



Bandwidth selection for a class of difference-based variance estimators in the nonparametric regression: A possible approach

M. Levine*

Department of Statistics, Purdue University, West Lafayette, IN 47907, USA

Received 20 April 2005; received in revised form 1 August 2005; accepted 1 August 2005

Available online 7 September 2005

Abstract

A possible approach to bandwidth selection for difference-based variance estimators in the nonparametric regression is proposed. The approach is based on the crossvalidation-type idea modified for the case of correlated data. The method is compared to an alternative plug-in style method. Difficulties in implementing the latter are highlighted and it is argued that the proposed method represents a better alternative. Simulations for several models are given that illustrate good practical performance of the proposed method.

© 2005 Elsevier B.V. All rights reserved.

Keywords: Bandwidth selection; Crossvalidation; Correlated data

1. Introduction

The model we consider in this article is the nonparametric regression model

$$y_i = g(x_i) + \sqrt{f(x_i)}\varepsilon_i, \quad i = 1, \dots, n, \quad (1)$$

where $g(x)$ and $f(x)$ are unknown mean and variance functions, respectively. It is assumed that $f(x) \in C^p$, $p \geq 2$, belong to some functional smoothness classes. The errors ε_i ,

* Tel.: +1 765 496 7571; fax: +1 765 494 0558.

E-mail address: mlevins@stat.purdue.edu.

$i = 1, \dots, n$ are independent standard normal random variables $N(0, 1)$. For convenience purposes, we further assume that $x \in [0, 1]$ and the (fixed) design is equispaced, t.i. $x_i = \frac{i}{n}$. Our problem of interest is estimating the variance function $f(x)$ in the presence of the (infinite dimensional) nuisance parameter $g(x)$.

It was pointed out in the late 1980s in [Carroll and Ruppert \(1988\)](#) that variance (and not only the mean) estimation is also an important problem. As a practical example, we may need to construct confidence(prediction) intervals for the mean function or to analyze immunoassay (building prediction and calibration intervals)—either task will require an estimate of the variance. At first, only the mean was assumed to be functionally dependent on the predictor variable and the error variance was assumed to be constant. A few of the articles that dealt with estimating the constant variance are [Rice \(1984\)](#), [Gasser et al. \(1986\)](#) and [Hall et al. \(1990\)](#). The heteroscedastic case is a fairly recent research topic. Some notable publications dedicated to it are [Müller and Stadtmüller \(1993\)](#), [Härdle and Tsybakov \(1997\)](#), [Fan and Yao \(1998\)](#) and [Dette \(2002\)](#). The functional nature of the variance means that different local estimation techniques, such as local polynomial estimation, are needed. The author in his dissertation (see [Levine, 2003](#)) proposed a class of Nadaraya–Watson-type local kernel estimators of the variance function.

It is well known that for any kernel estimator to become a workable procedure it has to be accompanied by a bandwidth selection procedure. The traditional application field for the kernel estimation—estimation of the mean function in nonparametric regression—contains many different bandwidth selection procedures.

Two most common classes of those procedures are

- plug-in-type procedures,
- data-driven procedures that minimize some estimator of the mean squared error (MSE).

However, the same problem with respect to the kernel variance estimators has received scant attention so far.

In this article, we work with the variance estimation method described in [Levine \(2003\)](#). First, we briefly describe the method. Next, we introduce a suitable crossvalidation-type method for use with this procedure. The next step is investigating its properties in Monte-Carlo settings. We will also show why an alternative approach via a plug-in-type procedure does not seem to be a good option for our variance estimation approach although it is very attractive conceptually. As a matter of fact, there are other possible nonconstant variance estimation procedures that have been suggested in recent years. We can mention, among others, the approach in [Fan and Yao \(1998\)](#) that is based on smoothing squared residuals from the local linear fit of the mean function. [Neumann \(1994\)](#) uses a similar idea, only in place of local linear regression the Gasser–Müller kernel estimation (see, e.g., [Gasser and Müller, 1979](#)) is used. A quite general method, based on the localized quadratic forms in observations, has been considered in [Müller and Stadtmüller \(1993\)](#). As for the bandwidth selection procedures, [Fan and Yao \(1998\)](#) uses the bandwidth selection method from [Fan and Gijbels \(1995\)](#); as possible alternatives, the crossvalidation bandwidth rule or the plug-in approach of [Ruppert et al. \(1995\)](#) are also mentioned. [Neumann \(1994\)](#) use a simple “leave-one-out” crossvalidation. Note that those are not methods designed specifically for

the bandwidth selection in the variance estimation context: since the problem of variance estimation in both of the aforementioned articles is being reduced to the local linear regression, the methods used are the classical ones designed in the local polynomial regression setting.

