validation bandwidth are negatively correlated. Heidenreich, Schindler and Sperlich (2010) studied therefore the performance of bandwidths which are simple linear combinations of a plug-in plus a cross-validation bandwidth. For kernel density estimation these bandwidths turned out to perform pretty well in all of their simulation studies.

Motivated by these positive results we will also try out such mixtures of estimated bandwidths in the context of kernel regression estimation. Like Heidenreich, Schindler and Sperlich (2010) we will only consider linear mixtures of two bandwidths. In particular, we again mix a CV bandwidth or a corrected ASE -based one with a plug-in or bootstrap method based bandwidth. Depending on the weighting factor $\alpha \in (0,1)$, the mixed methods are denoted as:

$$Mix_{method1,method2}(\alpha) = \alpha \cdot \hat{h}_{method1} + (1-\alpha)\cdot \hat{h}_{method2}, \qquad (2.44)$$

where $\hat{h}_\bullet$ denotes the optimal bandwidth to the respective method. We mix our bandwidth in the three following proportions, i.e. $\alpha = 1/2$, $\alpha = 1/3$ and $\alpha = 2/3$. As for all the others, we calculate the according ASE value for the resulting new bandwidths to assess the performance of the respective mix, see next Section.

2.5 Finite sample performance

Recall the MISE and MASE. Clearly, if $\int (f(x))^{-1} \mathrm{d}x$ is large, we expect a large integrated variance and therefore, the optimal bandwidth gives more weight on variance reduction and is therefore large. In cases of highly varying errors, i.e. a large σ^2, the same effect is observed. When the true underlying regression curve $m(\cdot)$ varies a lot, i.e. $\int (m''(x))^2 \mathrm{d}x$ is large, a large integrated squared bias is expected so that the optimal bandwidth gives more weight on bias reduction and therefore, chooses a small bandwidth. Clearly, some selection methods will do better in estimating the bias, others in estimating the variance. The same will hold for capturing the oscillation, say $m''(\cdot)$ or the handling of sparse data areas or skewed designs. As a conclusion, a fair comparison study requires a fair amount of different designs and regression functions.

For our data generating process we first have to choose the distribution of X. Then, we have to consider which are reasonable functions for $m(x)$. Finally, we have to assume a value for the variance of the error term. We generated noisy data following the models $Y_i = 1.5 \cdot \sin(k \cdot X_i) + \sigma \cdot \varepsilon_i$ with $\varepsilon \sim \mathscr{N}(0,1)$ for different k's, different σ's and a uniform design, i.e $X_i \sim U[-1,1]$, or a standard normal design, i.e. $X_i \sim N(0,1)$. We also considered the performance of the methods where $X_i \sim 1/2 \cdot \mathscr{N}(-0.6, 1/4) + 1/2 \cdot \mathscr{N}(0.3, 1/3)$. Because the results are almost identical to the uniform distribution, we do not show the results of this design in the consideration below. A list of all the models we used is given as:

Model	σ	Design	k	Model	σ	Design	k
1	1	uniform	6	7	0.5	uniform	4
2	1	normal	6	8	0.5	normal	4
3	0.5	uniform	6	9	1	uniform	2
4	0.5	normal	6	10	1	normal	2
5	1	uniform	4	11	0.5	uniform	2
6	1	normal	4	12	0.5	normal	2

Table 2.4: True Regression models

Random numbers following a normal mixture design are an example which may easily yield a large integrated asymptotic variance. Furthermore, the data are bimodal (so that two clusters are expected) and slightly skewed. Moreover, $\int (m''(x))^2 \mathrm{d}x$ becomes larger as k increases so that a larger integrated squared bias is expected as k increases. The different σ's affect the integrated variance of the local linear estimator.

The aim of this section is to compare the small sample performance of all methods discussed in the previous sections. Remember there different groups: cross-validation, corrected ASE, plug-in and bootstrap. We also compare these methods with different mixtures of the classical cross-validation (CV) criterion respectively several correcting ASE methods, with the rule-of-thumb and the direct plug-in estimate (PI1 and PI2 resp.). The mixing procedure is to include one half of the optimal bandwidth $\hat{h}_{CV}$ resp. an optimal bandwidth of a corrected ASE method

in different proportions with the optimal bandwidth of PI1 or PI2, then we assess the corresponding ASE value for the mixed bandwidth. The reason why this makes sense is that CV and corrected ASE methods tend to oversmooth while the PI methods tend to undersmooth the true $m(x)$.
All in all we present the following methods for estimation:

I cross-validation methods

1. CV: cross-validation
2. OSCV(L): one-sided cv (left)
3. OSCV(R): one-sided cv (right)
4. DoV: do-validation

II corrected ASE methods

5. Shib: Shibata's model selector
6. GCV: generalized cv
7. AIC: Akaikes information criterion
8. FPE: finite prediction error
9. Rice: Rice's T

III plug-in methods

10. PI1: rule-of-thumb plug-in
11. PI2: direct plug-in

IV bootstrap methods

12. SB: smoothed bootstrap
13. WB: wild bootstrap

V mixtures of two methods

VI ASE: infeasible ASE

There are certainly many ways how to compare the selection methods. Just when have in mind that different selectors are looking at different objective functions, it is already clear that it cannot be fair to use only one criterion. Consequently, we had to compare the performance by different performance measures, most of them based on the averaged squared error (ASE), as this is maybe the one the practitioner is mainly interested in. More specific, the considered measures are:

m_1: $mean(\hat{h}_{opt})$
mean of the selected bandwidths for the different methods

m_2: $std(\hat{h}_{opt})$
standard deviation of the selected bandwidths

m_3: $mean\left[ASE(\hat{h})\right]$
classical measure where the ASE of $\hat{m}$ is calculated (and averaged over the 500 repetitions)

m_4: $std\left[ASE(\hat{h})\right]$
volatility of the ASE's

m_5: $mean(\hat{h} - h_{ASE})$
'bias' of the bandwidth selectors, where h_{ASE} is the real ASE-minimizing bandwidth

m_6: $mean\left[(\hat{h} - h_{ASE})^2\right]$
squared L2-distance between the selected bandwidths and h_{ASE}

m_7: $mean\left[|\ \hat{h} - h_{ASE}\ |\right]$
L_1 distance between the selected bandwidths and h_{ASE}

m_8: $mean\left[ASE(\hat{h}) - ASE(h_{ASE})\right] = mean\left[|\ ASE(\hat{h}) - ASE(h_{ASE})\ |\right]$
L_1 distance of the ASE's based on selected bandwidths compared to the minimal ASE

m_9: $mean\left(\left[ASE(\hat{h}) - ASE(h_{ASE})\right]^2\right)$
squared L2-distance compared to the minimal ASE

In the following we will concentrate on the most meaningful measures, namely the bias of the bandwidths selectors (m_5), the means and standard deviations of the ASE's (m_3 and m_4), showed as box-plots, as well as the L_1-distance of the ASE's (m_8).
Without loss of generality, we used the Quartic Kernel throughout, i.e. $K(u) = \frac{15}{16}(1-u^2)^2 1_{\{|u|\leq 1\}}$. For both bootstrap procedures we tried several priors g but will present only results for the well working choice $g = 1.5 \cdot \hat{h}_{CV}$. The problems in choosing a bandwidth h which is too small already described in Section 2.3 appear by using the local linear estimator $\hat{m}_h(x)$. Hence, the correction of the bandwidth grid, given in (2.18), is done in every case where this estimator is used for calculation. All results are based on the calculations from 500 repetitions. In our simulation study we tried all methods for the sample sizes $n = 25$, $n = 50$, $n = 100$, and $n = 200$. We will first compare all methods without the mixtures. In order to summarize the different methods of choosing the optimal bandwidth, we first consider the selected bandwidths and the corresponding bias for each method separately. Afterward, we compare the methods by various measures.
Before we start with the numerical outcomes for the different methods we should briefly comment on the in practice also quite important questions of computational issues, in particular the complexity of implementation and computational costs, i.e. the time required to compute the optimal bandwidth along the considered methods. The fastest methods are the so-called corrected ASE methods. The second best in speed performance are the plug-in methods, where the rule-of-thumb plug-in is better than the direct plug-in. The fact that we only consider one-dimensional regression problems and local linear smoother allows for an implementation such that the CV methods behave also quite good but certainly worse than the plug-in. In our implementation and for the somewhat larger sample sizes (in the end, we only consider small or moderate ones) the slowest were the bootstrap based methods, in particular the smoothed bootstrap. The direct plug-in and the smoothed bootstrap method turned out to be quite complex in programming. Note that in general for more complex procedures the numerical results should be better than for the other methods to legitimate the computational effort.

2.5.1 Comparison of the bias and L1-distance for the different bandwidths (m_5, m_7)

Most of our numerical findings have been summarized in two figures: In Figure 2.7 we show the biases (m_5) and in Figure 2.8 the $L_1(h)$-distances (m_7) for all methods and models, but only for sample sizes $n = 25$ and $n = 200$.

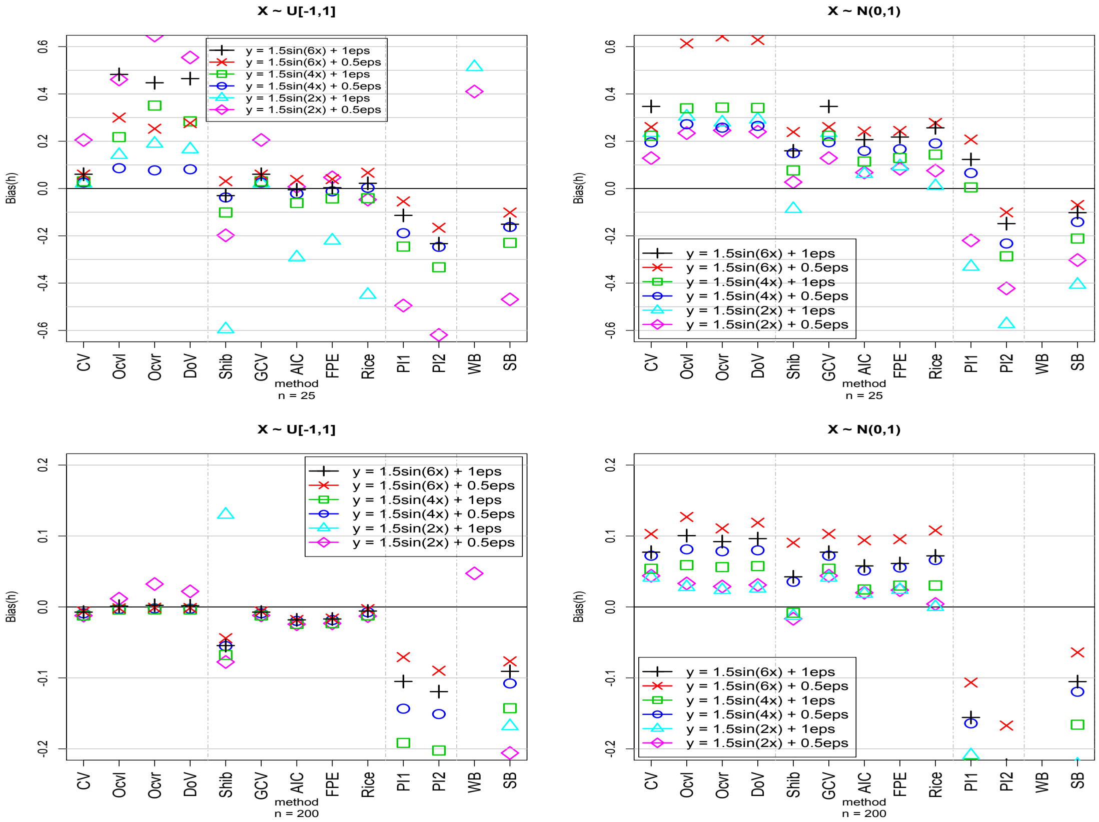


Figure 2.7: Comparison of the bias for sample sizes $n = 25$ (above) and $n = 200$ (below)

We first summarize the behavior of CV and GCV since they behave almost identically. For the standard normal distribution (see right panel in Figure 2.7), they are oversmoothing for all cases. For the uniform distribution the bias changes signs for increasing sample size, i.e. the bigger n the more tendency to undersmooth. Compared to all competitors, the L1-distances are relatively small for all models, see Figure 2.8. Because of the almost identical behavior of these two methods we will only show CV in the next subsections respectively in the pictures below.

OSCV-l, OSCV-r and DoV also oversmooth for the standard normal distribution but for larger sample sizes the behavior improves considerably and compared to the competitors. Conspicuous for the normal design is that for $n = 25$ with a high frequency of the sinus function the values of m_5 and m_7 are very high. For the uniform distribution with $n = 200$ we cannot see any clear tendency to over- respectively undersmoothing, and the L1-distance is almost zero, see also Figure 2.8. Because of the similar behavior of these three methods, and because DoV generally behaves best, we will only consider DoV in the following.

The bandwidth selection rules AIC, FPE and Rice from the second group are oversmoothing for the standard normal distribution. Only for $n = 100$, $k = 2$, and $\sigma = 1$ Rice undersmooths, and has an almost zero bias (not shown in the Figure 2.7). For the uniform design the three methods are almost always undersmoothing but in general show a good performance respective to the

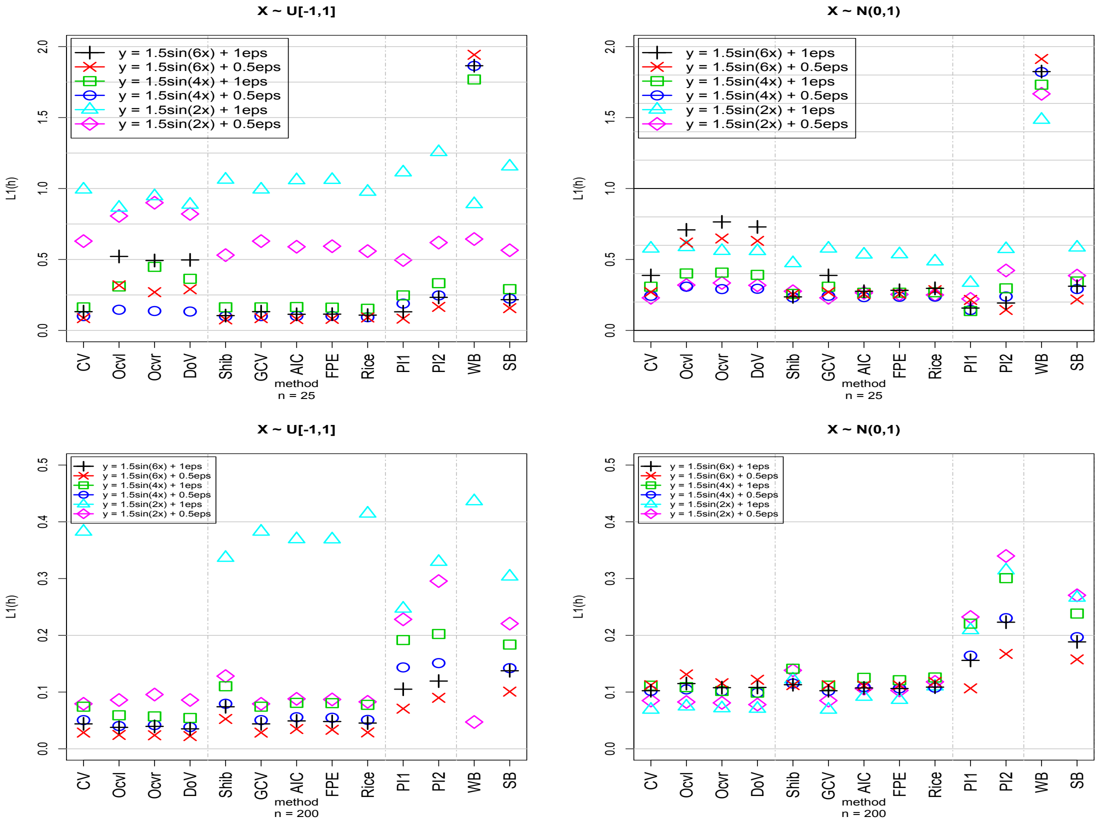


Figure 2.8: Comparison of the L1-distance for $n = 25$ (above) and $n = 200$ (below)

bias. The most noticeable for these three methods is that for $n = 25$ they behave better than CV, GCV and the one-sided CV methods, but for $n = 200$ the AIC, FPE and Rice are just as good as CV, GCV and the one-sided CV (see also Figure 2.8). In comparison AIC, FPE and Rice seem to benefit less from increasing sample sizes, i.e. although the bias respectively the L1(h)-distance is getting smaller in absolute value it is not getting smaller in the same magnitude like CV, GCV and the one-sided CV methods. In general, due to the bias, AIC, FPE and Rice show the best performance, i.e. they do not fail and are often the best. Because of the similar behavior of these three methods, and because Rice mostly behaves best, we will only consider Rice in the next sections.

The Shib selection method is almost always undersmoothing for the uniform design. For the standard normal distribution it is oversmoothing for $n = 25$ but for the bigger samples there is no clear tendency. The main difference to the other ASE corrected methods is that Shib bandwidths are worse for the uniform design, but a little bit better for the normal design.

The plug-in methods and SB are almost always undersmoothing over all designs and sample sizes. They all undersmooth with a bias which is large in absolute value. For the standard normal design, PI1 shows a good bias behavior for the smallest sample size $n = 25$ and is best for the high frequency models. In general we can state for PI1, PI2 and SB that for $n = 25$ they are as good as all the methods from group I and group II, but for increasing sample size the value of the bias and the $L_1(h)$-distance loose compared to the other selectors. Hence, in the end, PI1, PI2 and SB seem to be worse than all the methods from the first and the second group.

