Köhler, Max; Schindler, Anja; Sperlich, Stefan

Working Paper

A Review and Comparison of Bandwidth Selection Methods for Kernel Regression

Discussion Papers, No. 95

Provided in Cooperation with:
Courant Research Centre 'Poverty, Equity and Growth in Developing and Transition Countries', University of Göttingen

Suggested Citation: Köhler, Max; Schindler, Anja; Sperlich, Stefan (2011) : A Review and Comparison of Bandwidth Selection Methods for Kernel Regression, Discussion Papers, No. 95, Georg-August-Universität Göttingen, Courant Research Centre - Poverty, Equity and Growth (CRC-PEG), Göttingen

This Version is available at:
https://hdl.handle.net/10419/90468

Courant Research Centre

'Poverty, Equity and Growth in Developing and Transition Countries: Statistical Methods and Empirical Analysis'

Georg-August-Universität Göttingen
(founded in 1737)

Discussion Papers

No. 95

A Review and Comparison of Bandwidth Selection Methods for Kernel Regression

Max Köhler, Anja Schindler, Stefan Sperlich

September 2011

Wilhelm-Weber-Str. 2 · 37073 Goettingen · Germany
Phone: +49-(0)551-3914066 · Fax: +49-(0)551-3914059

Email: crc-peg@uni-goettingen.de Web: http://www.uni-goettingen.de/crc-peg

A Review and Comparison of Bandwidth Selection Methods for Kernel Regression

Max Köhler [*], Anja Schindler [†], Stefan Sperlich[‡]

September 8, 2011

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.[1]

AMS classification: 62G07, 62G05
Keywords: Kernel regression estimation, Bandwidth Selection, Plug-in, Cross Validation.

[*]Max Köhler, University of Goettingen, CRC Poverty, Equity and Growth, Göttingen, Germany
[†]University of Göttingen, Faculty of Economics, Göttingen, Germany
[‡]Université de Genéve, Département d'économétrie, Genéve, Switzerland

[1]The authors acknowledge financial support from the Spanish "Ministerio de Ciencia e Innovación" MTM2008-03010.

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

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

with i.i.d. random variables ε_i with mean zero and unit variance. Further, $\sigma^2(x) = var(Y|x)$ 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

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

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
$$\det \begin{pmatrix} S_{h,0}(x) & S_{h,1}(x) \\ S_{h,1}(x) & S_{h,2}(x) \end{pmatrix} = S_{h,0}(x)S_{h,2}(x) - (S_{h,1}(x))^2 \neq 0 \tag{3}$$
the minimizer of (2) is unique and given below.

- If $S_{h,0}(x)S_{h,2}(x) - (S_{h,1}(x))^2 = 0$ we distinguish between
  - $\diamond$ $x = X_k$ for a $k \in \{1,\ldots,n\}$ 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) is solved by $\beta_0 = Y_k$, and β_1 can be chosen arbitrarily.
  - $\diamond$ $x \neq X_k$ for all $k \in \{1,\ldots,n\}$. 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
$$\hat{m}_h(x) = \frac{1}{n}\sum_{i=1}^{n} W_{h,i}(x)Y_i \tag{4}$$
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 3 and Section 4 we introduce and discuss the various selection methods we could find in the literature, separately for the three different risk measures. In Section 5 we present in detail extensive simulation studies to compare all here discussed selection methods. Section 6 concludes.

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:

$$AMISE(\hat{m}_h(x) \mid X_1,\ldots,X_n) = \frac{1}{nh}||K||_2^2\sigma^2(b-a) + \frac{h^4}{4}\mu_2^2(K)\int_S (m''(x))^2 f(x)dx, \qquad (5)$$

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 (5). An obvious choice of for defining an optimal bandwidth is to say choose h such that (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 (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

$$ASE(h) = \frac{1}{n}\sum_{j=1}^{n} (\hat{m}_h(X_j) - m(X_j))^2\, w(X_j), \qquad (6)$$

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 1. For the sake of illustration we have to anticipate here some definitions given in detail at the beginning of our simulation Section 5. When we refer here and in the following illustrations of this section to certain models, for details please consult Section 5.

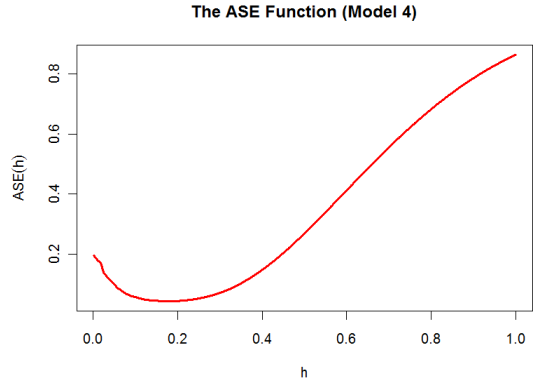

Figure 1: *ASE* with $w(X_j) = 1_{[X_6, X_{144}]}$ for $n = 150$ simulated data following Model 3

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

$$\lim_{n\to\infty} \frac{ASE(\hat{h})}{ASE(\hat{h}_0)} = 1 \tag{7}$$

almost surely. If (7) is fulfilled, it follows easily that

$$\frac{ASE(\hat{h})}{ASE(\hat{h}_0)} \xrightarrow{P} 1 \tag{8}$$

or nearly equivalently

$$\frac{\hat{h}}{\hat{h}_0} \xrightarrow{P} 1, \tag{9}$$

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.

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 (6) response Y_j for $m(X_j)$. This yields the substitution estimate

$$p(h) = \frac{1}{n}\sum_{j=1}^{n}(\hat{m}_h(X_j) - Y_j)^2 w(X_j). \tag{10}$$

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 (10) is the same as that of (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 (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.