The remaining method to be compared is the wild bootstrap "WB". From Figure 2.7 it can be seen that the values are often out of range except for model 11 for both sample sizes and model 9 for $n = 25$. In Figure 2.8 it can be seen that WB can only keep up with the other methods for model 9 and model 11. These two models are the smoothest of all. But WB is never the best method due to the bias and is best only for two special cases if we compare the L1(h)-distances (model 9 for $n = 25$ and model 11 for $n = 200$). For the wiggly designs WB fails completely and chooses always the largest bandwidth of our bandwidth grid.

2.5.2 Comparison of L1 and L2-distances for the different bandwidths (m_6, m_7)

We will now summarize the performance of the selection methods according to the measures $L_1(h)$ and $L_2(h)$. In order to see the most important results, it is sufficient to concentrate on $k = 6$ and $\sigma = 1$ as all further results are almost identical to these with respect to the ordering of the considered methods (compare once again Figure 2.7 and Figure 2.8). All in all we provide here the comparison of the selection methods along models 1, 2, 9 and 10. In Figure 2.9 we have plotted the resulting $L_1(h)$, and in Figure 2.10 the $L_2(h)$. For each of the four models we show the values for all sample sizes, i.e. for $n = 25, 50, 100, 200$.

Considering the wild bootstrap method "WB", we notice that it is only for model 9 (the smoothest) not out of the range of our plots. But even for this model we had to use a wider plotting range, because the $L_1(h)$ respectively $L_2(h)$ values turned out to be very large for basically all methods. "WB" can only compete with the other selection methods in this case, but for $n = 100$ and $n = 200$ is even here the worst of all methods. The cross validation, say "CV", method exhibits a pretty good performance for model 1; for sample size $n = 50$ it is indeed the best. For model 2 and model 10 it shows only bad performances for $n = 25$ but good ones for the larger sample sizes. For model 9 it has an average behavior. This changes if we extend the cross validation idea to one-sided and do-validation. Indeed, for models 1, 2 and 10 "DoV" (and one-sided cross validation, where do-validation is based on) behaves badly only for $n = 25$, because of the resulting lack of information. It already behaves well for $n = 50$ and very well for not saying excellently for larger samples with $n = 100$ and $n = 200$. For model 9 its $L_1(h)$- respectively $L_2(h)$-values are even very good for $n = 25$. But for this very smooth model and sample sizes $n = 50$, $n = 100$ and $n = 200$ the plug-in PI1 is the best selection method. For model 10 PI1 is the best just for $n = 25$. Finally, "Shib" and "Rice" have an average behavior for all models and sample sizes, only for model 1 they are best for small samples with $n = 25$.

Summarizing we can say that the cross-validation methods need a sample size of at least 50 to perform well if we have a model that is not that smooth. For really smooth regression problems, the plug-in "PI1" does well.

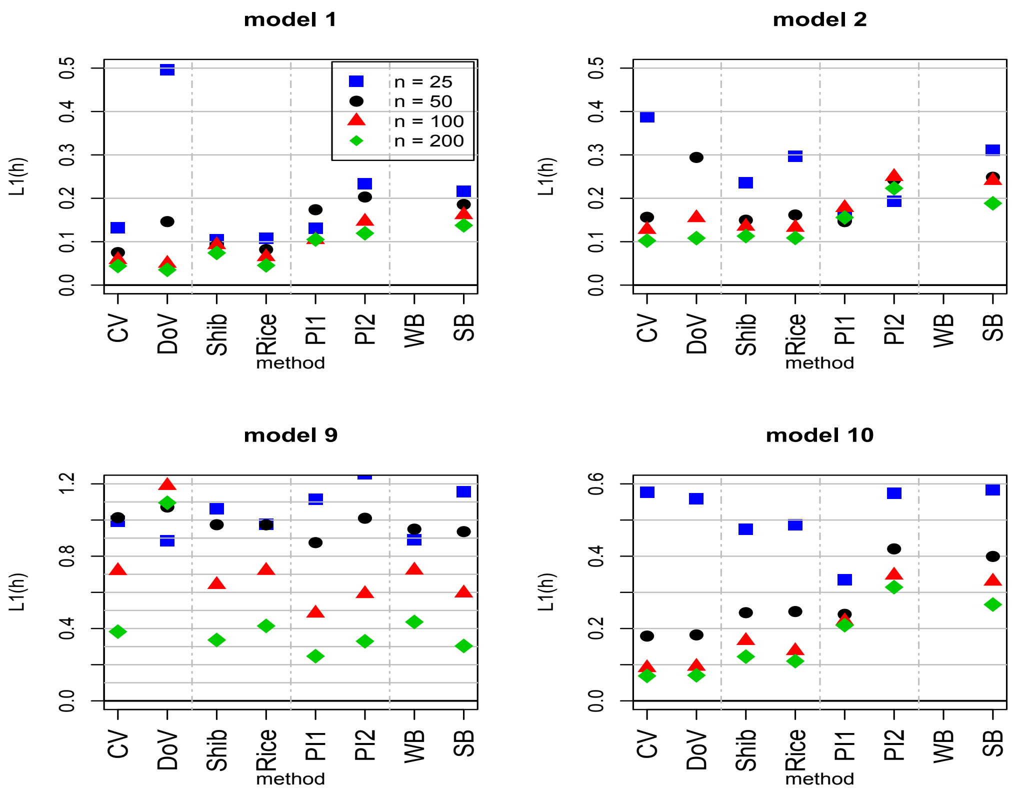


Figure 2.9: L1(h) for each four models varying the sample size

2.5.3 Comparison of the ASE-values (m_3, m_4)

In this subsection we summarize the results for the ASE-values of the different measures, i.e. the bandwidth that has been chosen for the respective method is inserted in the formula for the ASE. This is done because it enables us to compare rather the resulting regression performance than the bandwidths selected by the different methods. Needless to say, that the smallest ASE-value is reached with the benchmark, i.e. the true ASE optimal bandwidth. In our simulation we assumed twelve different models, i.e. we know the true value for $m(x)$ and the exact variance of the error term, what we do not in practice. For the same reasons we mentioned in the last subsection, the results for $k = 4$ and $\sigma = 0.5$ are skipped in the following. Hence, we compare only the boxplots of the selection methods along our models 1, 2, 9 and 10.

The main conclusions from the ASE-distributions can be summarized as follows. Varying the sample size, we can see from the boxplots, that for both designs, i.e. uniform design (see figure 2.11) and standard normal design (see figure 2.12), the means and median values for CV, DoV, Shib and Rice decrease with increasing sample sizes and decreasing frequencies. With respect to the inter quartile range (IQR henceforth) and the standard deviations it is almost the same with two exceptions. The first one is the IQR of DoV for model 9 and $n = 100$ is smaller than for $n = 200$, but there are more outliers for $n = 100$. The second one is Shib where the IQR

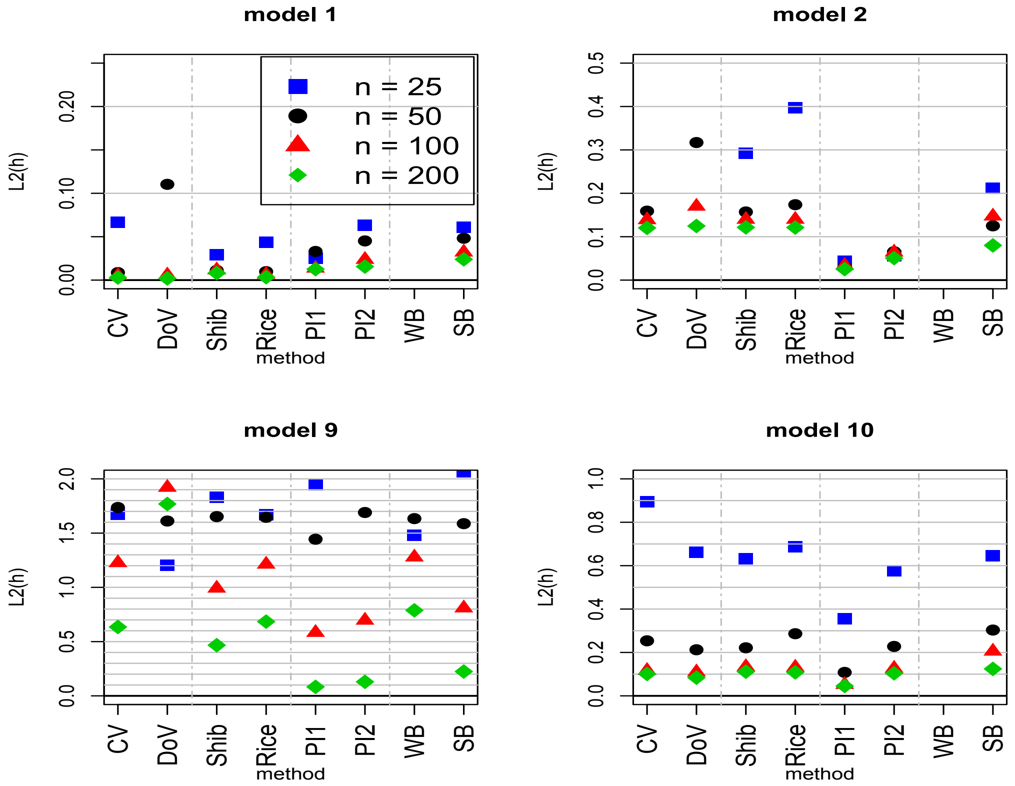


Figure 2.10: L2(h) for each four models varying the sample size

increases with decreasing frequency in the uniform design for $n = 25$, $n = 50$ and $n = 100$.

For the plug-in and the bootstrap methods the results look quite messy. With respect to the IQR and the standard deviations, WB and PI1 clearly improve with increasing sample size. For PI2 it is the same for model 1, 2 and 9, but for model 10 it is the other way round. For SB the IQR and the standard deviation are getting larger with increasing sample size.

Now, we compare the methods for model 1 (see Figure 2.11, first row). DoV benefits most from increasing sample size, i.e. for $n = 25$ DoV is worst of group I, group II and PI1, but for $n = 200$ DoV is the overall best. CV and Rice behave very similar, and they are the best selectors for $n = 25$, and 2nd best for $n = 200$. Shib shows a good behavior for smaller sample sizes, but for $n = 100$ and $n = 200$ it has the largest IQR of group I and group II. In general, the plug-in methods behave worse than groups I and II, and only a little bit better than group IV.

The most noticeable of model 9 is that WB is the overall best method, there PI2 and SB behave worst. That is because model 9 is the smoothest model, i.e. a large bandwidth is optimal in this case. For $n = 25$ and $n = 50$ DoV is the best of I, II, and III, but for larger sample sizes CV and Rice are doing better.

The results for model 2, the most wiggly design, can be seen in figure 2.12, first row. The most interesting changes, compared to model 1, occur in the first four methods. There we have more

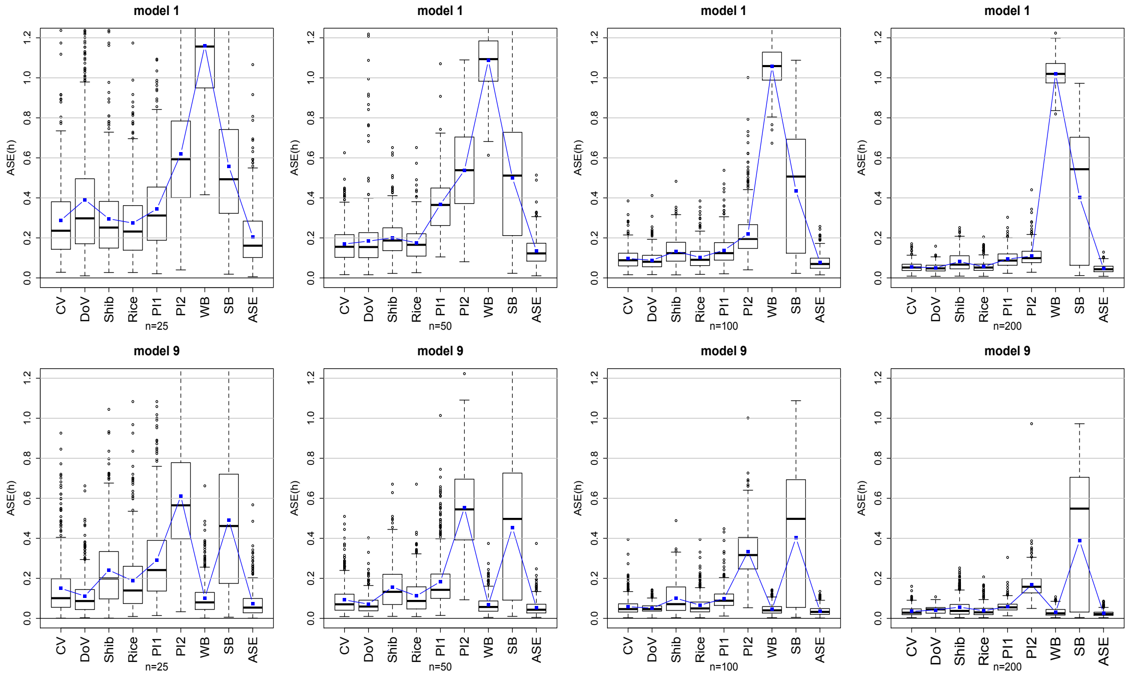


Figure 2.11: ASE-values for $X \sim U[-1,1]$ for all sample sizes

extreme outliers the bigger the sample size is. The reason for that is that these methods have problems with outliers in the covariate X. Therefore, these outliers appear, if there is a random sample having a big proportion of observations around zero but thin tails. The behavior of the methods from group I and II is very similar, i.e. the chosen method does not have a big effect on the results. Further outcomes are similar respectively identical to model 1.

Finally, we consider the results for model 10 (see figure 2.12, second row). We state the differences to model 2 (for both $X \sim N(0,1)$) and model 9 (for both $k = 2$). In contrast to model 2, the extremity of outliers does only increase a little bit with increasing sample size which is due to the fact that the model is smoother. The difference to model 9 is that WB is not the best method for model 10. This is maybe due to the fact that model 10 is more wiggly than model 9. But for both model 9 and model 10 selector WB does not fail completely in contrast to model 1 and model 2. For WB we can therefore state that if m is smooth enough this method can be used to estimate the bandwidth.

2.5.4 Comparison of the L_1 and L_2-distances of the ASE values (m_8, m_9)

If we look at Figures 3.10 and 2.14, we can conclude that there is nothing new with respect to the comparison of the considered bandwidth selection methods. One interesting fact should be mentioned: the L_1-distances do generally not decrease with increasing sample size. In model 2 the L_1-distances increase with increasing sample size for the plug-in and bootstrap methods. In model 2 all L_1 and L_2-distances for WB are out of range. For this model PI1 is the best method

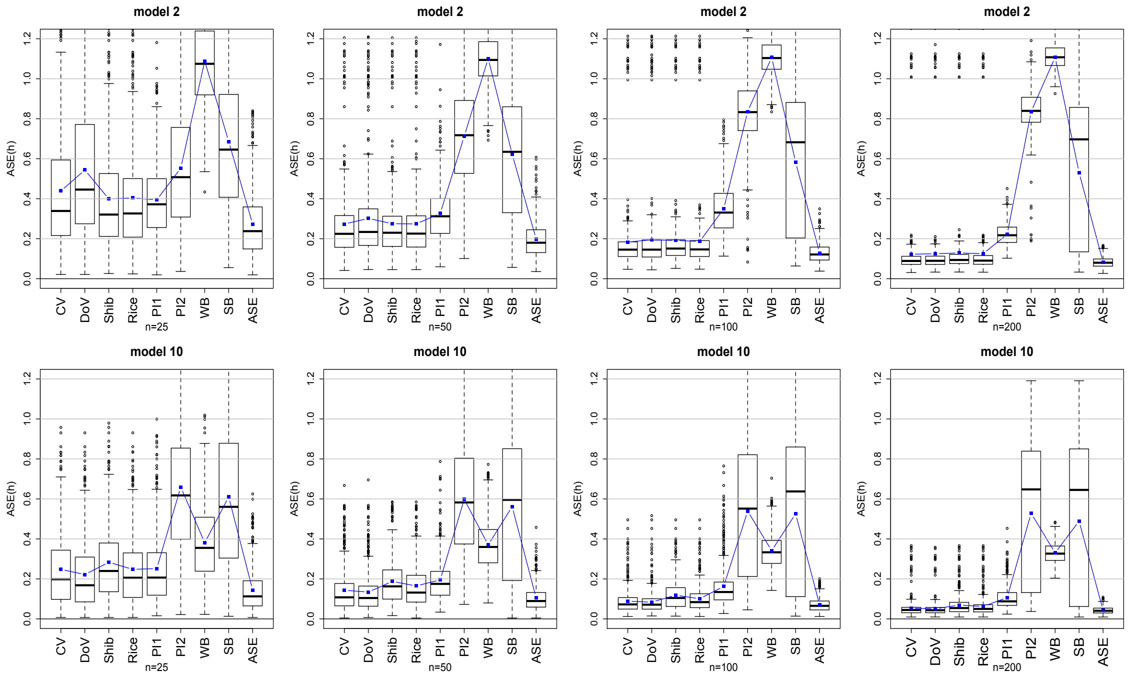


Figure 2.12: ASE-values for $X \sim N(0,1)$ for all sample sizes

for $n = 25$ but for all other sample sizes the CV and ASE-corrected methods behave better. PI2, WB and SB behave worse than the CV and ASE-corrected methods for all sample sizes.

One interesting fact for the CV and ASE-corrected methods is that there is a gap between $n = 25$ and the other sample sizes. That means, if we have a normal design respectively a more wiggly model (see model 1) combined with an extreme small sample size, PI1 will be a good method in bandwidth estimation. Another mentionable fact is that for model 9, the smoothest model, WB is the best method when looking at the L_1 and L_2 ASE values, see Figures 3.10, 2.14. For model 10 WB is good, but not better than CV or corrected ASE based methods. That means that the decision of using WB depends more on the smoothness of m than on the smoothness of the distribution of X.

We mentioned in the beginning of Section 2.5 that PI2 and SB are more complicated to implement, and especially SB has a notable computation time. If we look at all the results we can say that PI2 and SB behave badly due to all the performance measures. Hence, there is no reason for using these two methods for bandwidth estimation for the considered models.

2.5.5 Comparison of different mixtures

Finally we tried to mix two methods in order to get better results than with only one method. We tried to mix a method that tends to oversmooth with a method that tends to undersmooth the data. An obvious candidate is to mix the optimal bandwidth of the classical cross-validation (CV) respectively of a correcting ASE methods with one of the plug-in or a bootstrap optimal

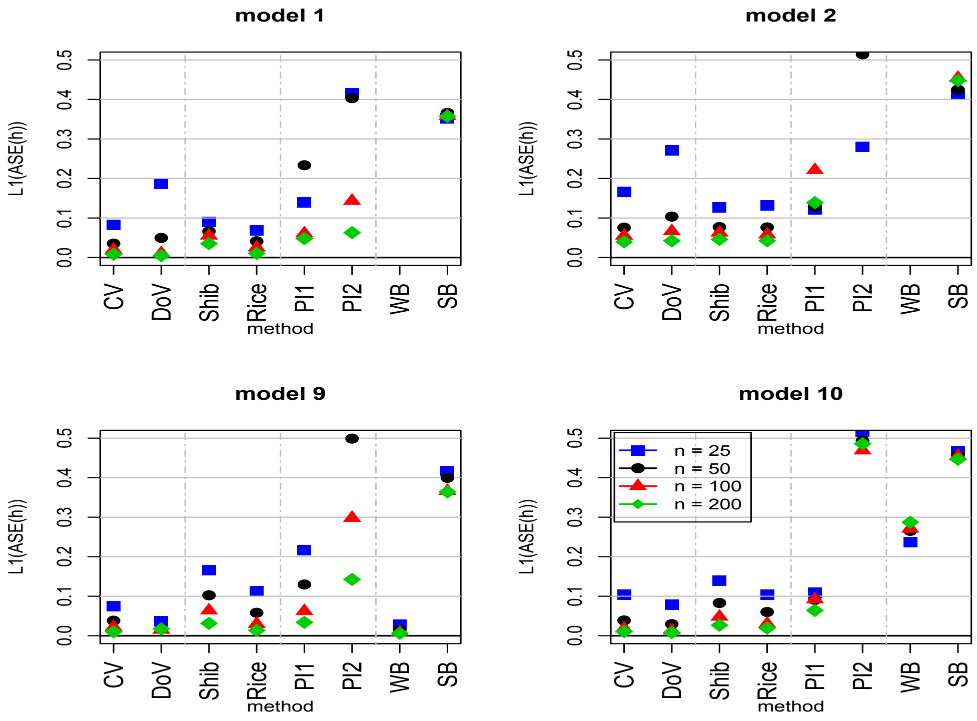


Figure 2.13: $L1(ASE)$ for each four models varying the sample size

bandwidth. Recall that CV and corrected ASE methods tend to oversmooth while the PI and bootstrap methods tend to undersmooth. The mixtures will be compared with DoV which in the end is also a mixture, namely the left- and the right-sided OSCV method, respectively.
Depending on the weighting factor $\alpha \in (0,1)$, the mixed methods are denoted as in formula (2.44) by $Mix_{method1,method2}(\alpha)$. We only try to mix methods having a good performance. We also considered other mixtures, but the best results are obtained by mixing CV and Rice with PI1. Hence, the results we present here are:

1 m11: $Mix_{CV,PI1}(1/2)$

2 m12: $Mix_{CV,PI1}(2/3)$

3 m13: $Mix_{CV,PI1}(1/3)$

4 m21: $Mix_{Rice,PI1}(1/2)$

5 m22: $Mix_{Rice,PI1}(2/3)$

6 m23: $Mix_{Rice,PI1}(1/3)$

In fact, we did simulation for basically all two-folded mixtures but skip the presentation of all the other methods for the sake of brevity and because they simply behave worse. Specifically, we decided to show the following six different mixtures: three CV-PI1, and three Rice-PI1 mixtures.
In the Figures 2.15 and 2.16 we added DoV for obvious reasons mentioned above and because this method exhibited a pretty good performance before. The bias behavior of PI1 is almost always worst, the only exception is model 2 with a sample size of 25, where CV and DoV have the biggest bias. As already mentioned, the aim to mix methods was, to get better results than

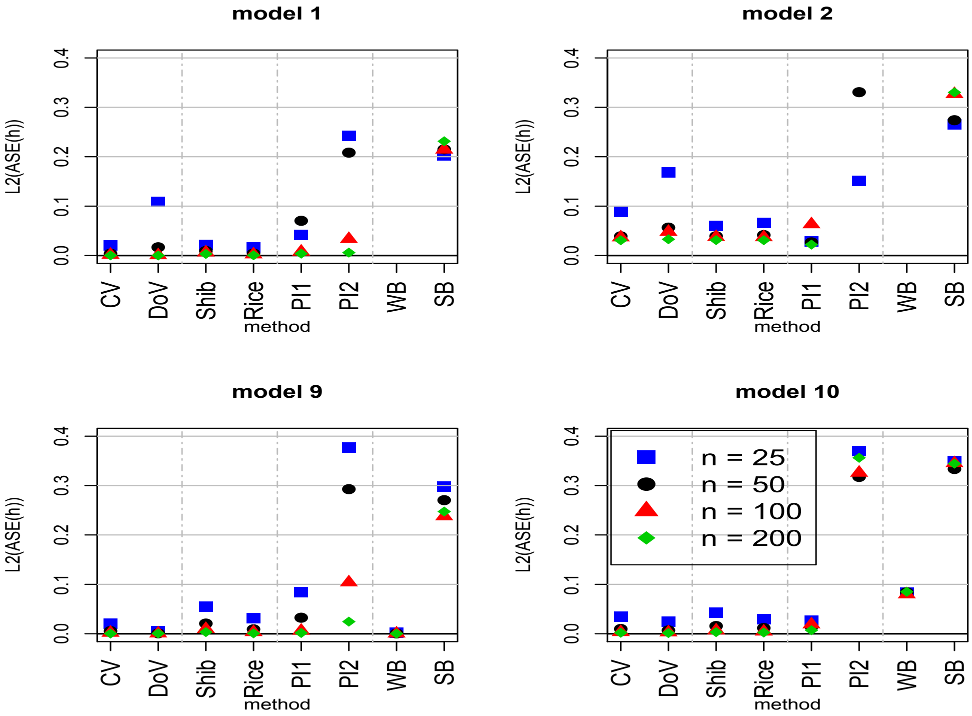


Figure 2.14: $L2(ASE)$ for each four models varying the sample size

with one single method. But, we see, that the bias values of the mixtures are indeed better than for PI1 but worse than for CV or Rice. Only for model 2, the most wiggly model, we can achieve the objective of improvement. For the L1 values we get similar results, see Figure 2.16. In conclusion we can say, that the additional effort of mixing different methods seems not to be justifiable.

2.6 Conclusions

The problem of bandwidth choice is basically as old as nonparametric estimation is. While in the meantime kernel smoothing and regression as been becoming a standard tool for explorative empirical research, and can be found in any statistical and econometric software package, the bandwidth selection can still be considered as an unsolved problem - at least for practitioners. Quite recently, Heidenreich, Schindler and Sperlich (2010) revised and compared more than thirty bandwidth selection methods for kernel density estimation. Although they could not really identify one method that performs uniformly better than all alternatives, their findings give clear guidelines at least for a certain class of densities like we typically expect and find them in social and econometric sciences.

This article is trying to offer a similar revision, comparison and guidelines for kernel regres-

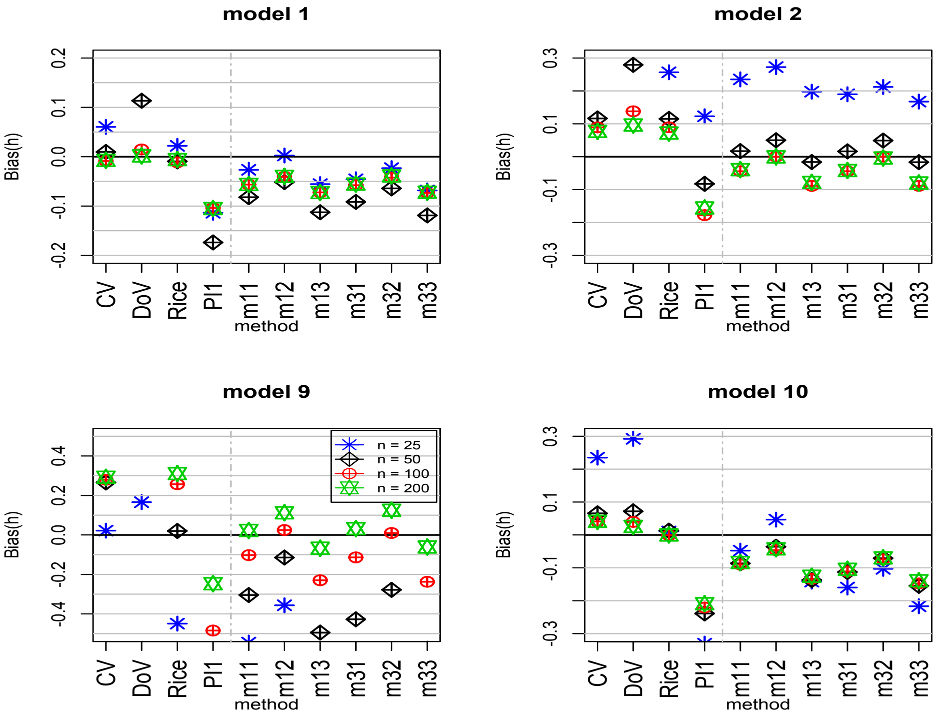


Figure 2.15: bias(h) for different mixtures

sion. Though it is true that especially for large and huge data sets, today spline regression, and in particular P-spline estimation is much more common than is the use of kernel regression, the latter is still a preferred tool for many econometric methods. Moreover, it has been experienced a kind of revival in the fairway of treatment and propensity score estimation, smoothed likelihood methods and small area statistics (in the latter as a competitor to spline methods for reasons of interpretation).

To the best of our knowledge we are the first providing such a comprehensive review and comparison study for bandwidth selection methods in the kernel regression context. We have discussed, implemented and compared almost twenty selectors, completed by again almost 20 linear combinations of two seemingly negatively correlated (with respect to signs of the bandwidth bias) selectors of which the six best have been shown here. For different reasons discussed in the introduction we concentrated our study on local linear kernel estimation.

We started with a review of the idea and definition of the methods, its asymptotics, implementation and computational issues. Probably the most interesting results are summarized in the last section, i.e. Section 2.5. We could see which methods behave quite similar and found a certain ranking of methods although – like in Heidenreich, Schindler and Sperlich (2010) – no bandwidth selector performed uniformly best. Different to their study on density estimation, for regression the mixtures of methods could not really improve compared to the single use of a selector, except the so-called do-validation. This even turned out to be maybe even the best

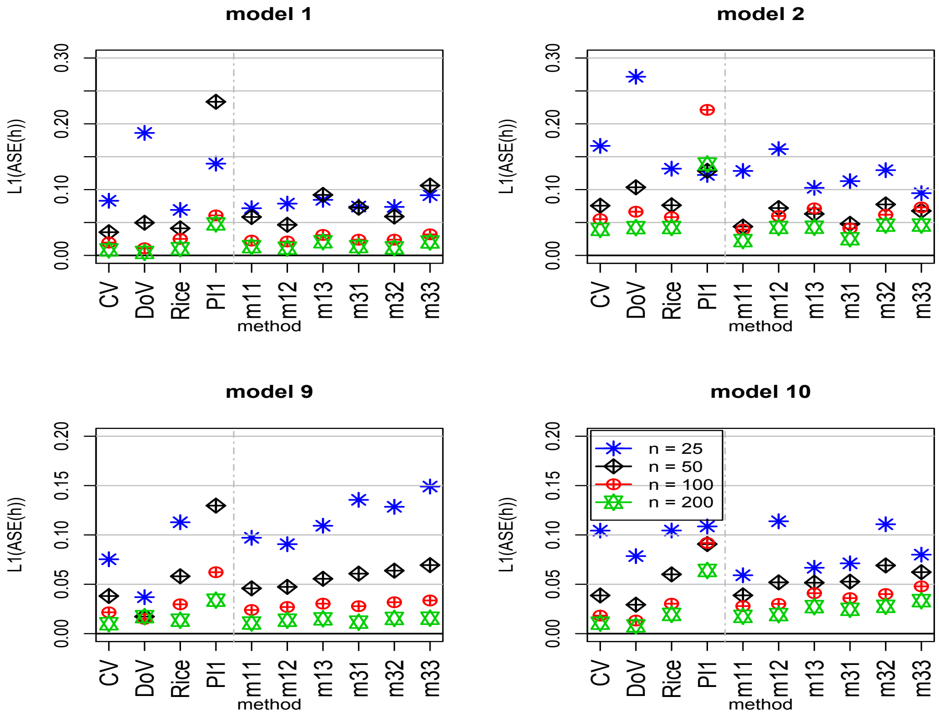


Figure 2.16: L1(ASE) for different mixtures

performing method though it is not alway easy to implement nor computationally very fast. For the rather small data sets considered, also the classical cross validation still performs well but should be replaced by generalized cross validation for increasing sample size. Note that for our context and estimator, CV and GCV behaved almost equivalently for the considered sample sizes. Nonetheless, already here and although we had rather wiggly as well as rather smooth functions under consideration, OSCV and especially DoV outperformed the classical CV. So it did for almost all models and sample sizes also compared to the other methods, at least when looking at the distribution of ASE, see Subsection 2.5.4. In our opinion, for the practitioner this is the most important measure. It should be mentioned that in the reduced set of selectors, the method proposed by Rice (1984) did also a pretty fair job for the models and sample sizes considered in this article.

References

AKAIKE, H. (1974) A New Look at the Statistical Model Identification, *IEEE Transactions on Automatic Control* **19**, 716-723.

AKAIKE, H. (1970) Statistical Predictor Information, *Annals of the Institute of Statistical Mathematics* **22**, 203-217.

CAO-ABAD, R. (1991). Rate of Convergence for the Wild Bootstrap in Nonparametric Regression, *The Annals of Statistics* **19**: 2226-2231.

CAO-ABAD, R. AND GONZÁLEZ-MANTEIGA, W. (1993). Bootstrap Methods in Regression Smoothing, *Nonparametric Statistics* **2**: 379-388.

CHEN, M.-J., FAN, J., MARRON, J.S. (1997) On automatic boundary corrections, *The Annals of Statistics* **Vol. 25, No. 4**, 1691-1708.

CHIU, S.-T. (1990) On the Asymptotic Distributions of Bandwith Estimates, *The Annals of Statistics* **18**, 1696-1711.

CLARK, R. M. (1977) Non-Parametric Estimation of a Smooth Regression Function, *Journal of the Royal Statistical Society, Series B* **39**: 107-113.

CRAVEN, P., WAHBA, G. (1979) Smoothing Noisy Data With Spline Functions, *Numerische Mathematik* **31**, 377-403.

FAN, J. (1992). Design-Adaptive Nonparametric Regression, *Journal of American Statistical Association*, **87**, 998-1004.

FAN, J., GIJBELS, I. (1992) Variable bandwidth and local linear regression smoothers, *The Annals of Statistics* **Vol. 20**, 2008-2036.

GASSER, T., KNEIP, A. AND KÖHLER, W. (1991). A Fast and Flexible Method for Automatic Smoothing, *Journal of the American Statistical Association* **86**: 643-652.

GASSER, T., MÜLLER, H.G.. Kernel Estimation of Regression Functions. *Smoothing techniques in curve estimation (Lecture Notes in Mathematics)*, **757**, 23-68.

GASSER, T., MÜLLER H.G. (1979) Kernel Estimation of Regression Functions, *Smoothing Techniques for Curve Estimation*, Lecture Notes in Mathematics 757, eds. T. Gasser and M. Rosenblatt, Heidelberg: Springer-Verlag, pp. 23-68 **Vol. 87 No. 420**, 998-1004.

GONZÁLEZ-MANTEIGA, W., MARTÍNEZ MIRANDA, M.D. AND PÉREZ GONZÁLEZ, A. (2004). The choice of smoothing parameter in nonparametric regression through Wild Bootstrap, *Computational Statistics & Data Analysis* **47**: 487-515.

HALL, P. (1990). Using the bootstrap to estimate mean square error and select smoothing parameters in nonparametric problems, *Journal of Multivariate Analysis* **32**: 177-203.

HÄRDLE, W. (1992). Applied Nonparametric Regression, *Cambridge University Press.*

HÄRDLE, W., HALL, P., MARRON, J.S. (1988) How far are automatically chosen Smoothing Parameters from their Optimum, *Journal of American Statistical Association* **83**, 86-95.

HÄRDLE, W. AND MAMMEN, E. (1993). Comparing nonparametric versus parametric regression fits, *The Annals of Statistics*, **21**, 1926-1947.

HÄRDLE, W., MARRON, J.S. (1985) Optimal Bandwidth Selection in Nonparametric Regression Function Estimation. *The Annals of Statistics* **13**, 1465-1481.