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

$$G(h) = \frac{1}{n}\sum_{j=1}^{n}(Y_j - \hat{m}_h(X_j))^2 \, \Xi\left(\frac{1}{n}W_{h,j}(X_j)\right) w(X_j), \tag{11}$$

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

$$\Xi(u) = 1 + 2u + O(u^2) \ , \ u \to 0. \tag{12}$$

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)

$$\text{with} \qquad \Xi_S(u) = 1 + 2u \ . \tag{13}$$

- Generalized cross validation (GCV) $\hat{h}_{GCV} = \underset{h \in \mathbb{R}^+}{\operatorname{argmin}} \, G_{GCV}(h)$, see Craven and Wahba (1979)

$$\text{with} \qquad \Xi_{GCV}(u) = (1-u)^{-2} \ . \tag{14}$$

- Akaikes information criterion (AIC) $\hat{h}_{AIC} = \underset{h \in \mathbb{R}^+}{\operatorname{argmin}} \, G_{AIC}(h)$, see Akaike (1974)

$$\text{with} \qquad \Xi_{AIC}(u) = \exp(2u) \ . \tag{15}$$

- The finite prediction error (FPE) $\hat{h}_{FPE} = \underset{h\in\mathbb{R}^+}{\operatorname{argmin}}\, G_{FPE}(h)$, see Akaike (1970)

$$\text{with} \qquad \Xi_{FPE}(u) = \frac{1+u}{1-u}\,. \tag{16}$$

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

$$\text{with} \qquad \Xi_T(u) = (1-2u)^{-1}\,. \tag{17}$$

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 (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 (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 (11) can be evaluated,

$$h_{min,G} = \max\left\{h_{min,l+1}, \ldots, h_{min,n-l}\right\}. \tag{18}$$

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 (18).

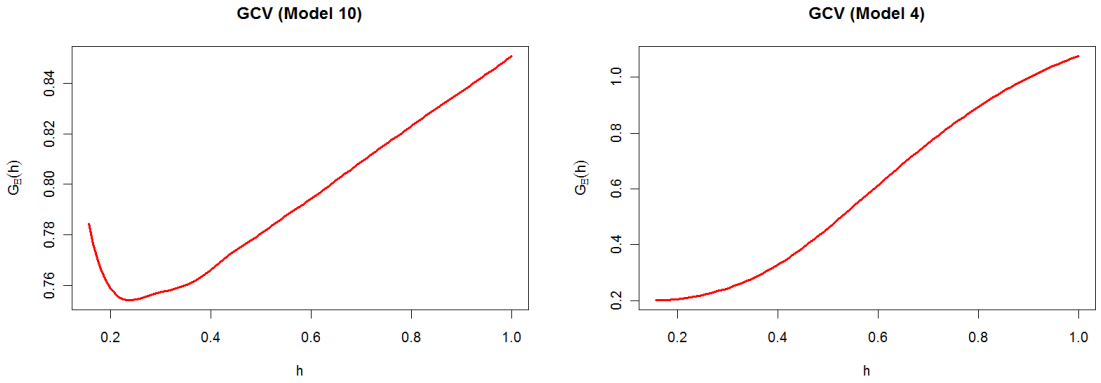

Figure 2: The Corrected ASE Functions for $n = 150$ independent data following Model 4 and Model 10, respectively.

Figure (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 5.

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

$$CV(h) = \frac{1}{n}\sum_{j=1}^{n}(Y_j - \hat{m}_{h,-j}(X_j))^2 w(X_j) \ , \tag{19}$$

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:

$$\hat{h}_{CV} = \underset{h\in\mathbb{R}^+}{\operatorname{argmin}}\, CV(h) \ . \tag{20}$$

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 $h < h_{min,j} := \min\left\{X_j - X_{j-1}, X_{j+1} - X_j\right\}$ 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, (19) is only defined if we fix $h > h_{min,CV}$ with

$$h_{min,CV} := \max\left\{h_{min,l+1},\dots,h_{min,n-l}\right\} \ . \tag{21}$$

Although this mitigates the problems at the lower bound of the bandwidth scale (i.e. for bandwidth

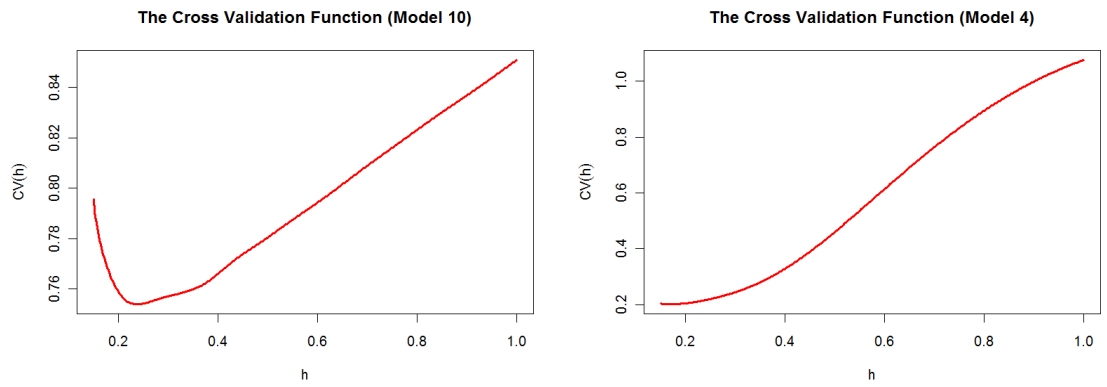

Figure 3: The CV functions for $n = 150$ simulated data following Model 4 and Model 10, respectively.

approaching zero), Figure 3 exhibits similar problems for the CV as we saw them for the corrected ASE criteria. Figure 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.

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

$$OSCV(b) = \frac{1}{n-2l} \sum_{i=l+1}^{n-l} (\tilde{m}_b^{-i}(X_i) - Y_i)^2, \tag{22}$$

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 (22) at least $l-1$ data points can be used. Denote the minimizer of (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, (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 (22) is approximately the same as minimizing

$$E\left[\frac{1}{n-2l} \sum_{i=l+1}^{n-l} (\tilde{m}_b(X_i) - m(X_i))^2\right]. \tag{23}$$

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 (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

$$\hat{h}_{OSCV} = C \cdot \hat{b}. \tag{24}$$

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 1, see also Figure 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.

As for the corrected ASE and CV bandwidth selection rules, the OSCV bandwidth selection rule is consistent. Now consider the i'th summand of (22). Analogously to prior discussions, (22) 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 (22)

Table 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: 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

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

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

this problem is much less serious for the OSCV than for the other methods. Due to the fact that

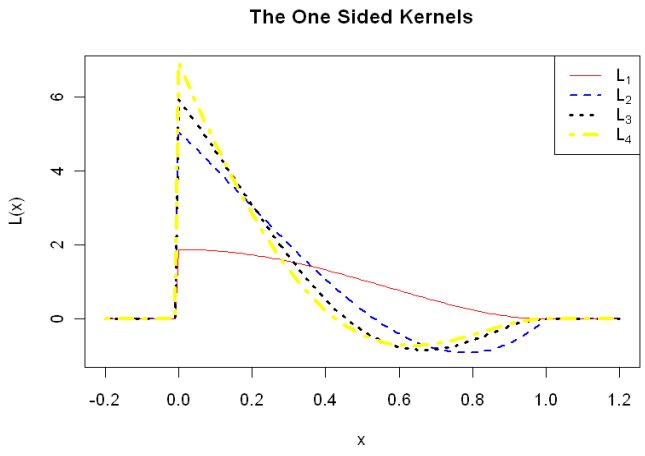

Figure 4: The One Sided Selection Kernels used for left OSCV.