HÄRDLE, W. AND MARRON, J.S. (1991). Bootstrap simultaneous error bars for nonparametric regression, *The Annals of Statistics*, **19**, 778-796.

HÄRDLE, W.; MÜLLER, M.; SPERLICH, S. AND WERWATZ, A. (2004). Nonparametric and Semiparametric Models, *Springer Series in Statistics*, Berlin.

HART, J. D. (1994). Automated Kernel Smoothing of Dependent Data by Using Time Series Cross-Validation, *Journal of the Royal Statistical Society, Series B* **56**: 529-542.

HART, J. D. AND YI, S.(1998). One-Sided Cross Validation, *Journal of American Statistical Association* **93**: 620-631.

HEIDENREICH, N.B., SCHINDLER, A. AND SPERLICH, S. (2010). Bandwidth Selection Methods for Kernel Density Estimation - A Review of Performance *SSRN Discussion Paper: papers.ssrn.com/sol3/papers.cfm?abstract_id=1726428*

HURVICH, C. M., SIMONOFF, J.S. AND TSAI C. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion, *Journal of the Royal Statistical Society, Series B* **60**: 271-293.

MALLOWS, C.L. (1973) Some Comments on C_p, *Technometrics* **15**, 661-675.

MAMMEN, M., MARTÍNEZ-MIRANDA, M.D., NIELSEN, J.P. AND SPERLICH, S. (2011). Do-validation for Kernel Density Estimation. *Journal of the American Statistical Association, forthcoming*.

MARRON, J.S. (1986). Will the Art of Smoothing ever become a Science, *Function Estimates* (Contemporary Mathematics 59), Providence, RI: American Mathematical Society, pp. 169-178.

NADARAYA, E.A. (1964) On Estimating Regression, *Theory of Probability and its Application* **9**, 141-142.

NIELSEN, J. P. (1999) Scand. Actuarial, **1**, 93.95.

PARK, B.U. AND MARRON, J.S. (1990). Comparison of Data-Driven Bandwidth Selectors, *Journal of the American Statistical Association* **85**: 66-72.

PRIESTLEY, M. B., CHAO, M.T. (1972) Non-parametric function fitting, *Journal of the Royal Statistical Society, Series B* **34**, 385-392.

RICE, J. (1984) Bandwith Choice for Nonparametric Regression, *The Annals of Statistics* **Vol. 12, No. 4**, 1215-1230.

RUPPERT, D.; SHEATHER, S.J. AND WAND, M.P. (1995). An Effective Bandwidth Selector for Local Least Squares Regression, *Journal of the American Statistical Association* **90(432)**: 1257-1270.

RUPPERT, D., AND WAND, M.P. (1994). Multivariate Locally Weighted Least Squares Regression, *The Annals of Statistics* **22**, 1346-1370.

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

SHIBATA, R. (1981) An Optimal Selection of Regression Variables, *Biometrika* **68**, 45-54.

WATSON, G.S. (1964) Smooth Regression Analysis, *Sankhyā, Series A* **26**, 359-372.

YANG, L. AND TSCHERNIG, R. (1999) Multivariate bandwidth selection for local linear regression, *Journal of the Royal Statistical Society, Series B* **61**: 793-815.

YI, S. (2001). Asymptotic Stability of the OSCV Smoothing Parameter Selection, *Communications in Statistics - Theory and Methods*, **30**, 2033-2044.

Chapter 3

Improving Nonparametric Regression by Prior-transformation ?

Abstract

In this essay a transformation for nonparametric kernel regression is tested for achieving a better ASE behavior. Three different bandwidth selection methods with an expected good behavior are assessed, i.e. the performance of the untransformed and the transformed version are compared. The idea of the transformation is to obtain good results for the estimation of the local linear estimator with the help of a global bandwidth, although the density of the explanatory variable needs a local bandwidth selection. The topic of this essay was suggested and supported by Prof. Dr. Stefan Sperlich.

3.1 Introduction

Suppose that there are random pairs (X_i, Y_i), $i = 1, 2, \ldots n \in \mathbb{N}$, where the X_i's are explanatory variables drawn from a continuous distribution with density function f. Without loss of generality, $X_1 < X_2 < \ldots < X_n$ is assumed. The Y_i's are response variables generated by the following model:

$$Y_i = m(X_i) + \sigma(X_i)\varepsilon_i, \qquad \text{where } \{\varepsilon_i\} \text{are i.i.d. } N(0,1) \tag{3.1}$$

Further, $\sigma^2(x) = var(Y|x)$ is finite, and the ε_i are independent of all X_j. Assume that one aims to estimate $m(x) = E(Y \mid X = x)$ for an arbitrary point $x \in \mathbb{R}$. Let $K : \mathbb{R} \to \mathbb{R}$ be a kernel function that fulfills $\mu_0(K) = 1$, $\mu_1(K) = 0$ and $\mu_2(K) < \infty$, where $\mu_l(K) = \int_{-\infty}^{\infty} u^l K(u)\,\mathrm{d}u$. Furthermore, denote $K_h(u) := \frac{1}{h}K(u/h)$, where $h \in \mathbb{R}^+$ is the smoothing parameter (bandwidth).
The aim is to use a global smoothing parameter h although the probability density function of X has a sharp peak, i.e. a region with many observations, where a small bandwidth is better suited for estimation. It also has a smooth area where a big bandwidth is required, otherwise, the estimation of m for values x of this region would be equal to zero. The problem of global bandwidth selection in these cases is illustrated in figure 3.1 where a typical representative of the described distribution, $X \sim ln\mathcal{N}(0,1)$ is chosen. Although a local bandwidth selection would be suitable, global bandwidth selection using a transformation is applied in this article.

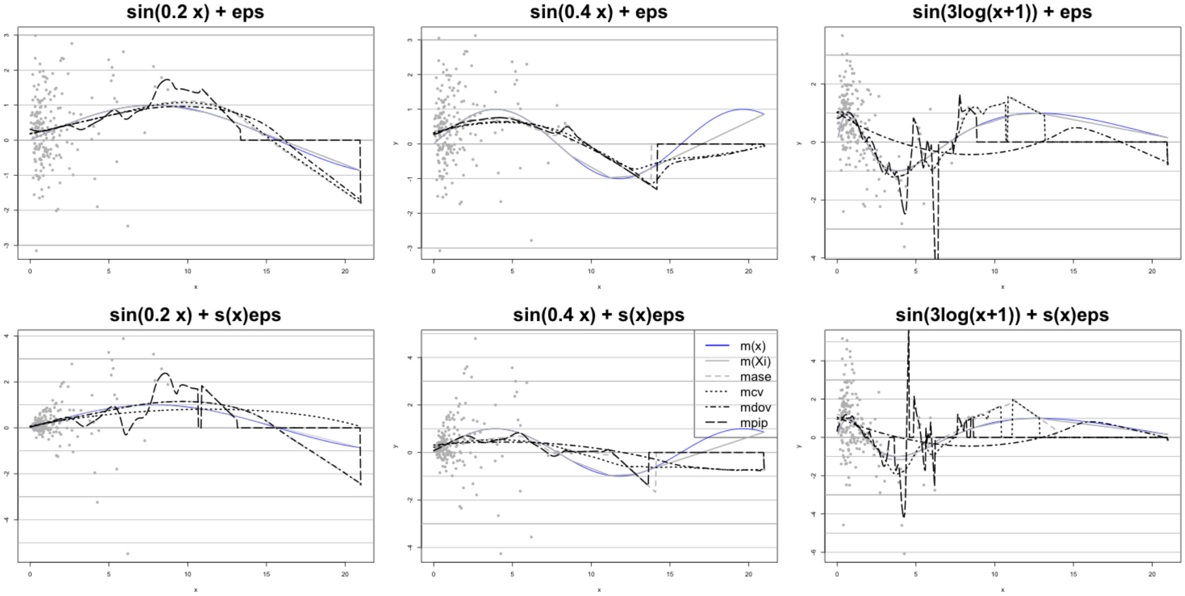


Figure 3.1: $\hat{m}_h(x)$ for distribution $X \sim ln\mathcal{N}(0,1)$, assuming different sinus functions for $m(x)$. The first row shows homoscedastic models, and the second row shows heteroscedastic models. The dashed, dotted and dashed-dotted lines show the estimation of $\hat{m}_h(x)$ by different bandwidth selection methods: CV, DoV, and PIP (see description below).

In the literature there is a fully nonparametric linear transformation approach with global bandwidth selection, which is introduced in Park et al. (1997). They propose a transformation to reduce the MISE (mean integrated squared error) in kernel regression. The idea is to regress Y_i on x_i in the first stage, where they use a fixed design of x and calculate the estimator $\hat{m}_b(x)$ with a global bandwidth b. In the second stage they propose two alternatives: regress Y_i on $\hat{m}_b(x_i)$ using bandwidth h and the transformed data for the kernel weights or the same bandwidth b from the first stage and use the original scale for calculating the kernel weights. In their simulation study they compare the second alternative of the transformation estimator with different local polynomial estimators $\hat{m}_{p,h}(x)$.The conclusion is that there are only slight differences between the transformation and the local cubic estimator. The overall performance of the local cubic estimator is better, but for some special cases its behavior is really bad. Hence, they use the transformation approach for their practical example.

In the article of Park et al. (1997) the problem of avoiding local bandwidth selection is not been addressed. Further, in this thesis the local linear estimator in estimating $m(x)$ and the different methods are compared. Hence, the transformation, tested in this article follows another approach. The idea is to map the data x into the interval $[0,1]$ with the help of a bijective strictly monotonic increasing function first. For this transformation the parametric cumulative distribution function of X is used:

$$x \to F_X(x)$$

Obviously, a nonlinear transformation is applied to the data. The advantage is that the interval including the highest part of the observation is extended and the part with only a few observations is compressed, of course proportional to the transformed range $[0,1]$. This results in almost uniformly distributed transformed sample as shown in figure 3.1.

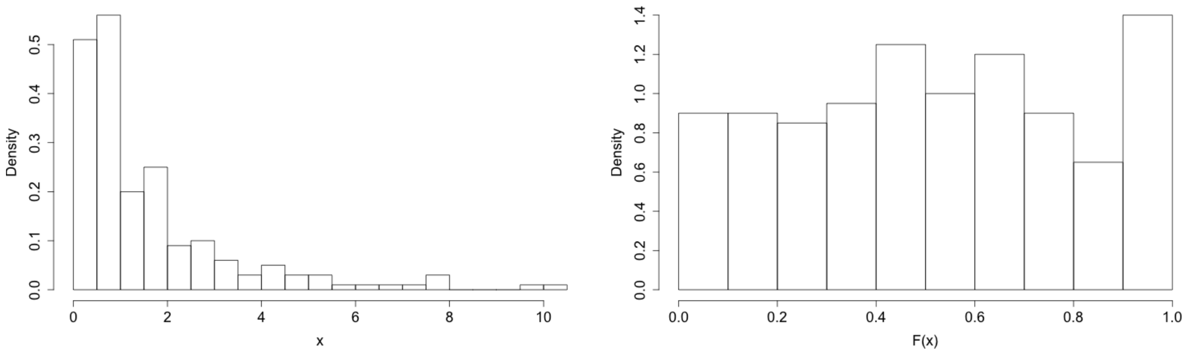


Figure 3.2: Histogram of a lognormally distributed sample X (left) vs. histogram of the transformed sample $F_X(X)$ (right)

Now the following question arises: How to obtain the suitable transformation? In the simulation study below, X is assumed to follow some distribution, i.e. it is not necessary to guess it. But, if someone wants to apply this approach, one has to guess the suitable parametric probabilty density function, by plotting a histogram of X. The next step is to regress Y on $F_X(x)$ nonparametrically, to find a global bandwidth h. In the following $F_X(x)$ will be denoted by z and hence $Z_i = F_X(X_i)$. For the estimation the well-known local linear kernel estimator $\hat{m}_h(z)$ is used. It can be expressed as a weighted sum of the Y_i, i.e.

$$\hat{m}_h(z) = \frac{1}{n}\sum_{i=1}^{n} W_{h,i}(z)Y_i \tag{3.2}$$

with weights

$$W_{h,i}(z) = \begin{cases} \frac{nS_{h,2}(z)K_h(z-Z_i)-nS_{h,1}(z)K_h(z-Z_i)(Z_i-z)}{S_{h,0}(z)S_{h,2}(z)-S_{h,1}(z)^2} & \text{if } S_{h,0}(z)S_{h,2}(z) \neq S_{h,1}(z)^2 \\ n & \text{if } S_{h,0}(z)S_{h,2}(z) = S_{h,1}(z)^2, \text{ for } z = z_i \\ 0 & \text{otherwise} \end{cases}$$

where $S_{h,j}(z) = \sum_{i=1}^{n} K_h\left(z-Z_i\right)\left(Z_i-z\right)^j$. The local linear estimator for an arbitrary point $z \in \mathbb{R}$ is only defined, if $W_{h,i}(z) > 0$ for at least one i. Note that the matrix with entries $\{W_{h,i}(Z_j)\}_{i,j}$ gives the so-called hat-matrix in kernel regression.

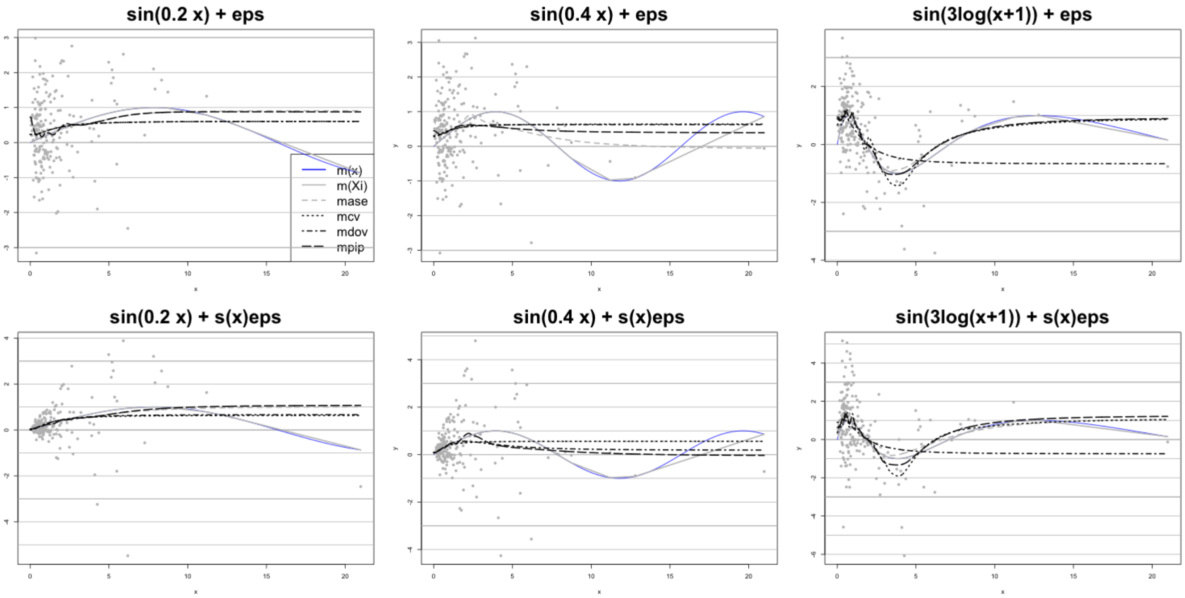
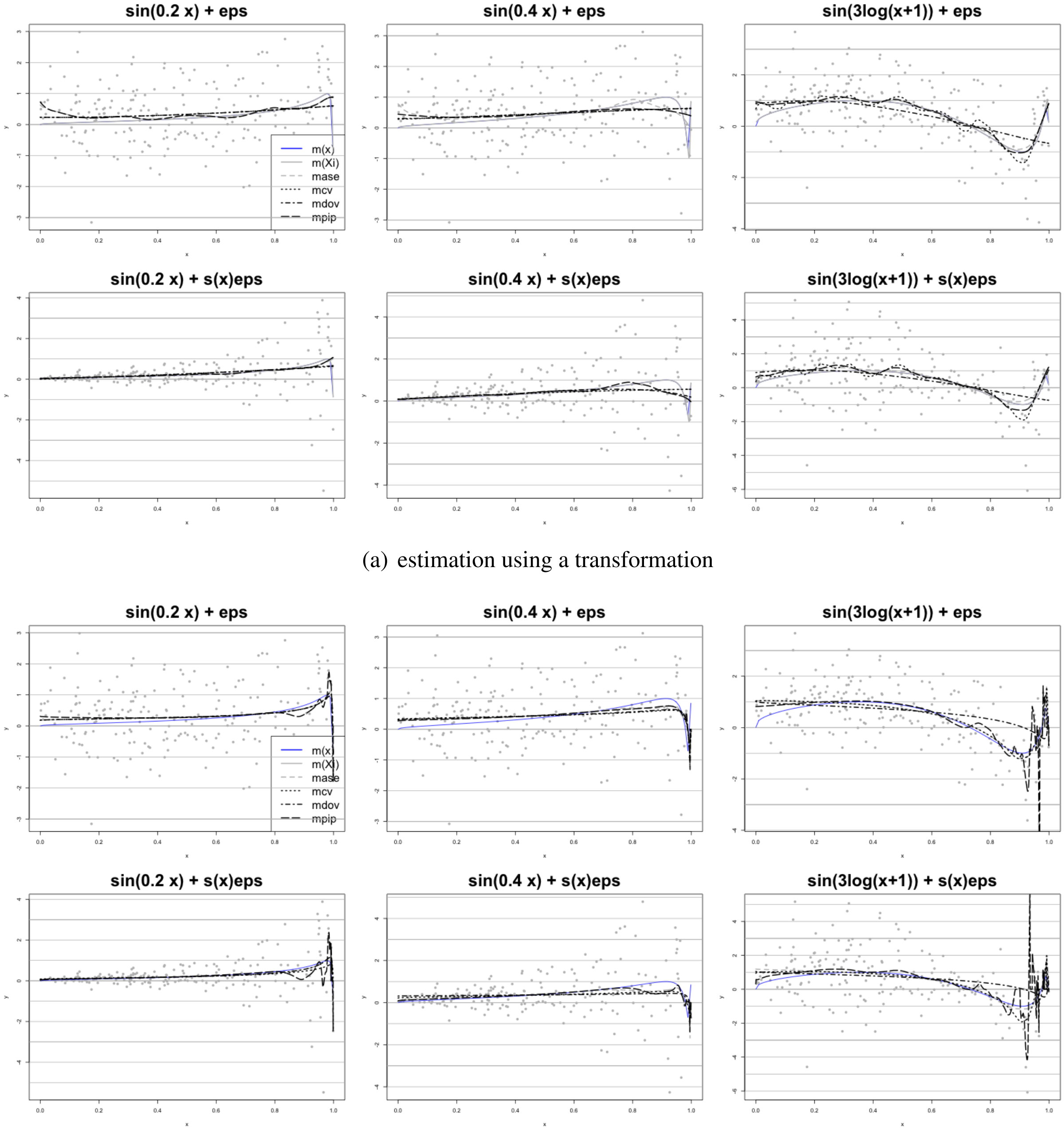


Figure 3.3: $\hat{m}_h(x)$ using a transformation for $X \sim ln\mathcal{N}(0,1)$, assuming different sinus functions for $m(x)$. The first row shows homoscedastic models, and the second row shows heteroscedastic models. The dashed, dotted and dashed-dotted lines show the estimation of $\hat{m}_h(x)$ by different bandwidth selection methods: CV, DoV, and PIP (see description below).