$\tilde{m}_b(x)$ 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 (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 (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.

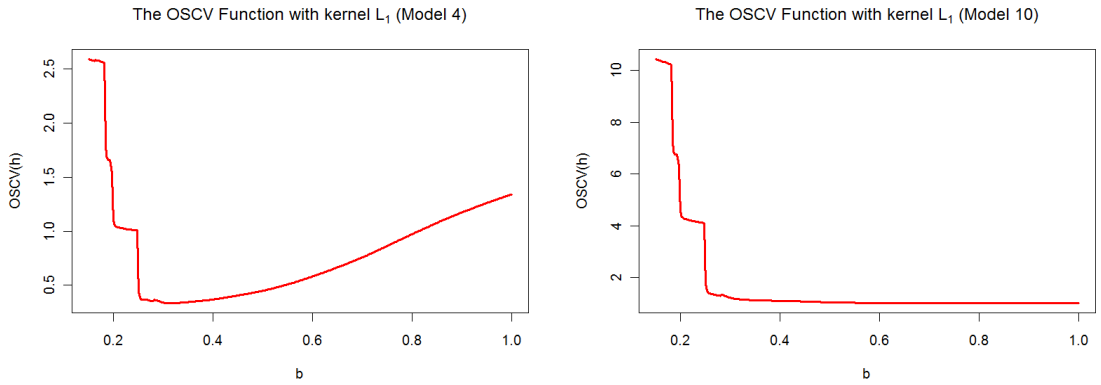

Figure 5: The OSCV Functions based on 150 independent data (X_i, Y_i).

Note that the regression estimator used at the bandwidth selection stage, namely $\tilde{m}_b(x)$ in (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 (24) does not change. There is evidence that the difference of 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.

$$ARE(K,L) = \lim_{n\to\infty} \frac{E((\hat{h}_{OSCV} - \hat{h}_0)^2)}{E((\hat{h}_{CV} - \hat{h}_0)^2)} \tag{25}$$

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 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

$$\widehat{ARE}(K,L)=\frac{\sum_{i=1}^{30}((\hat{h}_{OSCV})_i-(\hat{h}_0)_i)^2}{\sum_{i=1}^{30}((\hat{h}_{CV})_i-(\hat{h}_0)_i)^2}. \tag{26}$$

The results in the case of $n=150$ are given in Table 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 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 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-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.

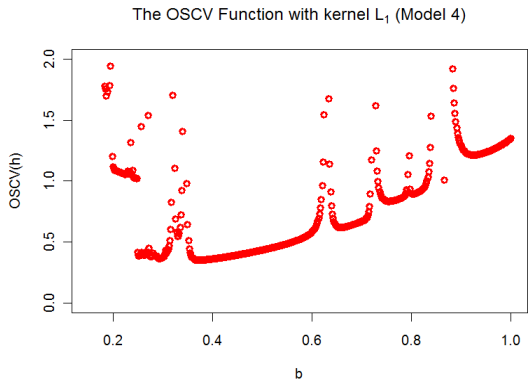

Figure 6: The left OSCV function using kernel L_4.

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

$$\hat{m}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) Y_i \tag{27}$$

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 (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 (9).

Härdle and Marron (1985) considered the Nadaraya-Watson estimator in a multivariate random design setting. They proved the optimality condition (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 $L > 0$, $\int ||u||^{\xi} |K(u)| \, du < \infty$, the Hölder continuity of f and m and that the density f is bounded from below and compactly supported.

If conditions (8) and (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 $n \to \infty$,

$$n^{3/10}(\hat{h} - \hat{h}_0) \xrightarrow{\mathscr{L}} N(0, \sigma^2) \tag{28}$$

and

$$n^{3/10}(ASE(\hat{h}) - ASE(\hat{h}_0)) \xrightarrow{\mathscr{L}} C\chi_1^2 \tag{29}$$

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 (28) is of order $n^{1/10}$ and condition (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

$$n^{3/10}(\hat{h} - h_0) \approx \sum_{j=1}^{\lfloor n/2 \rfloor} (V_j - 2) w_K(j), \tag{30}$$

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

$$n^{3/10}(\hat{h}_{CV} - h_0) \approx \sum_{j=1}^{\lfloor n/2 \rfloor} (V_j - 2) w_K(j). \tag{31}$$

When Hart and Yi (1998) computed the first twenty weights $w_K(j)$ $(j = 1, 2, \ldots, 20)$ and for the quartic kernel K and $n = 100$, they observed that $w_K(1)$ and $w_K(2)$ are large and negative but $w_K(3), \ldots, w_K(20)$ 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 (31) for OSCV, namely

$$n^{3/10}(\hat{h}_{OSCV} - h_0) \approx \sum_{j=1}^{\lfloor n/2 \rfloor} (V_j - 2) w_L(j). \tag{32}$$

When they calculated the weights $w_L(j)$ $(j = 1, 2, \ldots, 20)$ in (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,

$$\lim_{n\to\infty} E((n^{3/10}(\hat{h}_{OSCV} - h_0))^2) = C^2 V(L), \tag{33}$$

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 (33) is of order $n^{1/5}$.

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 (5). Minimizing w.r.t. h, leads to the AMISE-optimal bandwidth (h_{AMISE}), given by:

$$h_{AMISE} = \left(\frac{||K||_2^2 \cdot \int_S \sigma^2(x)\, dx}{\mu_2^2(K) \cdot \int_S (m''(x))^2 f(x) dx \cdot n} \right)^{1/5}, \tag{34}$$

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:

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

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 smooth bootstrap (giving h_{SB}), again the unknown expressions in (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 $(1, 2, \ldots, N_{max})$ 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 (35), see Subsection 4.2.

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 (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

$$\begin{aligned}\hat{\sigma}_Q^2(N) &= \frac{1}{n-5N}\sum_{i=1}^{n}\sum_{j=1}^{N}(Y_i-\widehat{m}_{Q_j}(X_i))^2\mathbf{1}_{\{X_i\in\chi_j\}} \\ &= \frac{1}{n-5N}\sum_{i=(j-1)n/N+1}^{jn/N}\sum_{j=1}^{N}\left(y_i-\hat{\beta}_{0j}+\hat{\beta}_{1j}x_i-\hat{\beta}_{2j}x_i^2-\hat{\beta}_{3j}x_i^3-\hat{\beta}_{4j}x_i^4\right)^2 \qquad (36)\end{aligned}$$

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.

4.2 Direct plug-in bandwidth selection

In this approach the unknown quantities in (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

$$\hat{\theta}_{22}(g)=n^{-1}\sum_{i=1}^{n}\left[\hat{m}_g^{(2)}(X_i)\right]^2, \qquad (37)$$

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.

$$\hat{\theta}_{22}^{\alpha}(g_{AMSE})=\frac{1}{n}\sum_{i=1}^{n}\left[\hat{m}^{(2)}(X_i)\right]^2\mathbf{1}_{\{(1-\alpha)a+\alpha b<X_i<\alpha a+(1-\alpha)b\}}, \qquad (38)$$

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

$$g_{AMSE}=\left[C_2(K)\frac{\sigma^2\cdot(b-a)}{|\theta_{24}|n}\right]^{1/7} \qquad (39)$$

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 (36)). For θ_{24} we use:

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

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

The nonparametric estimator for σ^2 is:

$$\hat{\sigma}^2 = \nu^{-1}\sum_{i=1}^{n}\left[Y_i-\hat{m}_{\lambda_{AMSE}}(X_i)\right]^2, \tag{40}$$

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:

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

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 smooth 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

$$\hat{\sigma}_g^2(x) = \frac{1}{n}\sum_{i=1}^{n} W_{gi} Y_i^2 - [\hat{m}_g(x)]^2 = Var^*(Y^* \mid X^* = x). \tag{41}$$

For the estimation of $\hat{\sigma}$ assuming homoscedasticity, we average (41) over $x = X_i^*$. Additionally, a nonparametric estimator for θ_{22} is calculated as in formula (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 (35) to calculate a smooth bootstrap bandwidth $\hat{h}_{SB}$ which is certainly still a function of the pilot bandwidth.

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

$$MASE(g,h) = \frac{1}{n}\sum_{i=1}^{n} MSE_{h,g}^*(X_i) \tag{42}$$

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 (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.

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 $\hat{h}_{PDI}$

$$\frac{\hat{h}_{DPI} - h_{MISE}}{h_{MISE}} \xrightarrow{P} D, \tag{43}$$

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 (39), (40) with

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

conditional on $X_1, \ldots, X_n$. 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 (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 (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 smooth 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}.

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:

$$Mix_{method1,method2}(\alpha) = \alpha \cdot \hat{h}_{method1} + (1-\alpha) \cdot \hat{h}_{method2}, \tag{44}$$

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.

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 $\varepsilon \sim \mathcal{N}(0,1)$ 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 $X_i \sim 1/2 \cdot \mathcal{N}(-0.6, 1/4) + 1/2 \cdot \mathcal{N}(0.3, 1/3)$. 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

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: smooth 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 L_2 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 L_2 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 3 appear by using the local linear estimator $\hat{m}_h(x)$. Hence, the correction of the bandwidth grid, given in (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 smooth bootstrap. The direct plug-in and the smooth 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.

5.1 Comparison of the bias and L_1-distance for the different bandwidths (m_5, m_7)

Most of our numerical findings have been summarized in two figures: In Figure 7 we show the biases (m_5) and in Figure 8 the $L_1(h)$-distances (m_7) for all methods and models, but only for sample sizes $n = 25$ and $n = 200$.

We first summarize the behavior of CV and GCV since they behave almost identically. For the standard normal distribution (see right panel in Figure 7), they are oversmoothing for all cases.

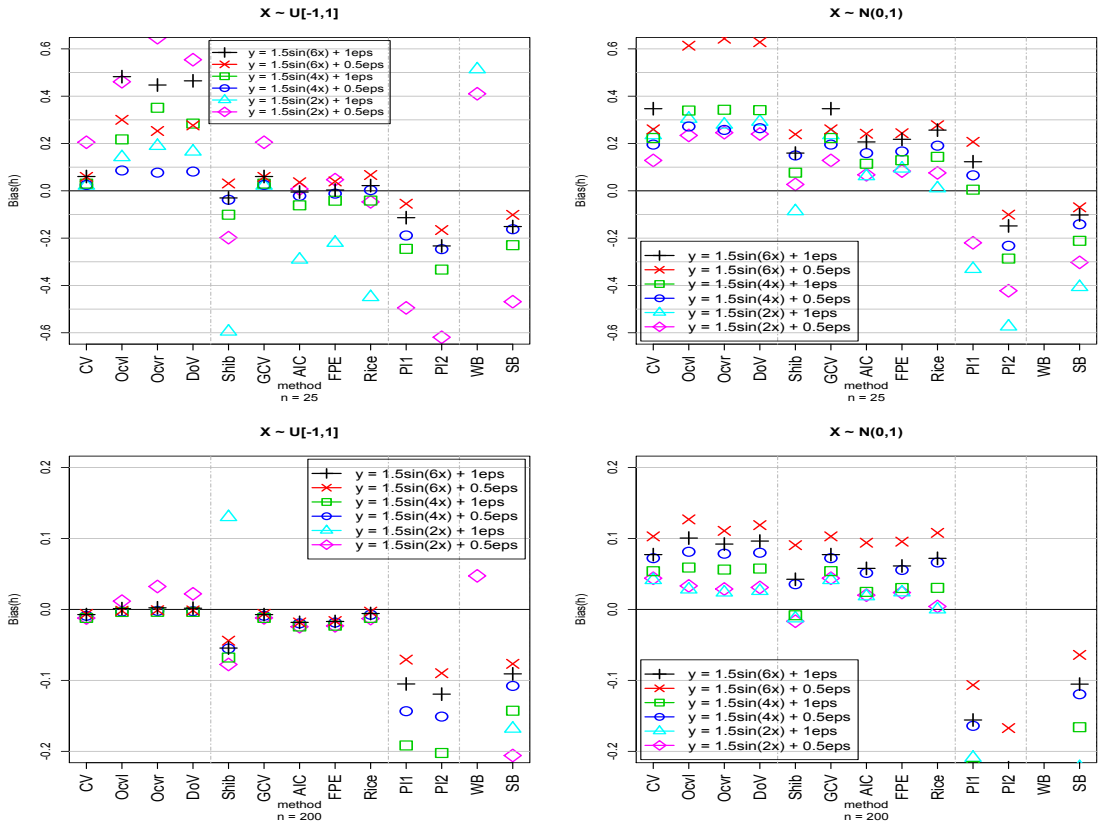

Figure 7: Comparison of the bias for sample sizes $n = 25$ (above) and $n = 200$ (below)