Because the transformation is only done for X_i but not for Y_i, the values of the true function m are the same at the transformed scale, and hence, $m(x) = E(Y \mid X = x) = E(Y \mid Z = z)$. On the transformed range a global bandwidth h is chosen for estimating $\hat{m}_h(z)$. If the values $\hat{m}_h(Z_i)$ are plotted against the original scale, it can be seen, that the estimator $\hat{m}$ is very wiggly in the area with many observations and very smooth in the region with less observations . This corresponds to local bandwidth selection, but if one looks at the figure 3.3 it can be seen, that in the region with sparse data the true m has been oversmoothed. Of course, there are no jumps anymore, but the question arises if the estimator using the transformation is better than the untransformed estimation.



(a) estimation using a transformation

(b) estimation without using any transformation

Figure 3.4: $\hat{m}_h(x)$ for $X \sim ln\mathcal{N}(0,1)$ on the transformed scale (a) and (b)

The reason for the oversmoothing becomes obvious when looking at figure 3.4. The true m on the transformed scale that we aim to estimate in the transformation approach, is very smooth in the subinterval $[0,0.9)$ because of the stretching. The oscillations of the sinus functions in the

left and middle column in figure 3.4 are in the subinterval $[0.9, 1]$. Hence, again the problem arises that it would be better to estimate the optimal $\hat{m}_h$ with the help of a local bandwidth. But by comparing subfigure 3.4(a) to subfigure 3.4(b) it could be assumed, that the estimation with the transformation approach works better. Obviously, the estimation quality depends largely upon the true m, i.e. the smoothness of m in the area on the transformed scale. The transformed estimation will be better if on the original scale m is more wiggly in the area with many observations and smooth in the area with sparse data.
From this point the pivotal question, i.e. the tradeoff between untransformed and transformed estimation, becomes clear: Is it better to have jumps in the region with sparse data or to oversmooth there too much?

For the purpose of assessing the estimation quality and answering this question, suitable risk measures are needed. In Köhler, Schindler, Sperlich (2011) there is an extensive recapitulation of different risk measures. To evaluate the estimation results the *ASE* (averaged squared error) is used, which is defined by

$$ASE(h) = \frac{1}{n}\sum_{j=1}^{n}\left(\hat{m}_h(X_j) - m(X_j)\right)^2 w(X_j), \tag{3.3}$$

where an additional trimming or weight function w is introduced to eliminate summands $(\hat{m}_h(X_j) - m(X_j))^2$ where X_j is near to the boundary. Having the explanatory variables ordered, $w(X_j) = 1_{[X_{l+1}, X_{n-l}]}$ can simply be set for a given l. By this means, the variability of the *ASE* score function can be greatly reduced, see Gasser and Müller (1979). Denote the minimizer of *ASE* by $\hat{h}_0$.
The *ASE* for the transformed data are given as the true m does not depend on the Z_j but is still a function of the original data X_j. As already mentioned above, the values of m do not change on the transformed scale and are now dependent on the observations, i.e. for each observation j a value of m is assigned. Hence, the true m for a specific observation j with $j = 1, 2, \ldots, n$ is denoted by m_j. Furthermore, the optimal bandwidth of the transformed estimation is denoted by ht. If we plot the estimated $\hat{m}_{ht}(Z_j)$ on the original scale, the global bandwidth ht from the estimation on the transformed scale is on the original scale no longer global, and hence, the $\hat{m}$ does no longer depend on the bandwidth parameter h but on the observation j and is denoted by $\hat{m}_j$. The procedure for obtaining the ASE-optimal bandwidth is the following: transform the values X_j as described. Then estimate the n-dimensional vectors $\hat{m}_h(Z) = \hat{Y}$ for each h. The true *ASE* with respect to one bandwidth ht, chosen on the transformed scale, is therefore calculated as:

$$ASE(ht) = \frac{1}{n}\sum_{j=1}^{n}\left(\hat{m}_{ht}(Z_j) - m_j\right)^2 w(Z_j) = \frac{1}{n}\sum_{j=1}^{n}\left(\hat{m}_j - m_j\right)^2 w_j \tag{3.4}$$

because the transformation with a cumulative distribution function is strictly monotonic increasing, the boundary points of sorted vector X are the boundary points of the vector Z and therefore only dependent on the observation j, denoted by w_j.
There are several bandwidth selection methods to find the optimal estimator for $m(x)$. An extensive simulation study comparing numerous of them is given in Köhler, Schindler and Sperlich

(2011). In section 3.2 the bandwidth selection methods, used for the simulation study in section 3.3, are presented in brief.

3.2 Bandwidth selection methods

The aim is to find a suitable global bandwidth for estimating model (3.1). Because of the findings in Köhler, Schindler and Sperlich (2011) the pool of possible bandwidth selection methods had been reduced. For the simulation study below only the following are used:

CV: cross-validation method, introduced by by Clark (1977)

DoV: do-validation (by Mammen, Martínez-Miranda, Nielsen and Sperlich (2011)), i.e. mixture of left and right one-sided cross-validation, see Hart and Yi (1998)),

Shib: Shibata's model selector, see Shibata (1981)

Rice: Rice's T, see Rice (1984)

PIrot: plug-in rule-of-thumb, see Ruppert, Sheather and Wand (1995).

To know what are the features of the different methods, a short revision for this methods will be given in the following subsections. For a more precisely description, see the literature, just mentioned. For an extensive review see Köhler, Schindler and Sperlich (2011).

3.2.1 Choosing the smoothing parameter based on ASE

The score functions for CV, DoV, Shib, and Rice are methods to estimate the *ASE* function in practice when the true function m is not known. The approach for estimating the *ASE* function is plugging into (3.3) response Y_j for $m(X_j)$. This yields the substitution estimate

$$p(h) = \frac{1}{n}\sum_{j=1}^{n}(\hat{m}_h(X_j) - Y_j)^2 w(X_j). \tag{3.5}$$

It can easily be shown, that this is a biased estimator of $ASE(h)$, see for example Härdle (1992), chapter 5. One can accept a bias that is independent of h as in this case the minimizer of (3.5) is the same as that of (3.3). Unfortunately this is not the case for $p(h)$. The following methods are approaches to correct for the bias.

The Corrected ASE

It is clear that $h \downarrow 0$ leads to interpolation, i.e. $\hat{m}_h(X_j) \to Y_j$, so that the function to be minimized, namely $p(h)$, could become arbitrarily small. On the other hand, this would surely cause a very large variance of $\hat{m}_h$ what indicates that such a criterion function would not balance bias and variance. Consequently, the corrected ASE penalizes when choosing h too small in an (at least asymptotically) reasonable sense. Hence, it is defined as following

$$G(h) = \frac{1}{n} \sum_{j=1}^{n} (Y_j - \hat{m}_h(X_j))^2 \, \Xi \left(\frac{1}{n} W_{h,j}(X_j) \right) w(X_j), \tag{3.6}$$

where $w(X_j) = 1_{[X_{l+1}, X_{n-l}]}$ is used to trim near the boundary. $\Xi(.)$ is a penalizing function with first-order Taylor expansion

$$\Xi(u) = 1 + 2u + O(u^2) \ , \ u \to 0. \tag{3.7}$$

The smaller the bandwidth h has been chosen the larger gets $W_{h,j}(X_j)$ and the penalizing factor $\Xi\left(\frac{1}{n}W_{h,j}(X_j)\right)$ increases. By conducting a first-order Taylor expansion of G and disregarding lower order terms it is easy to show that $G(h)$ is roughly equal to $ASE(h)$ up to a shift that is independent of h. There is a number of methods whose penalizing functions satisfy the expansion (3.7). In the simulation study two of them are used whose penalizing functions are given by:

- Shibata's model selector $\hat{h}_S = \underset{h \in \mathbb{R}^+}{\operatorname{argmin}} \, G_S(h)$, see Shibata (1981)

$$\text{with} \qquad \Xi_S(u) = 1 + 2u \, . \tag{3.8}$$

- Rice's T (T) $\hat{h}_T = \underset{h \in \mathbb{R}^+}{\operatorname{argmin}} \, G_T(h)$, see Rice (1984)

$$\text{with} \qquad \Xi_T(u) = (1 - 2u)^{-1} \, . \tag{3.9}$$

All these corrected ASE bandwidth selection rules are consistent for $n \to \infty$ and $nh \to \infty$ as $h \downarrow 0$. In practice they certainly exhibit some deficiencies. To mitigate the problems that may occur for too small bandwidths, we will fix a data-adaptive lower bound for $\hat{h}$. Notice that for $h \leq h_{min,j} := \min\left\{X_j - X_{j-1}, X_{j+1} - X_j\right\}$ (recall that the explanatory variables are ordered for the sake of presentation), we get $\frac{1}{n}W_{h,j}(X_j) = 1$ and $\frac{1}{n}W_{h,i}(X_j) = 0$ for all $i \neq j$. In this case the j'th summand of (3.6) is not defined if we choose $\Xi(.) = \Xi_{GCV}(.)$ or $\Xi(.) = \Xi_{FPE}(.)$ but is $\Xi(1)$ finite for all other penalizing functions such that the j'th summand of (3.6) gets zero. This shows that for sufficient small bandwidths h the score function $G(h)$ is either not defined or can be arbitrarily small. This does surely not solve the problem of balancing bias and variance of the local linear estimator. Therefore, first the infimum of the set of all bandwidths is calculated for which (3.6) can be evaluated,

$$h_{min,G} = \max\left\{h_{min,l+1}, \ldots, h_{min,n-l}\right\}. \tag{3.10}$$

When minimizing $G(h)$ for any of the above listed criteria, only the bandwidths h is used that fulfills $h > h_{min,G}$, all taken from the grid in (3.10).

Cross-validation

In the following the CV method introduced by Clark (1977) are presented. The proposed the score function

$$CV(h) = \frac{1}{n}\sum_{j=1}^{n}(Y_j - \hat{m}_{h,-j}(X_j))^2 w(X_j) \ , \tag{3.11}$$

where $\hat{m}_{h,-j}(X_j)$ is the leave one out estimator which is simply the local linear estimator based on the data $(X_1,Y_1),\dots(X_{j-1},Y_{j-1})$, $(X_{j+1},Y_{j+1}),\dots,(X_n,Y_n)$. In analogy to the ASE function, the weights $w(\cdot)$ are used to reduce the variability of $CV(h)$. Again, the trimming $w(X_j) = 1_{[X_{l+1},X_{n-l}]}$ is applied to get rid of boundary effects. It can easily be shown that this score function is a biased estimator of $ASE(h)$ but the bias is independent of h. This motivates the until today most popular data-driven bandwidth selection rule:

$$\hat{h}_{CV} = \operatorname*{argmin}_{h\in\mathbb{R}^+} CV(h) \ . \tag{3.12}$$

As for the corrected ASE bandwidth selection rules, the CV bandwidth selection rule is consistent but in practice, curiously has especially serious problems as $n \to \infty$. The reason is that this criterion hardly stabilizes for increasing n and the variance of the resulting bandwidth estimate $\hat{h}$ is often huge. Clearly, for $h < h_{min,j} := \min\left\{X_j - X_{j-1}, X_{j+1} - X_j\right\}$ there are similar problems as for the corrected ASE methods as then the local linear estimator $\hat{m}_h(X_j)$ is not defined. Therefore, (3.11) is only defined if we fix $h > h_{min,CV}$ with

$$h_{min,CV} := \max\left\{h_{min,l+1},\dots,h_{min,n-l}\right\} \ . \tag{3.13}$$

The One-Sided Cross Validation

As mentioned above the main problem of CV is the lack of stability resulting in large variances of its estimated bandwidths. As has been already noted by Marron (1986), the harder the estimation problem the better CV works. Based on this idea, Hart and Yi (1998) developed a new modification of CV.

Consider the estimator $\hat{m}_{\hat{h}_{CV}}$ with kernel K with support $[-1,1]$ that uses the CV bandwidth $\hat{h}_{CV}$. Furthermore, a second estimator $\tilde{m}_b$ with smoothing parameter b based on a (selection) kernel L with support $[0,1]$ is defined as:

$$OSCV(b) = \frac{1}{n-2l}\sum_{i=l+1}^{n-l}(\tilde{m}_b^{-i}(X_i) - Y_i)^2, \tag{3.14}$$

where $\tilde{m}_b^{-i}(X_i)$ is the leave-one-out estimator based on kernel L. Note that l must be at least 2. This ensures that in each summand of (3.14) at least $l-1$ data points can be used. Denote the minimizer of (3.14) by $\hat{b}$. The OSCV method makes use of the fact that a transformation $h : \mathbb{R}^+ \to \mathbb{R}^+$ exists, such that $E(h(\hat{b})) \approx E(\hat{h}_{CV})$ and $Var(h(\hat{b})) < Var(\hat{h}_{CV})$. More precisely, (3.14) is an unbiased estimator of

$$\sigma^2 + E\left[\frac{1}{n-2l}\sum_{i=l+1}^{n-l}(\tilde{m}_b(X_i) - m(X_i))^2\right] .$$

Therefore, minimizing (3.14) is approximately the same as minimizing

$$E\left[\frac{1}{n-2l}\sum_{i=l+1}^{n-l}(\tilde{m}_b(X_i)-m(X_i))^2\right]. \tag{3.15}$$

In almost the same manner it can be argued that minimizing $MASE(h)$ is approximately the same as minimizing $CV(h)$. Note that for calculating the $MASE$ (mean averaged squared error) the expectation of the ASE is taken.

The minimizer of (3.15) is denoted by b_n and the $MASE(h)$ minimizer by h_n. Using the results in Fan (1992) for minimizing the MASE-expressions, dividing the minimizers and taking limits yields

$$\frac{h_n}{b_n}\to\left[\frac{||K||_2^2}{(\mu_2^2(K))^2}*\frac{(\mu_2^2(L))^2}{||L||_2^2}\right]^{1/5}=:C,$$

see Yi (2001). Note that the constant C only depends on known expressions of kernels K and L. One can therefore define the data driven bandwidth selector

$$\hat{h}_{OSCV}=C\cdot\hat{b}. \tag{3.16}$$

According to which selection kernel L is used one gets different OSCV-values. Detailed discussions and simulation studies, which selection kernel L is optimal are given in Hart and Yi (1998), and in Köhler, Schindler and Sperlich (2011). Because of this discussion, in the simulation study below, the optimal Kernel from Hart and Yi (1998) is used, given by:

$$L(x)=(1-x^2)(6.92-23.08x+16.15x^2)1_{[0,1]},$$

with the respective kernel dependent constants $\mu_2^2(L)=-0.07692307$, $||L||_2^2=5.486053$, and $C=0.5192593$, which did also best in the simulation study in choosing L in Köhler, Schindler and Sperlich (2011).

As for the corrected ASE and CV bandwidth selection rules, the OSCV bandwidth selection rule is consistent. Now consider the i'th summand of (3.14). Analogously to prior discussions, (3.14) is only defined if $b>b_{min,lOSCV}=\max\{X_{l+1}-X_l,\ldots,X_{n-l}-X_{n-l-1}\}$, so that for minimizing (3.14) only bandwidths $b>h_{min,CV}$ are considered.

Note that the regression estimator used at the bandwidth selection stage, namely $\tilde{m}_b(x)$ in (3.16), uses only the data X_i that are smaller than the regression point x. This explains the notion left OSCV. For implementing the right OSCV, the kernel $R(u):=L(-u)$ is used. Note that this kernel has support $[-1,0]$ and therefore $\tilde{m}_b(x)$ uses only data at the right side of x. The transforming constant C in (3.16) does not change. There is evidence that the difference of left and right sided OSCV is negligible.

Even though one-sided cross validation from the left or from the right should not differ (from a theoretical point of view), in practice they do. To stabilize the behavior, Mammen, Martinez-Miranda, Nielsen, and Sperlich (2011) proposed to merge them to a so-called double one-sided or simply do-validation (half from the left-sided, half from the right-sided OSCV bandwidth) for kernel density estimation and obtained amazingly good results with that procedure. Also in Köhler, Schindler and Sperlich (2011) it can be seen, that DoV (do validation) has a good performance, especially for large sample sizes. Hence, it is used in the simulation study, below.