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 L_1-distances are relatively small for all models, see Figure 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 L_1-distance is almost zero, see also Figure 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 7). For the uniform design the three methods are almost always undersmoothing but in general show a good performance respective to the 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 8). In comparison AIC, FPE and Rice seem to benefit less from increasing sample sizes, i.e. although the bias respectively the $L_1(h)$-distance is getting smaller in absolute value it is not getting smaller in the same magnitude like CV, GCV and the

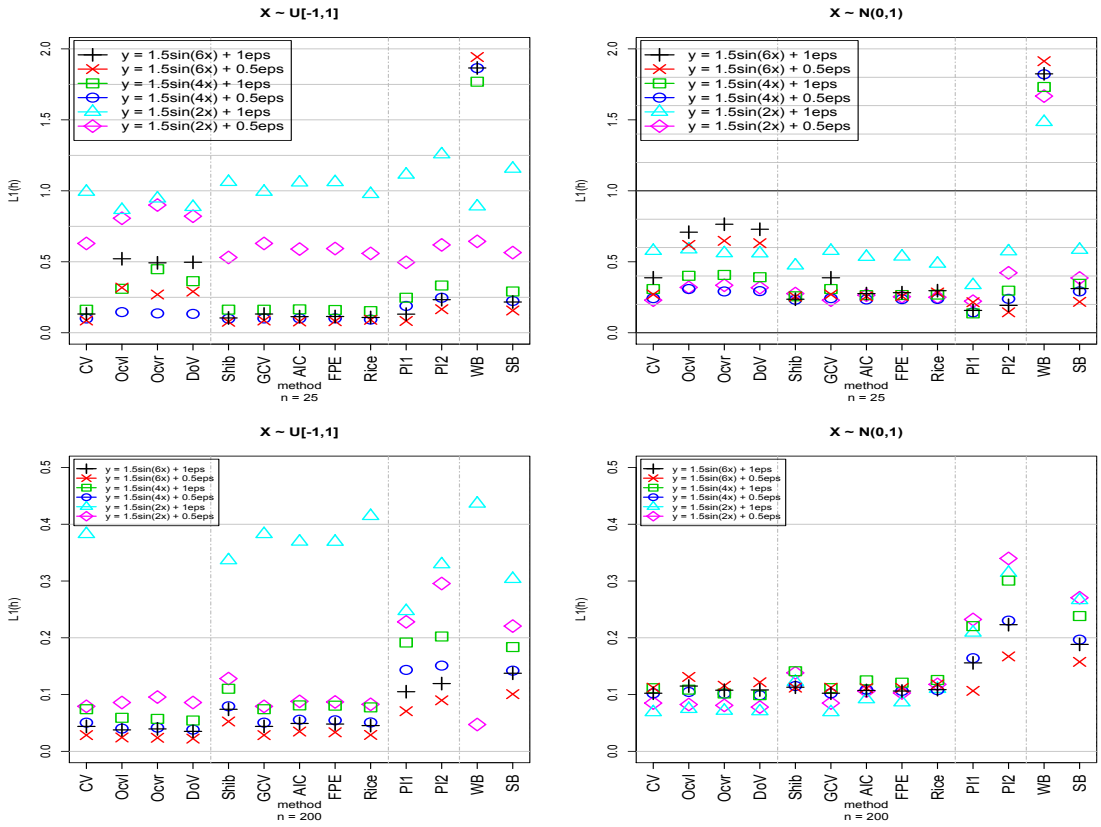

Figure 8: Comparison of the L_1-distance for $n = 25$ (above) and $n = 200$ (below)

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 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 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 $L_1(h)$-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.

5.2 Comparison of L_1 and L_2-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 7 and Figure 8). All in all we provide here the comparison of the selection methods along models 1, 2, 9 and 10. In Figure 9 we have plotted the resulting $L_1(h)$, and in Figure 10 the $L_2(h)$. For each of the four models we show the values for all sample sizes, i.e. for $n = 25, 50, 100, 200$.

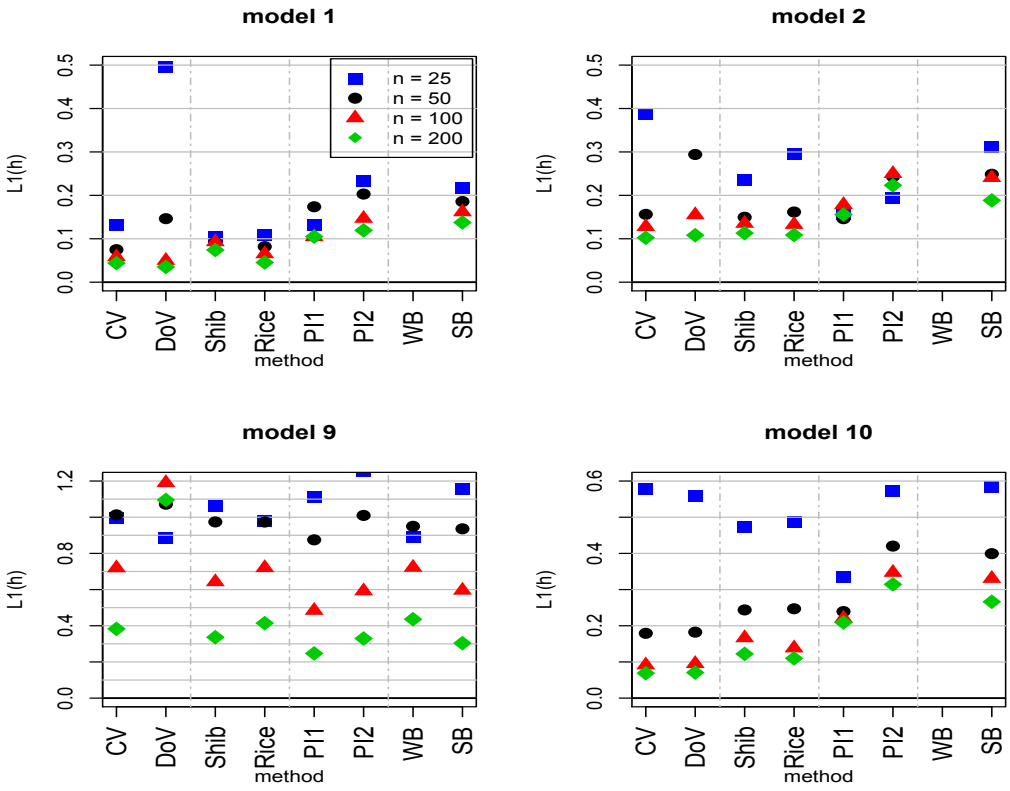

Figure 9: L1(h) for each four models varying the sample size

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

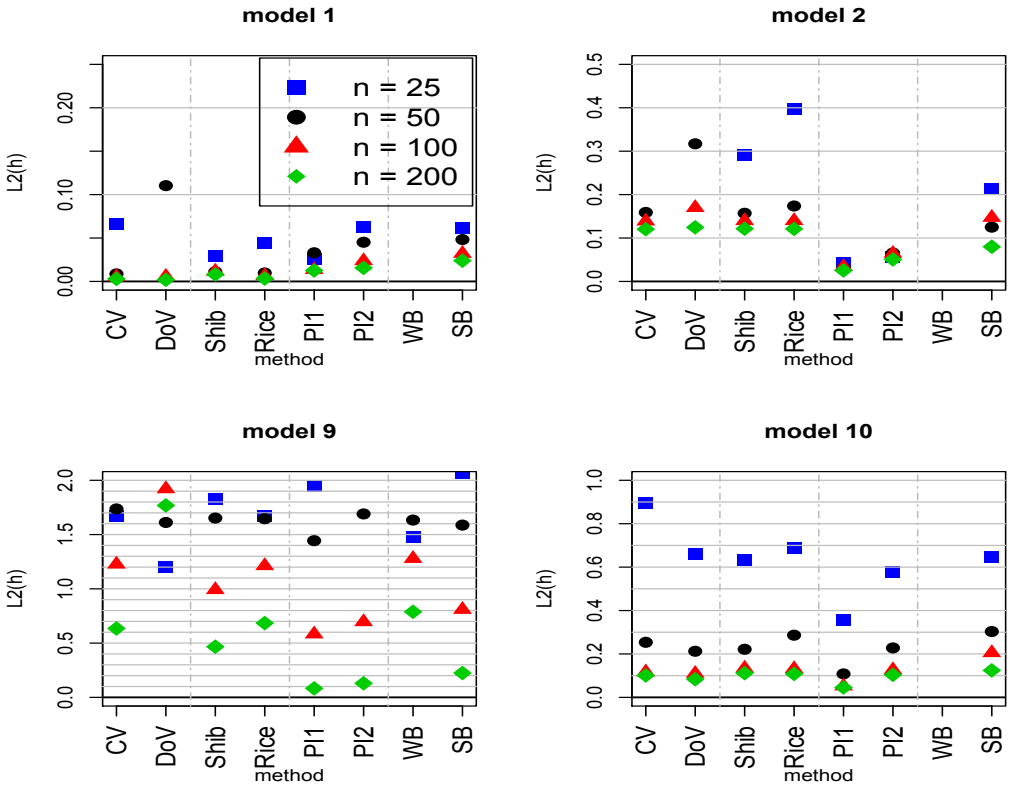

Figure 10: L2(h) for each four models varying the sample size

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.

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.