3.2.2 Plug-in methods: Choosing the bandwidth based on (A)MISE

In contrast to the cross-validation and corrected-ASE methods, the plug-in methods try to minimize the MISE (mean integrated squared error) or the AMISE (asymptotic MISE) of the local linear estimator $\hat{m}_h(x)$ in order to calculate the optimal bandwidth. The conditional weighted AMISE is given by:

$$AMISE(\hat{m}_h(x) \mid X_1,\dots,X_n) = \frac{1}{nh}||K||_2^2 \int_S \sigma^2(x)dx + \frac{h^4}{4}\mu_2^2(K)\int_S (m''(x))^2 f(x)dx, \qquad (3.17)$$

where the first summand is the mean integrated asymptotic variance, and the second summand the asymptotic mean integrated squared bias; cf. Ruppert, Sheather, and Wand (1995). The notation is the following: $f(x)$ indicates the density of X, $||K||_2^2 = \int K(u)^2du$, $\mu_l(K) = \int u^l K(u)du$, and f the unknown density of the explanatory variable X with the compact support $S = [a,b] \subset \mathbb{R}$. Finding a reasonable bandwidth means to balance the variance and the bias part of (3.17), therefore, an optimal bandwidth is obtained by minimizing (3.17) with respect to h. Clearly, the AMISE consists mainly of unknown functions and parameters. Consequently, the selection methods' main challenge is to find appropriate substitutes or estimates. Minimizing w.r.t. h, leads to the AMISE-optimal bandwidth (h_{AMISE}), given by:

$$h_{AMISE} = \left(\frac{||K||_2^2 \cdot \int_S \sigma^2(x)\,dx}{\mu_2^2(K)\cdot \int_S (m''(x))^2 f(x)dx \cdot n} \right)^{1/5}, \qquad (3.18)$$

where $S = [a,b] \subset \mathbb{R}$ is the support of the sample X of size n. Obviously, the two unknown quantities in 3.18 are $\int_S \sigma^2(x)dx$ and $\int_S (m''(x))^2 f(x)dx$, that have to be replaced by appropriate estimates. By using the quartic kernel 3.18 reduces to:

$$h_{AMISE} = \left(\frac{35 \cdot \int_S \sigma^2(x)\,dx}{\theta_{22}\cdot n} \right)^{1/5}, \qquad \theta_{rs} = \int_S m^{(r)}(x) m^{(s)}(x) f(x)dx, \qquad (3.19)$$

where $m^{(l)}$ denotes the lth derivative of m. In Köhler, Schindler and Sperlich (2011), three different strategies in obtaining the optimal bandwidth were described. Because of the performance in their simulation study, the rule-of-thumb bandwidth selector h_{rot} is used. This method is very straightforward, because the unknown quantities are replaced by parametric OLS estimators. In the simulation study below, a heteroscedastic model is tested. Hence, in the first step a parametric FGLS (feasible generalized least squares), more precisely a piece-wise polynomial regression fit is used as estimator to replace the unknown quantities.

For the sake of presentation assume the sample to be sorted in ascending order. The parametric FGLS-fit is a blocked quartic fit, i.e. the sample of size n is divided in N blocks $\chi_j = \left(X_{(j-1)n/N+1},\dots,X_{jn/N}\right)$, $(j = 1,\dots,N)$. For each of these blocks we fit the model:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4 + \varepsilon_i \quad i = (j-1)n/N+1,\dots,jn/N,$$

giving

$$\widehat{m}_{Q_j}(x) = \hat{\beta}_{0j} + \hat{\beta}_{1j}x_i + \hat{\beta}_{2j}x_i^2 + \hat{\beta}_{3j}x_i^3 + \hat{\beta}_{4j}x_i^4.$$

Then, the formula for the blocked quartic parametric estimator $\hat{\theta}_{22}$, is given by:

$$\hat{\theta}^{Q}_{22}(N) \;=\; \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{N}\widehat{m}^{(2)}_{Q_j}(X_i)\widehat{m}^{(2)}_{Q_j}(X_i)\mathbf{1}_{\{X_i\in\chi_j\}} = \frac{1}{n}\sum_{i=(j-1)n/N+1}^{jn/N}\;\sum_{j=1}^{N}\left(2\hat{\beta}_{2j}+6\hat{\beta}_{3j}x_i+12\hat{\beta}_{4j}x_i^2\right)^2,$$

and the estimator for σ_i^2

$$\hat{\sigma}^2_{i,Q} \;=\; \sum_{j=1}^{N}(Y_i-\widehat{m}_{Q_j}(X_i))^2\mathbf{1}_{\{X_i\in\chi_j\}} = \sum_{j=1}^{N}\left(y_i-\hat{\beta}_{0j}+\hat{\beta}_{1j}x_i-\hat{\beta}_{2j}x_i^2-\hat{\beta}_{3j}x_i^3-\hat{\beta}_{4j}x_i^4\right)^2\mathbf{1}_{\{X_i\in\chi_j\}}$$

To choose N we follow Ruppert, Sheather, and Wand (1995), respectively Mallows (1973): take the $\widehat{N}$ from $(1,2,\ldots,N_{max})$ that minimizes

$$C_p(N)=\frac{RSS(N)\cdot(n-5N_{max})}{RSS(N_{max})}-(n-10N),\text{ with } N_{max}=max\left[min(\lfloor n/20\rfloor,N^*),1\right],$$

where $RSS(N)$ is the residual sum of squares of a blocked quartic N-block-OLS, and with $N^*=5$ in our simulations. The idea of the rule-of-thumb bandwidth selector is to replace the unknown quantities in (3.19) directly by parametric estimates, i.e. for θ_{22} use The resulting rule-of-thumb bandwidth selector h_{rot} is given by

$$h_{rot}=\left(\frac{35\cdot\int_S\hat{\sigma}^2_{i,Q}dx}{\hat{\theta}^{Q}_{22}(N)\cdot n}\right)^{1/5},$$

which now is completely specified and feasible due to the various pre-estimates.

In Köhler, Schindler and Sperlich (2011) a blocked quartic fit was used, because for the so-called direct-plug-in method, that did not perform well, i.e. it chooses a bandwidth that is too small. For this method parametric estimators with a nonzero fourth derivative of $\widehat{m_Q}$ is needed. The rule-of-thumb plug-in method only requires a nonzero second derivative. For that reason, and because better results can be obtained, the estimation with a blocked parabolic FGLS-estimator θ_{22} is applied, given by:

$$\hat{\theta}^{P}_{22}(N) \;=\; \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{N}\widehat{m}^{(2)}_{P_j}(X_i)\widehat{m}^{(2)}_{P_j}(X_i)\mathbf{1}_{\{X_i\in\chi_j\}} = \frac{1}{n}\sum_{i=(j-1)n/N+1}^{jn/N}\;\sum_{j=1}^{N}\left(2\hat{\beta}_{2j}\right)^2,$$

and the estimator for σ_i^2

$$\hat{\sigma}^2_{i,P} \;=\; \sum_{j=1}^{N}(Y_i-\widehat{m}_{P_j}(X_i))^2\mathbf{1}_{\{X_i\in\chi_j\}} = \sum_{j=1}^{N}\left(y_i-\hat{\beta}_{0j}+\hat{\beta}_{1j}x_i-\hat{\beta}_{2j}x_i^2\right)^2\mathbf{1}_{\{X_i\in\chi_j\}}$$

The resulting rule-of-thumb bandwidth selector h^P_{rot} is given by

$$h^P_{rot}=\left(\frac{35\cdot\int_S\hat{\sigma}^2_{i,P}dx}{\hat{\theta}^{P}_{22}(N)\cdot n}\right)^{1/5}.$$

3.3 Simulation study

The simulation was done with the help of dyncall, developed by Adler and Philipp (2010), the specific R-package is called rdyncall, see Adler (2011). The library dyncall that is a function call interface for the C programming language among others. The package rdyncall works as following: first the C program is compiled and the compiled files are loaded via dyn.load to the R-program. Later it is possible to call the functions of the C program to obtain the results. In the simulation of this article, the data generation process is written in R, because it is very convenient to create pseudo random samples, there. After this, several C functions are called, to pass the data vectors and obtain the results in a short time, because the calculation in C is done very fast.

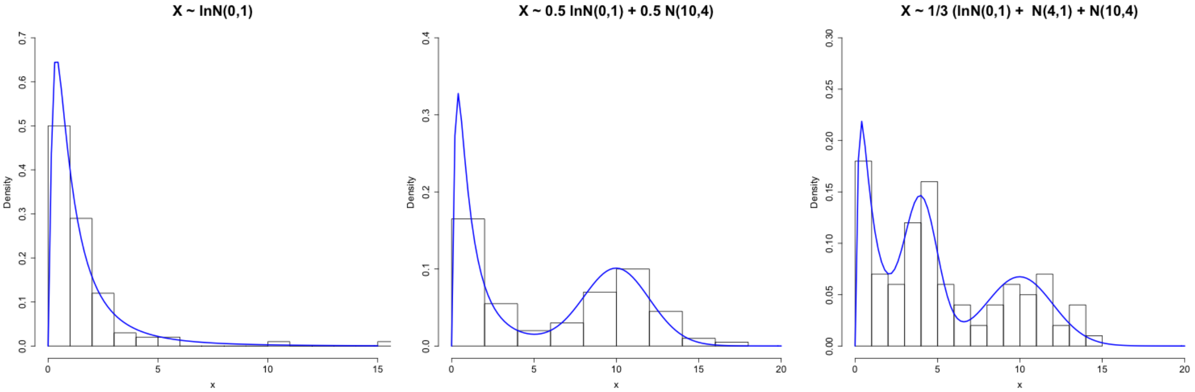


Figure 3.5: Histograms and true parametric probability density functions of X

For the data generating process, first the distribution of X has to be chosen, second, reasonable functions for $m(x)$ are needed and finally, a value for the variance of the error term has to be assumed. Because of the aim to test transformations in bandwidth selection methods for kernel regression, three different distributions of X that usually need a local bandwidth selection are assumed. Hence, the data for X are generated as $X \sim lnN(0,1)$, $X \sim 0.5(lnN(0,1)+N(10,4))$ and $X \sim 1/3(lnN(0,1)+N(4,1)+N(10,4))$. In figure 3.5 there is an illustration of the three cases via histograms of X with sample size 100 and the parametric density function from which the data were generated.

Further, three different functions for the true function $m(x)$ are assumed, shown in figure 3.6. These functions are two simple sinus functions $m_1(x) = sin(0.2x)$, and $m_2(x) = sin(0.4x)$ having different frequencies, and a sinus function having decreasing oscillations with increasing x: $m_3(x) = sin(3log(x+1))$ which assumed to have better results in the transformed estimation than in the untransformed one, because it is smoother in the area with sparse data than in the region with many data points. In figure 3.7 the true functions m are plotted against the transformed scale, remember the values of m do not change. There, the possible problems of global

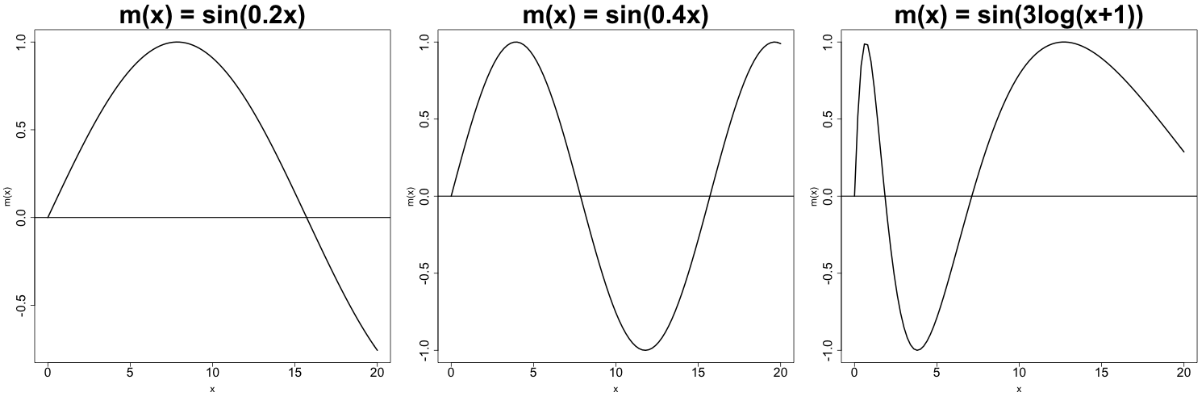


Figure 3.6: True functions of m(x)

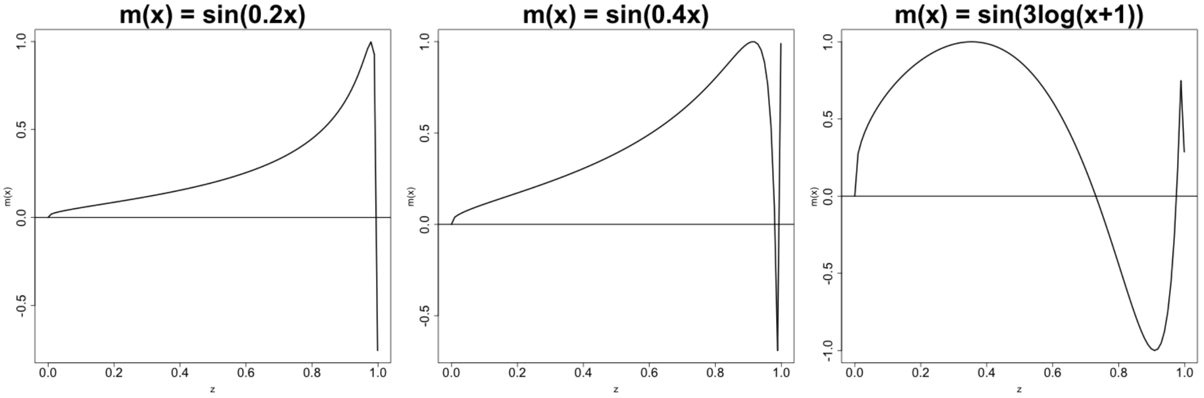


Figure 3.7: True functions of m(x) on the transformed scale

bandwidth selection, mentioned in the introduction, can be seen even in the case of estimating $m_3(x)$. For the simulation study two different true variances of the error term are assumed, the first is $\sigma_1^2 = 1$, assuming homoscedasticity. In order to see the performance of the methods in the case of heteroscedasticity, a variance $\sigma_2^2 = (2 \cdot |E(y|x)|)^2$ will be tested that is not independent from the observations, and $\varepsilon \sim \mathcal{N}(0,1)$. These assumptions result in 18 different true models, shown in figure 3.8 (in the Appendix) or summarized in table 3.1.

As already stated above, the purpose is to find the best estimate for $m(x)$ in model (3.1) by using the local linear estimator $\hat{m}_h(x)$. The first step is to evaluate the performance of the methods CV, DoV, Shib, Rice, and the two PIrot methods by using the respective transformation, i.e. the distribution function $F_X(x)$. The methods showing the best ASE behavior are chosen to be compared with their untransformed version.

More precisely, the first estimation step is to transform the data X. Then the regression Y_i on $F_X(X_i) = Z_i$ for $i = 1, 2, \ldots, n$ with the the local linear estimator $\hat{m}_{ht}(Z_i)$ is done resulting in the calculation of the optimal bandwidth $\widehat{ht}$ for each method with the respective true ASE-value,

$X \sim lnN(0,1)$		$X \sim 0.5(lnN(0,1)+N(10,4))$		$X \sim \frac{1}{3}(lnN(0,1)+N(4,1)+N(10,4))$	
No.	Model	No.	Model	No.	Model
1	$y = m_1(x) + \sigma_1\varepsilon$	7	$y = m_1(x) + \sigma_1\varepsilon$	13	$y = m_1(x) + \sigma_1\varepsilon$
2	$y = m_1(x) + \sigma_2\varepsilon$	8	$y = m_1(x) + \sigma_2\varepsilon$	14	$y = m_1(x) + \sigma_2\varepsilon$
3	$y = m_2(x) + \sigma_1\varepsilon$	9	$y = m_2(x) + \sigma_1\varepsilon$	15	$y = m_2(x) + \sigma_1\varepsilon$
4	$y = m_2(x) + \sigma_2\varepsilon$	10	$y = m_2(x) + \sigma_2\varepsilon$	16	$y = m_2(x) + \sigma_2\varepsilon$
5	$y = m_3(x) + \sigma_1\varepsilon$	11	$y = m_3(x) + \sigma_1\varepsilon$	17	$y = m_3(x) + \sigma_1\varepsilon$
6	$y = m_3(x) + \sigma_2\varepsilon$	12	$y = m_3(x) + \sigma_2\varepsilon$	18	$y = m_3(x) + \sigma_2\varepsilon$

Table 3.1: True regression models

calculated like in (3.4). Without loss of generality, the Quartic Kernel is used throughout, i.e. $K(u) = \frac{15}{16}(1-u^2)^2 1_{\{|u|\leq 1\}}$.
There are certainly many ways to compare the selection methods. Just when have in mind that different selectors are looking at different objective functions. To assess the results different performance measures are used, most of them based on the averaged squared error (ASE). Note that for the specific distributions, given above the weight function w in the formula of the ASE (see equations (3.3) and (3.4)) does not cut points on the right side of the distribution of X, because these are points in the area of sparse data, and hence, the main interesting points. The second reason is that there is no boundary on the the right side to worry about, because the distributions of X have exponential decreasing tails on the right side. Hence the weighting function changes to: $w(X_j) = 1_{[X_{l+1}, X_n]}$.
The considered performance measures are:

pm_1: $mean\left[ASE(\hat{h})\right]$
classical measure where the ASE of $\hat{m}$ is calculated (and averaged over the 1000 repetitions)

pm_2: $IQR\left[ASE(\hat{h})\right]$
interquartile range as measure of the volatility of the ASE's

pm_3: $mean(\hat{h} - \hat{h}_{ASE})$
'bias' of the bandwidth selectors, where h_{ASE} is the real ASE-minimizing bandwidth

pm_4: $mean\left[ASE(\hat{h}) - ASE(\hat{h}_{ASE})\right] = mean\left[\mid ASE(\hat{h}) - ASE(\hat{h}_{ASE}) \mid\right]$
L_1 distance of the ASE's based on selected bandwidths compared to the minimal ASE

pm_5: $mean\left(\left[ASE(\hat{h}) - ASE(\hat{h}_{ASE})\right]^2\right)$
squared L_2 distance compared to the minimal ASE

All results are based on the calculations from 1000 repetitions. In our simulation study we tried all methods for the sample sizes $n = 25$, $n = 50$, $n = 100$, $n = 200$, and $n = 500$.