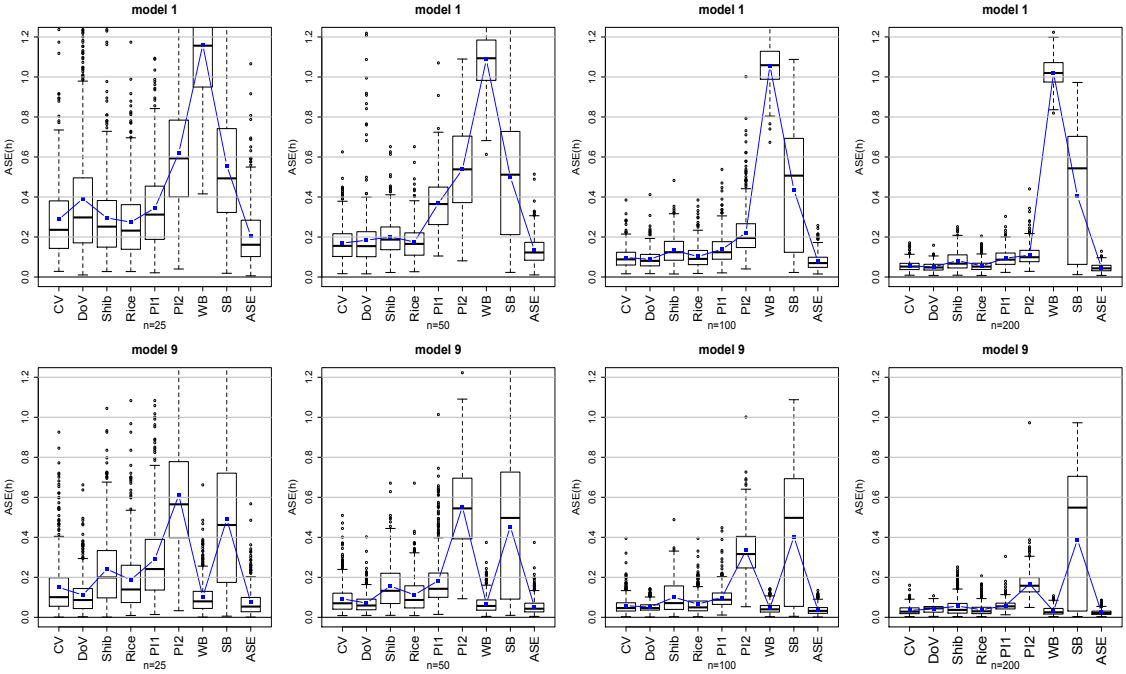

Figure 11: ASE-values for $X \sim U[-1,1]$ for all sample sizes

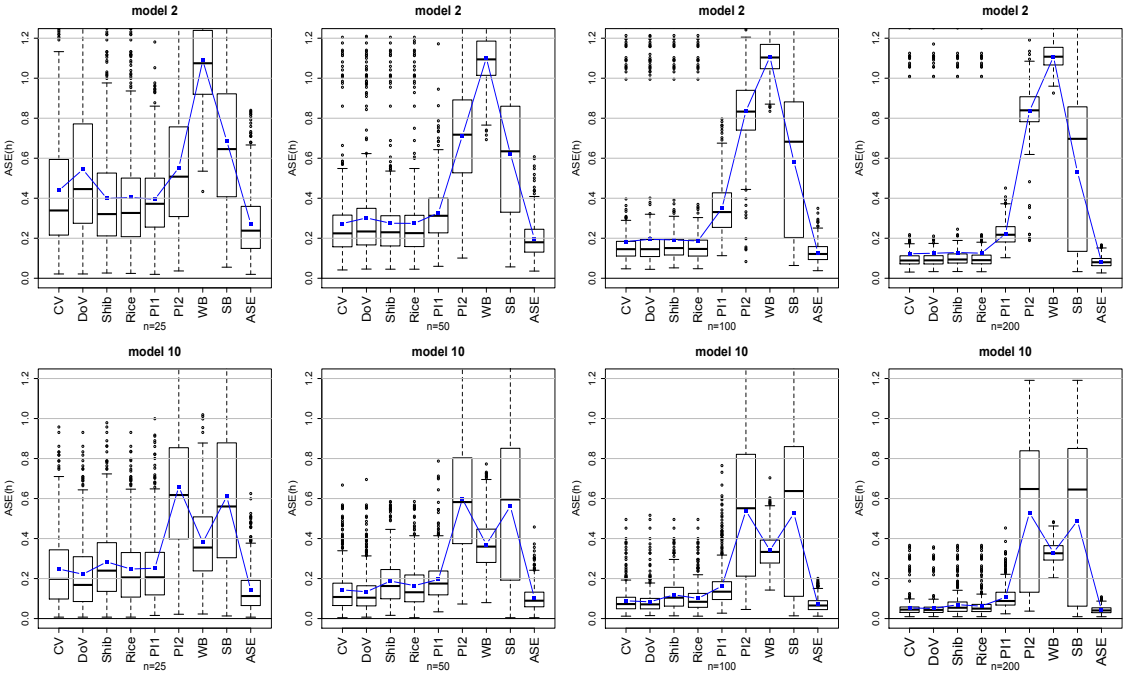

Figure 12: ASE-values for $X \sim N(0,1)$ for all sample sizes

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 11) and standard normal design (see figure 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 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 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 12, first row. The most interesting changes, compared to model 1, occur in the first four methods. There we have more 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 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.

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

If we look at Figures 13 and 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

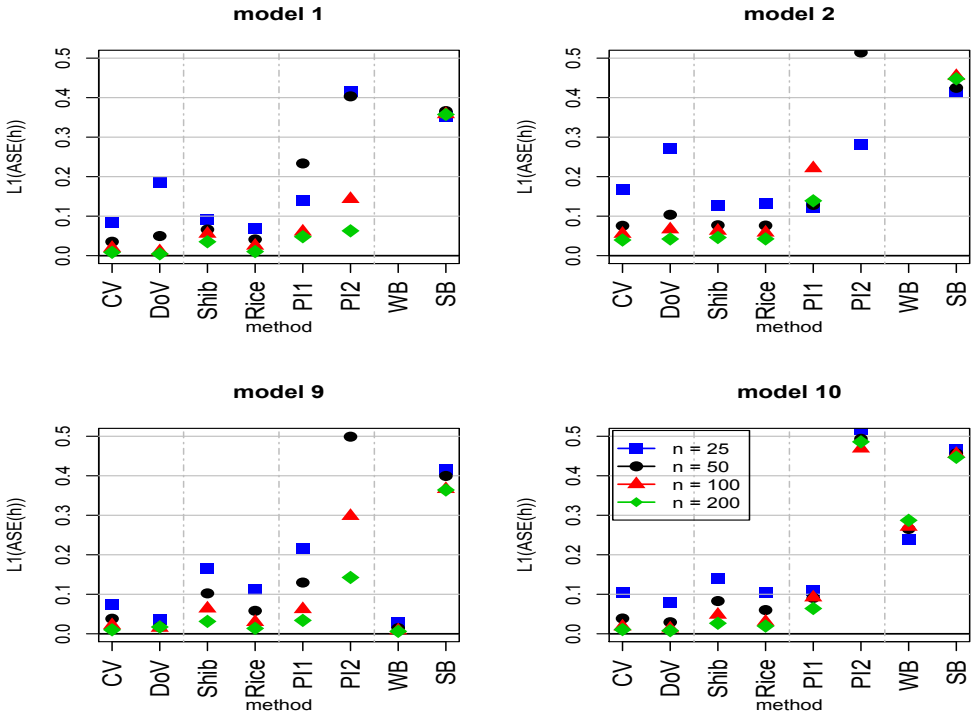

Figure 13: $L_1(ASE)$ for each four models varying the sample size

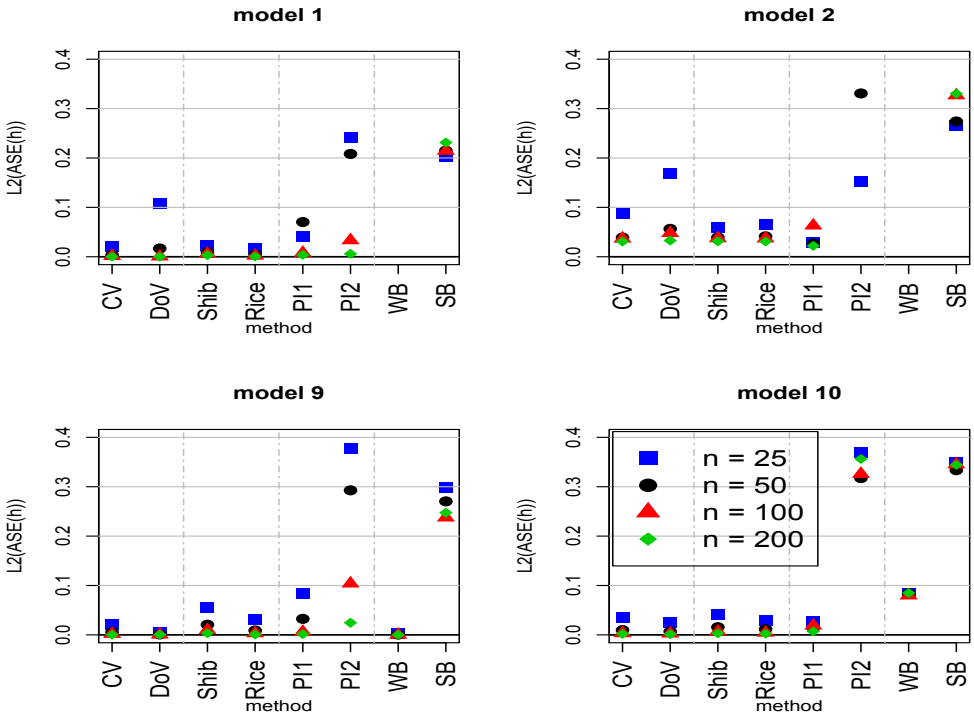

Figure 14: $L_2(ASE)$ for each four models varying the sample size

2 all L_1 and L_2-distances for WB are out of range. For this model PI1 is the best method 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 13, 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 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.

5.5 Comparison of different mixtures

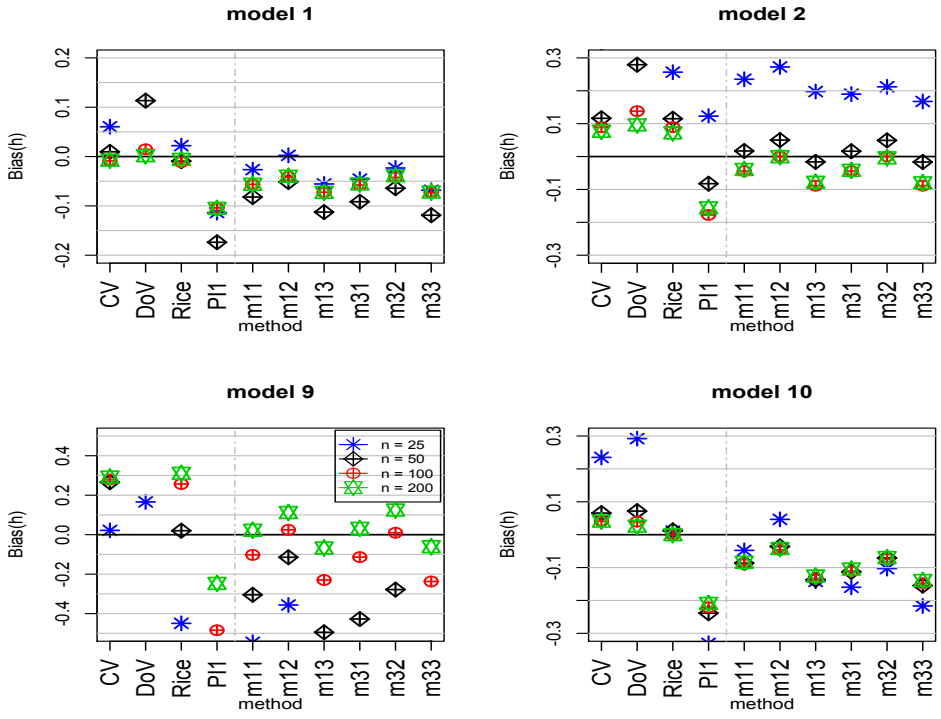

Figure 15: bias(h)

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 bandwidth. Recall that CV and corrected ASE methods tend to oversmooth while the PI and

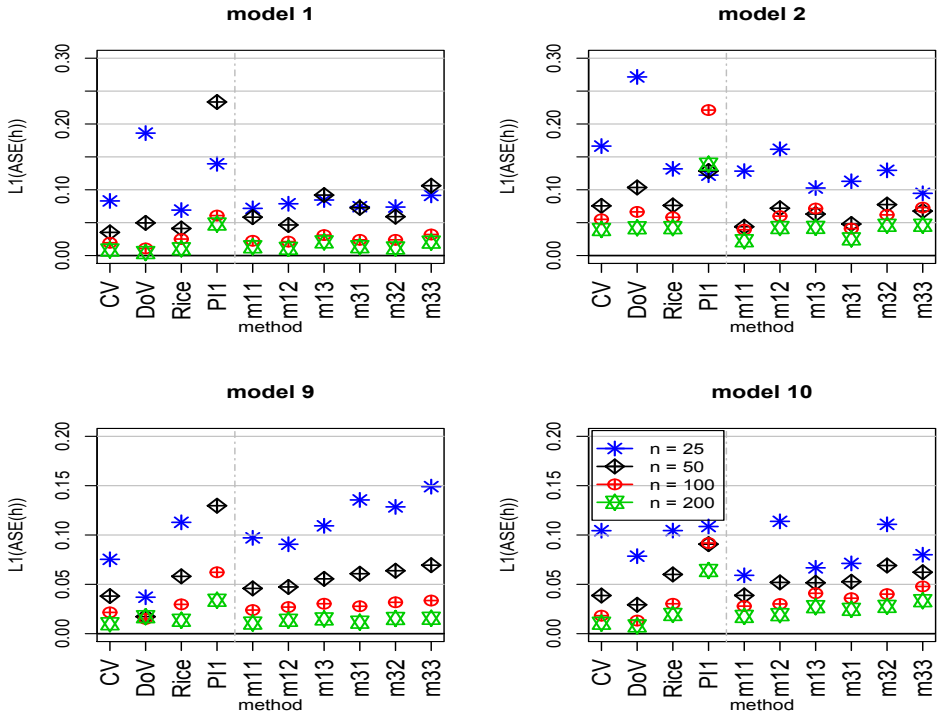

Figure 16: L1(ASE)

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 (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: $Mix_{CV,PI1}(1/2)$	3 m13: $Mix_{CV,PI1}(1/3)$	5 m22: $Mix_{Rice,PI1}(2/3)$
2 m12: $Mix_{CV,PI1}(2/3)$	4 m21: $Mix_{Rice,PI1}(1/2)$	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 15 and 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 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 16. In conclusion we can say, that the additional effort of mixing different methods seems not to be justifiable.

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 regression. 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 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 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 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., MARTINEZ-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.