3.3.1 Finite sample performance of the different methods using a transformation

First, the behavior of the bandwidth selection methods by using the transformation is tested. Two performance measures for all models and designs of X are illustrated: the bias (pm_3) of the selected bandwidth and the L1-distance (pm_4) of the ASE values.
The bias for the different bandwidths (pm_3) can be seen in figure 3.9 in the Appendix. The plug-in methods do worst in that cases, because they have the tendency to undersmooth, in general. The plug-in method with the blocked quartic fit has almost always the biggest bias in absolute value, for all models. CV has the best bias performance in general, i.e. independent from the model or sample size it does not fail.
The L1-distances (pm_4) can be seen in figure 3.10 in the Appendix. Also with respect to the L1-distance of the ASE-values, PIQ is the worst method of all. Its values for $n = 25$ are always out of the plotting range. DoV shows the best performance, only for model 11 and model 12 PI with the blocked parabolic fit behaves best. For the other models PIP is still the second best method. It can also be seen that the behavior of Shib and Rice is worse than for CV, DoV and PIP, in general. From that reason, the decision was only to compare CV, DoV and PIP in the next subsection, where the quality of the transformation is evaluated.

3.3.2 Finite sample performance of the transformation

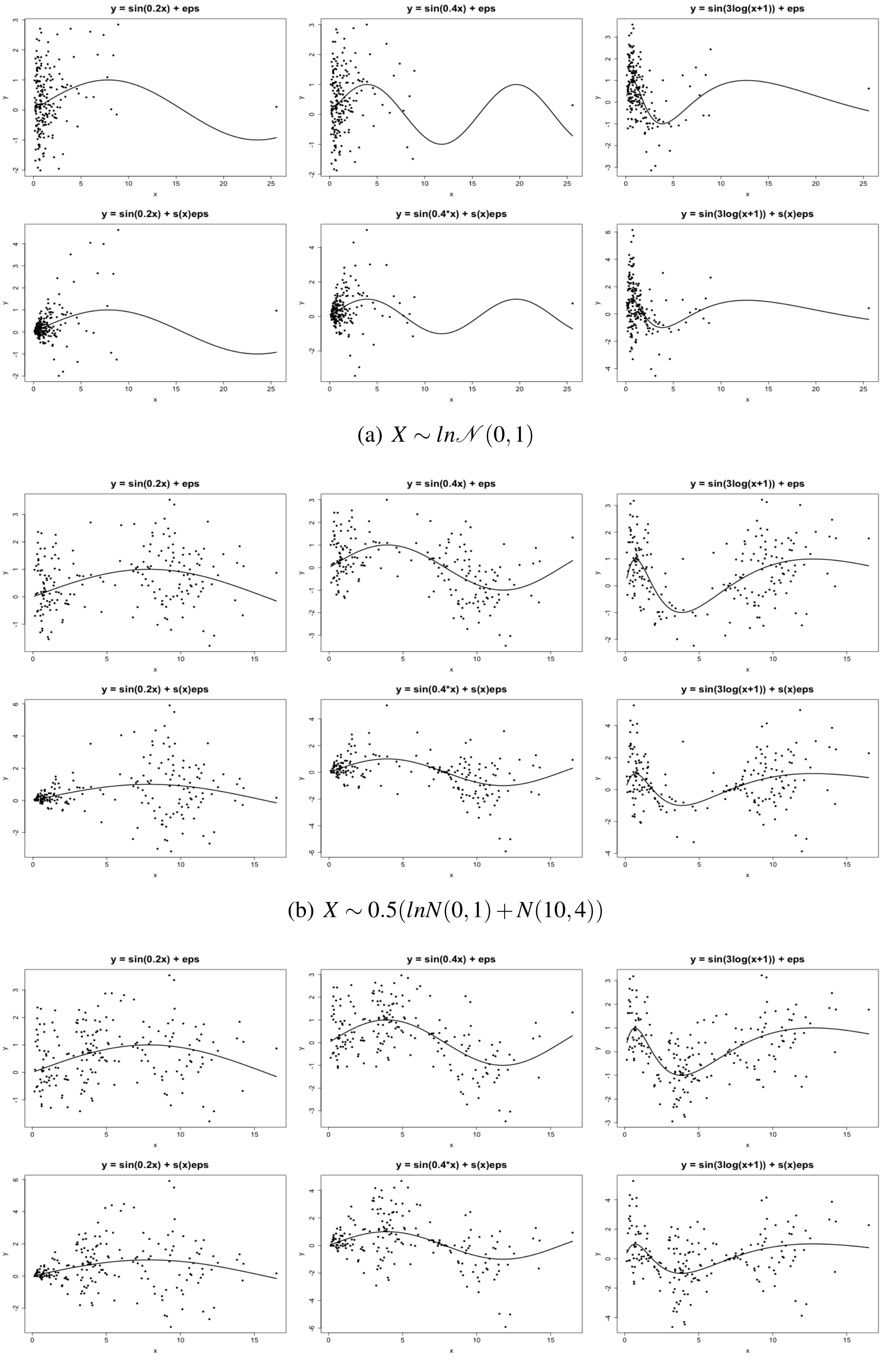
The next step in the simulation study is to test the transformation. As already mentioned above, several problems can arise by estimating a global bandwidth on the original data. Because of the finding in the last subsection, we will compare only three methods: CV, DoV and PIP , last letter “T” means with transformation. For this purpose, only the box plots of the ASE-values will be shown. The L1- and the L2-distances of the ASE values pm_4 and pm_5, used in the last subsection, are not suitable to compare the transformed against the untransformed version of the methods, because two different ASE-optimal bandwidths are chosen, and hence the ASE-values for the respective bandwidths are different. More precisely, the worse the ASE-value for the ASE-optimal bandwidths the better the L1 respectively L2 distance for a method. Furthermore, the performance measures based on the bandwidths are completely different, because for the estimation two different bandwidth grids are necessary.
Next, the results for the ASE-values of the different measures are summarized, i.e. the bandwidth that has been chosen for the respective method is inserted in the formula for the ASE. This is done because it enables to compare rather the resulting regression performance than the bandwidths selected by the different methods. The respective performance measures are pm_1 for the mean of the ASE-values, seen as squares in the box-plots and pm_2 the interquartile range (IQR) of the ASE-values, seen by the height of the boxes and of course the median of the ASE-values can be seen.
The basic question in evaluating the transformation is: Does the transformation work better than just using the untransformed version of the respective methods?
First the most important case: $X \sim lnN(0,1)$ (see figure 3.11) is considered. In general, it can be seen that the performance of the transformed versions of the methods is clearly better. As ex-

pected the best results are obtained by using $m_3(x) = sin\,(3log(x+1))$. For the true regression functions $m_1(x) = sin(0.2x)$ and $m_2(x) = sin(0.4x)$ the transformation works better for smaller n and for the heteroscedastic models. But although the transformation behave worse in case of homoscedasticity, it does not fail completely. By looking at the methods it can be seen that for $n = 25, 50, 100$ the DoV resp. DoVT is always the best. For $n = 200$ and $n = 500$ CV respectively CVT sometimes behaves better. If somebody would ask which method to use in general, the answer would be DoVT. Of course the results for the regression model $y = m_2(x) + \varepsilon$ for $n = 500$ but for all the other 29 cases the ASE values of DoVT are best or not far from the best. Focussing on $X \sim 1/3(lnN(0,1)N(10,4))$ (see figure 3.12), it can be seen that the untransformed methods behave better, and for $X \sim 1/3(lnN(0,1)+N(4,1)+N(10,4))$ (see figure 3.13) both the transformed and untransformed version are relatively equal. Obviously, the reason is because the untransformed methods have big problems especially in the case of $X \sim lnN(0,1)$. But if the problems with the distribution of X decrease, the problems with the regression function m, described above become more important. Hence, for $m_3(x)$ the performance is also acceptable for $X \sim 1/3(lnN(0,1)N(10,4))$ and $X \sim 1/3(lnN(0,1)+N(4,1)+N(10,4))$

3.4 Conclusions

In the simulation study different bandwidth selection methods for relatively smooth functions of $m(x)$ has been compared. The best methods were CV, DoV and PIP. If the transformed vs. the untransformed estimation was compared, the result was, that the best performance could be obtained by the DoV methods. The estimation could be improved by using the proposed transformation, especially DoVT turns out to be very good. The main conclusion is, that using the proposed transformation in combination with global bandwidth selection on the transformed scale is a very stable procedure to obtain a good estimation of $m(x)$, if the distribution of X is not very smooth and would actually require local bandwidth selection.

3.5 Appendix

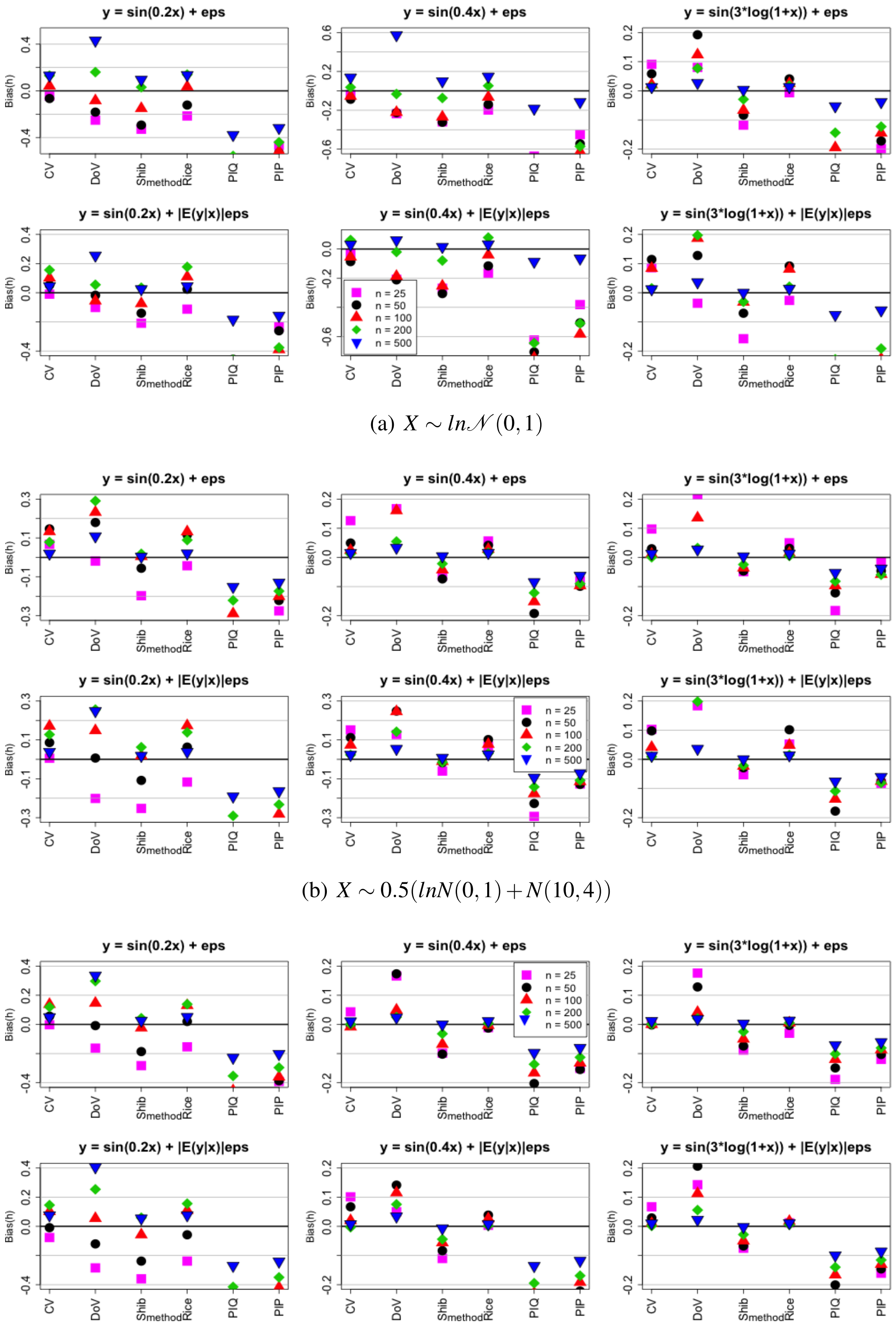


(a) $X \sim ln\mathcal{N}(0,1)$

(b) $X \sim 0.5(lnN(0,1)+N(10,4))$

(c) $X \sim 1/3(lnN(0,1)+N(4,1)+N(10,4))$

Figure 3.8: True regression models for distribution (a), (b) and (c)

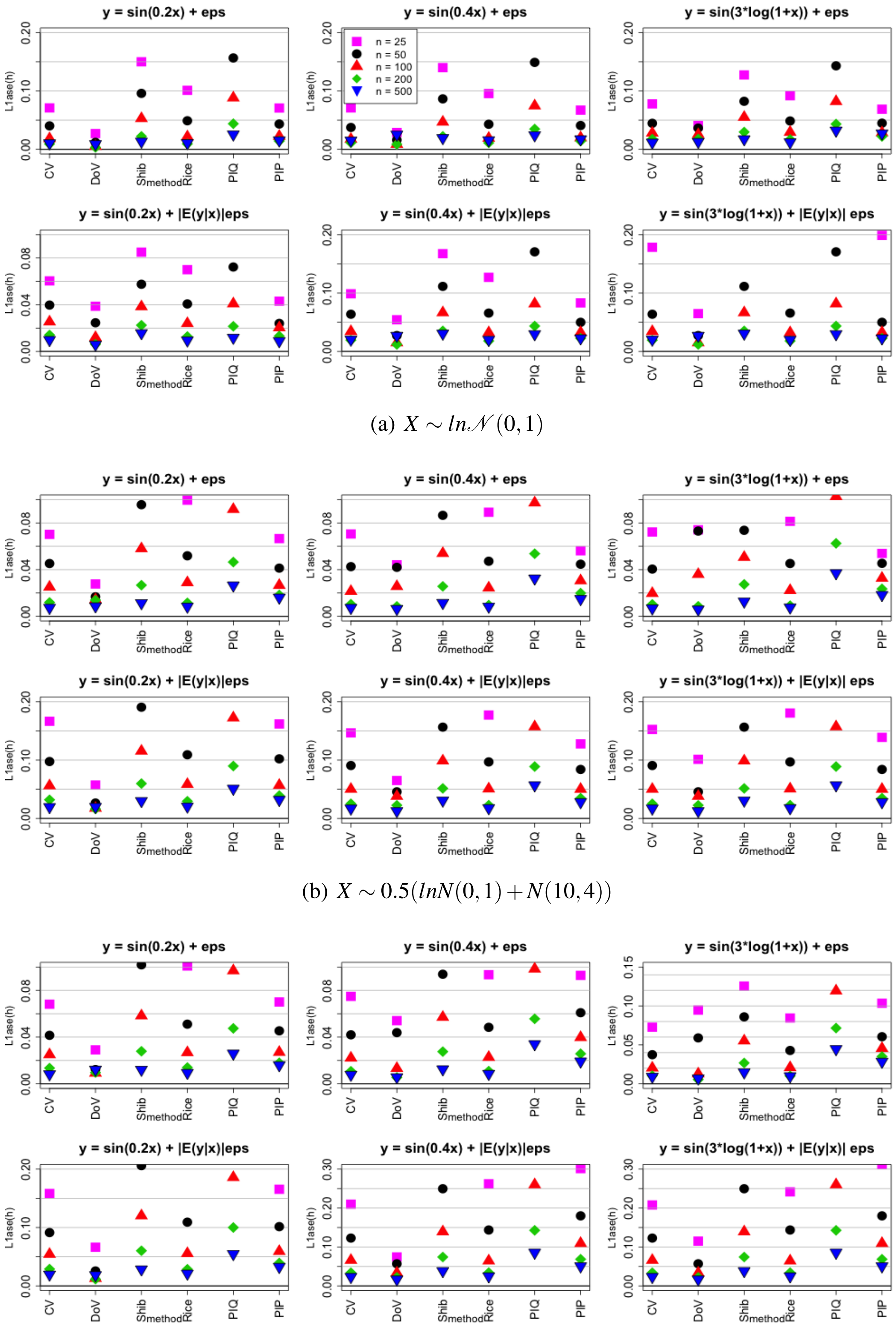


(a) $X \sim ln\mathscr{N}(0,1)$

(b) $X \sim 0.5(lnN(0,1)+N(10,4))$

(c) $X \sim 1/3(lnN(0,1)+N(4,1)+N(10,4))$

Figure 3.9: Bias of the selected bandwidths for the estimation with prior transformation, for distribution (a), (b) and (c)

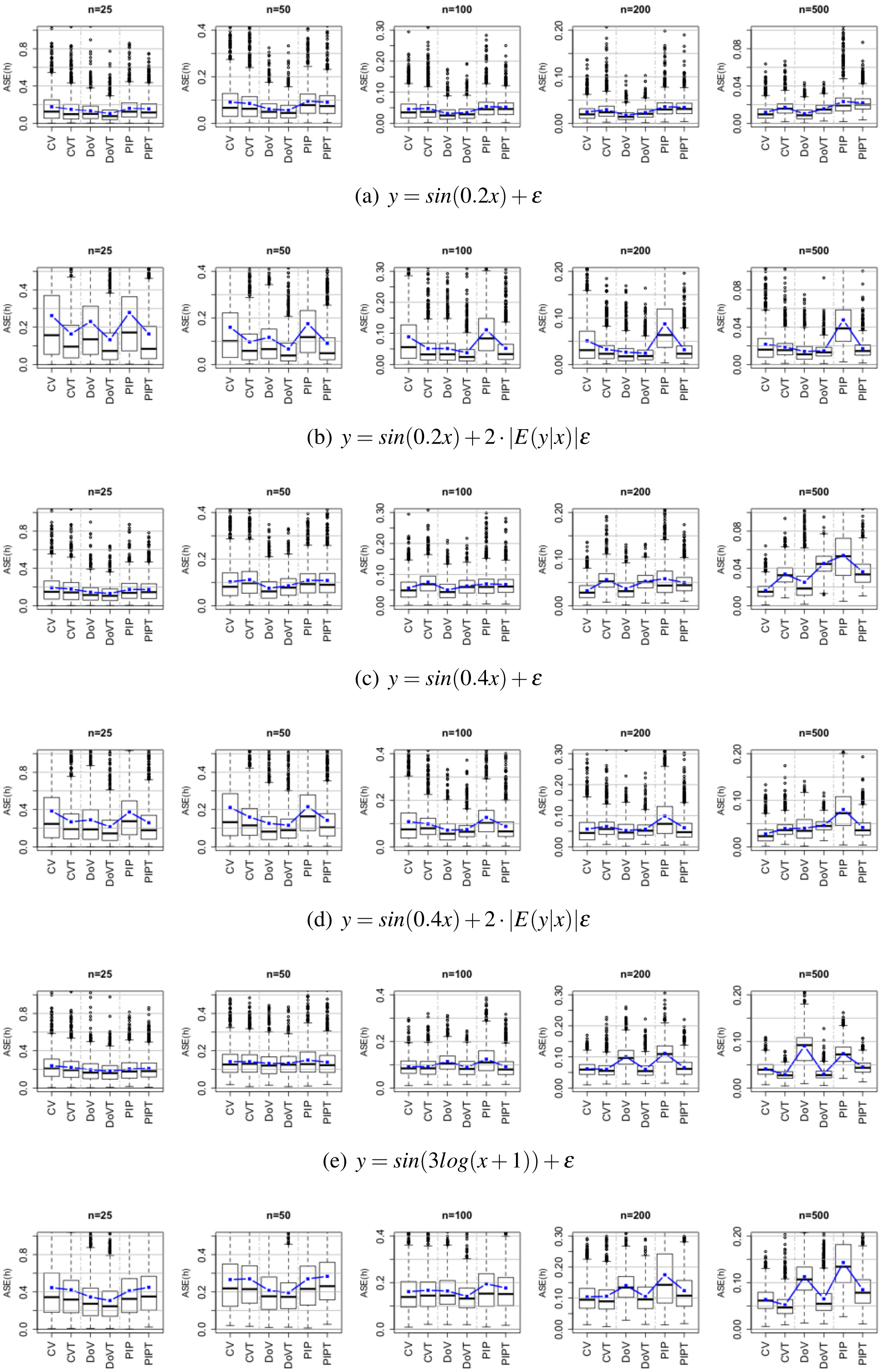


(a) $X \sim ln\mathcal{N}(0,1)$

(b) $X \sim 0.5(lnN(0,1)+N(10,4))$

(c) $X \sim 1/3(lnN(0,1)+N(4,1)+N(10,4))$

Figure 3.10: L1-distances of the ASE-values for the estimation with prior-transformation, for distribution (a), (b) and (c)



(a) $y = sin(0.2x) + \varepsilon$

(b) $y = sin(0.2x) + 2 \cdot |E(y|x)|\varepsilon$

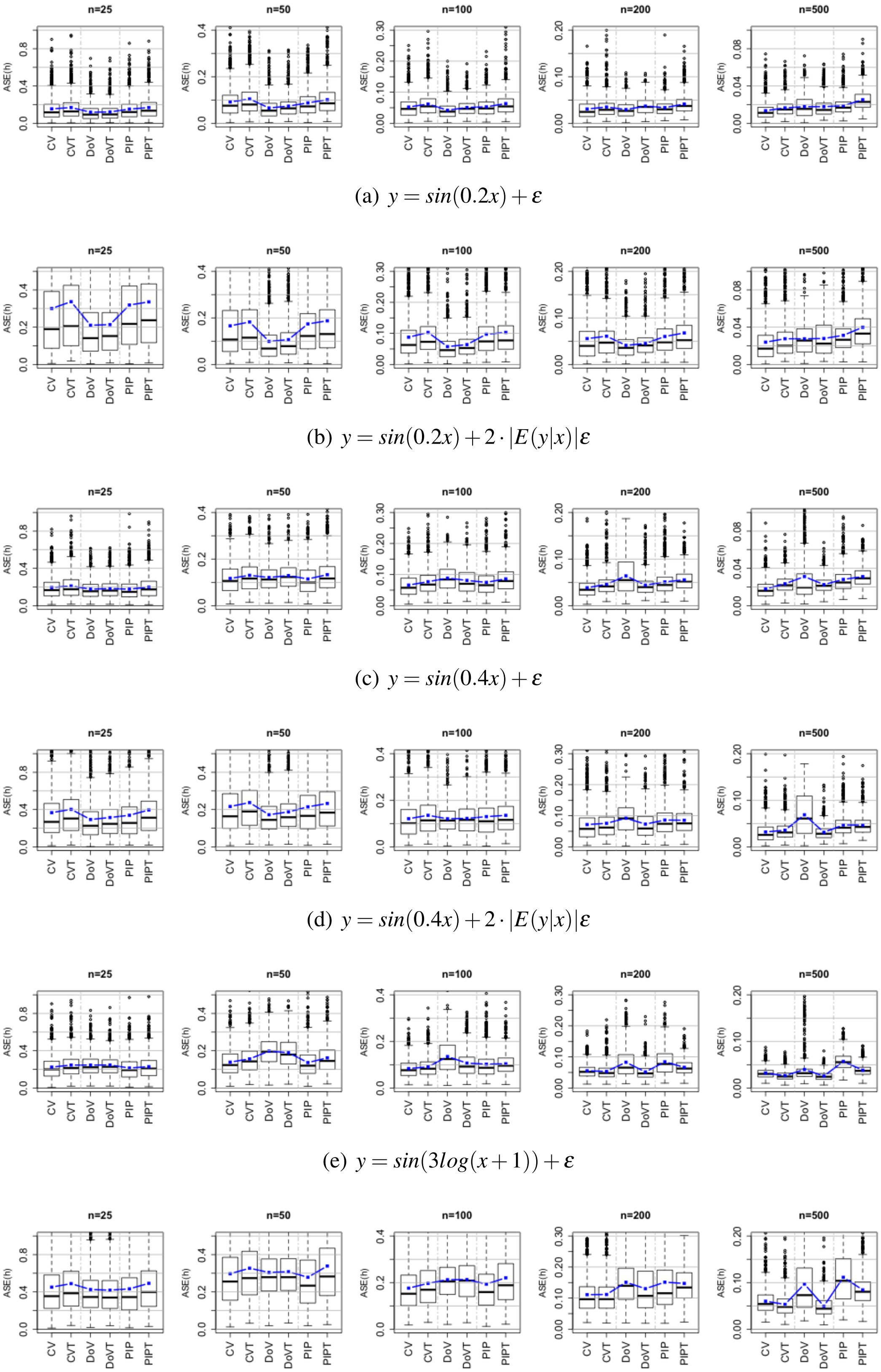
(c) $y = sin(0.4x) + \varepsilon$

(d) $y = sin(0.4x) + 2 \cdot |E(y|x)|\varepsilon$

(e) $y = sin(3log(x+1)) + \varepsilon$

(f) $y = sin(3log(x+1)) + 2 \cdot |E(y|x)|\varepsilon$

Figure 3.11: Box-plots and means (■) of the ASE-values for distribution $X \sim ln\mathscr{N}(0,1)$ for model (a) - (f)



(a) $y = sin(0.2x) + \varepsilon$

(b) $y = sin(0.2x) + 2 \cdot |E(y|x)|\varepsilon$

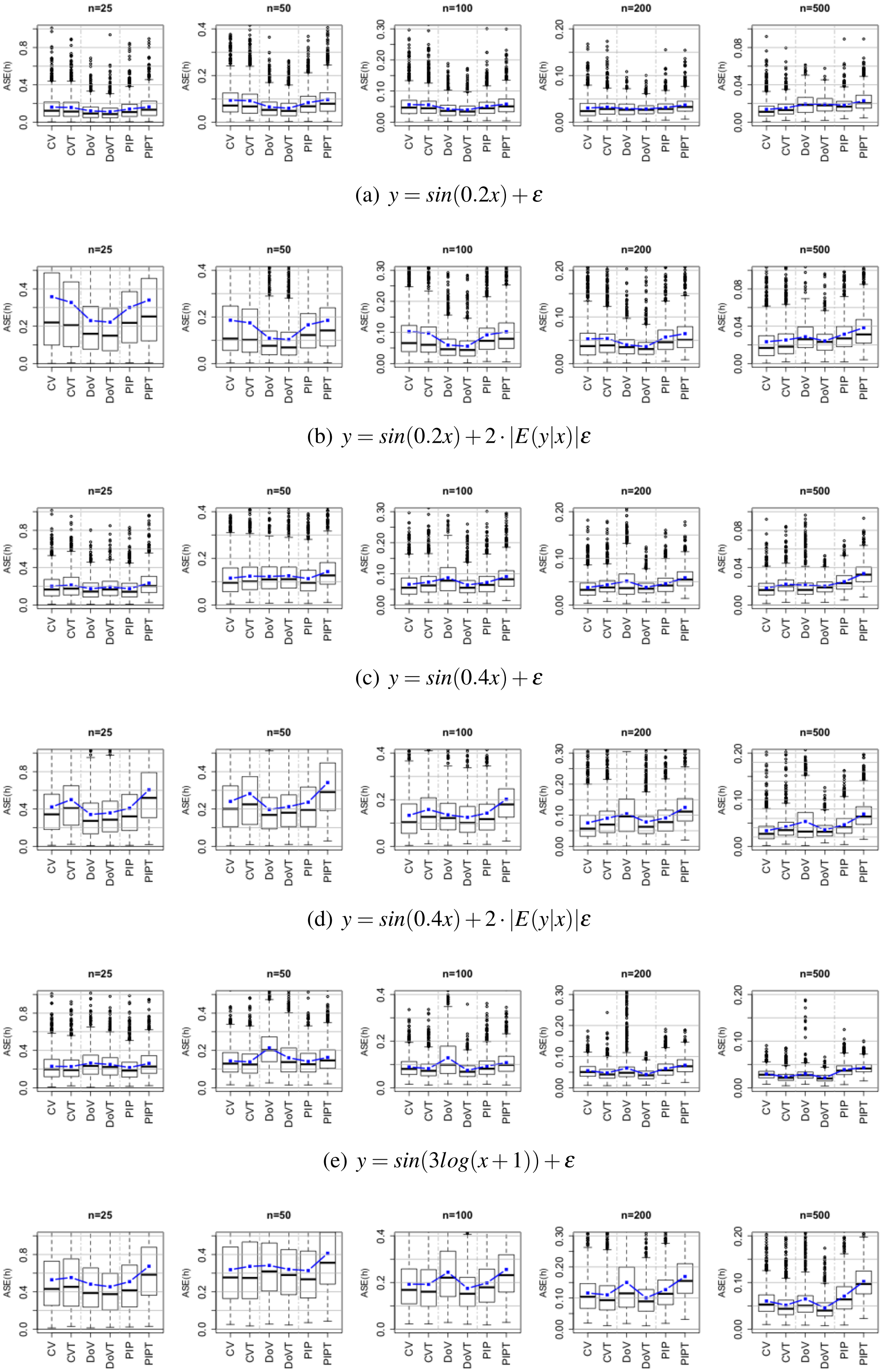
(c) $y = sin(0.4x) + \varepsilon$

(d) $y = sin(0.4x) + 2 \cdot |E(y|x)|\varepsilon$

(e) $y = sin(3log(x+1)) + \varepsilon$

(f) $y = sin(3log(x+1)) + 2 \cdot |E(y|x)|\varepsilon$

Figure 3.12: Box-plots and means (■) of the ASE-values for distribution $X \sim 0.5(lnN(0,1) + N(10,4))$ model (a) - (f)



(a) $y = sin(0.2x) + \varepsilon$

(b) $y = sin(0.2x) + 2 \cdot |E(y|x)|\varepsilon$

(c) $y = sin(0.4x) + \varepsilon$

(d) $y = sin(0.4x) + 2 \cdot |E(y|x)|\varepsilon$

(e) $y = sin(3log(x+1)) + \varepsilon$

(f) $y = sin(3log(x+1)) + 2 \cdot |E(y|x)|\varepsilon$

Figure 3.13: Box-plots and means (■) of the ASE-values for distribution $X \sim 1/3(lnN(0,1) + N(4,1) + N(10,4))$ model (a) - (f)

References

ADLER, D. (2011) Foreign Library Interface, http://cran.r-project.org

ADLER, D. AND PHILIPP, T. (2010) dyncall Library, http://dyncall.org

BOX, G.E.P. AND COX, D.R. (1964) An Analysis of Transformation, *Journal of the Royal Statistical Society, Series B* **26**: 211-246.

CARROLL, R.J. AND RUPPERT, D.(1984). Power Transformation When Fitting Theoretical Models to Data, *Journal of American Statistical Association* **79**: 321-328.

CLARK, R. M. (1977) Non-Parametric Estimation of a Smooth Regression Function, *Journal of the Royal Statistical Society, Series B* **39**: 107-113.

GASSER, T., MÜLLER, H.G.. Kernel Estimation of Regression Functions. *Smoothing techniques in curve estimation (Lecture Notes in Mathematics)*, **757**, 23-68.

HART, J. D. AND YI, S.(1998). One-Sided Cross Validation, *Journal of American Statistical Association* **93**: 620-631.

KÖHLER, M.; SCHINDLER, A. AND SPERLICH, S. (2011). A Review and Comparison of Bandwidth Selection Methods for Kernel Regression, *http://www.uni-goettingen.de/en/67061.html* **95**, Discussion paper.

MALLOWS, C.L. (1973) Some Comments on C_p, *Technometrics* **15**, 661-675.

MAMMEN, M., MARTÍNEZ-MIRANDA, M.D., NIELSEN, J.P. AND SPERLICH, S. (2011). Do-validation for Kernel Density Estimation, *Journal of the American Statistical Association, forthcoming.*

MARRON, J.S. (1986). Will the Art of Smoothing ever become a Science, *Function Estimates* (Contemporary Mathematics 59), Providence, RI: American Mathematical Society, pp. 169-178.

NYCHKA, D. AND RUPPERT, D. (1994). Nonparametric Transformations for Both Sides of a Regression Model, *Journal of the Royal Statistical Society, Series B* **57**: 519-532.

PARK, B.U.; KIM, W.C.; RUPPERT, D.; JONES, M.C.; SIGNORINI, D.F. AND KOHN, R. (1997). Simple Transformation Techniques for Improved Non-parametric Regression, *Scandinavian Journal of Statistics* **24**: 145-163.

RICE, J. (1984) Bandwith Choice for Nonparametric Regression, *The Annals of Statistics* **Vol. 12, No. 4**, 1215-1230.

RUPPERT, D.; SHEATHER, S.J. AND WAND, M.P. (1995). An Effective Bandwidth Selector for Local Least Squares Regression, *Journal of the American Statistical Association* **90(432)**: 1257-1270.

SHIBATA, R. (1981) An Optimal Selection of Regression Variables, *Biometrika* **68**, 45-54.

WANG, N. AND RUPPERT, D. (1995). Nonparametric Estimation of the Transformation in the Transform-Both-Sides Regression Model, *Journal of the American Statistical Association* **90(430)**: 522-534.

Chapter 4

Summary and Outlook

The main conclusions are already given at the end of each chapter. Obviously, bandwidth selection in nonparametric estimation is a very extensive field, where many authors have done research in it. In chapter 1 and chapter 2, our aim was to compare easy to implement estimation methods having a good performance with respect to the typically used error criteria. Further this methods do not need extra strong assumptions, the computation time is very low. Further, these two chapters are giving an overview over almost all existing methods in global bandwidth selection for the kernel estimation of densities of a random variable respectively univariate regression with the help of second order kernels. It can be seen, that the decision which method to use depends on the smoothness and the distribution of the data. Of course some methods are only appropriate for very specific cases or just not adequate for global bandwidth selection, but many methods are reasonable to use. However, practitioners often use the well known standard routines or just guess optimal bandwidth. In chapter 3, we tested if a special transformations can improve the estimation. From the conclusions we see, that in some cases there is an improvement and moreover, the estimation with transformation does never fail. But it remains to find out, if there are possible transformations, that generally doing better than an untransformed version.