2. Method description

Once again, we consider model (1). What follows is a brief review of the method to be used to estimate the variance. For a more detailed treatment of this subject please see [Levine \(2003\)](#).

Definition 1. A difference sequence of order r is a sequence of real numbers d_i , $i=0, \dots, r$ such that its elements sum up to zero:

$$\sum_{i=0}^r d_i = 0 \quad (2)$$

while the sum of squares is 1

$$\sum_{i=0}^r d_i^2 = 1. \quad (3)$$

The first step is to construct the building blocks for the variance estimator. We call them pseudoresiduals of order r and use the following definition:

Definition 2. A pseudoresidual of order r is

$$\Delta_{r,i} = \sum_{j=0}^{r-1} d_j y_{j+i}, \quad (4)$$

where $\{d_i\}_{i=1}^r$ is a difference sequence as defined in (2) and (3).

Example. The (unique) difference sequence of order 2 is $\left\{\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right\}$. The corresponding pseudoresidual is $\Delta_i = (y_i - y_{i-1})/\sqrt{2}$.

Remark 3. It is also possible to define the pseudoresidual Δ_i as being symmetric around y_i :

$$\Delta_{r,i} = \sum_{j=0}^{r-1} d_j y_{i - \left[\frac{r+1}{2}\right] + j + 1}. \quad (5)$$

Note that this choice does not influence any asymptotic results.

Next, we define the variance estimator as the weighted local average of the squared pseudoresiduals of order r . Each pseudoresidual is a normalized difference of $r + 1$ observations;

it can be viewed as the rough first-step estimate of the variance function $f(x)$. If this view is adopted, the next step of taking a weighted local average is nothing else but smoothing of several rough estimates to produce a more precise one. The following definition should provide a clear insight into the nature of this estimator.

Definition 4. Let us assume that the kernel function $K(\cdot)$ is a proper density on $[-1, 1]$. Then, the variance estimator of order r is defined as

$$\hat{f}_h(x) = \frac{\sum_{i=1}^{n-r} A_{r,i}^2 K\left(\frac{x-x_i}{h}\right)}{\sum_{i=1}^{n-r} K\left(\frac{x-x_i}{h}\right)}. \quad (6)$$

Remark 5. By definition, any kernel function $K(\cdot)$ must satisfy $\int K(u) du = 1$ but it need not be the proper density. (see, e.g. Härdle, 1990). We can also define (6) with this more general kernel in mind. In such a case, the “best” bandwidth has to be defined in the minimax sense over a certain class of variance functions \mathcal{F} and the rates of convergence are then different from those we obtain. The functional class \mathcal{F} has to be wide enough to contain the set of p times continuously differentiable functions we use in this article. For example, we can assume that $\mathcal{F} = W_2^p$ —an L_2 -based Sobolev space.

Remark 6. This estimator is conceptually equivalent to a NW kernel estimator though of course $A_{r,i}^2$ are not independent.

Various properties of this estimator were derived in Levine (2003) including closed form expressions of the asymptotically optimal bandwidth and integrated mean squared error (MISE). Ignoring the higher-order bias terms that depend on r , Levine (2003) showed that for fixed r , $h \rightarrow 0$, $n \rightarrow \infty$ and $nh \rightarrow \infty$ the asymptotically optimal (in the sense of Parzen (1962) and Rosenblatt (1956)) bandwidth $h \sim O(n^{-1/(2p+1)})$ while $MISE \sim O(n^{-2p/(2p+1)})$. These rates of convergence are the same as those derived in Müller and Stadtmüller (1993) for similarly defined variance estimator class; however, compared to Müller and Stadtmüller (1993), Levine (2003) also derives its asymptotic distribution. Another possible comparison benchmark here may be Fan and Yao (1998), which bases the variance estimation on the squared residuals from the estimation of the mean function $g(x)$. If we look at the higher-order asymptotics, it becomes clear that this estimator class demonstrates superior to Fan and Yao (1998) second-order asymptotic properties. To be more specific, let us say that $p = 2$. Then, while the contribution to MSE from estimating the mean function $g(x)$ is of the order $O(n^{-4/5})$ in Fan and Yao (1998), it has the order $O(n^{-2})$ in Levine (2003).

Unfortunately, these results do not provide us with a bandwidth selection algorithm which is an obvious priority before the method can be used in practice. It is worth noting that, as far as we know, there are no results on the bandwidth selection for the difference-sequence-type estimators (6). When estimating the variance function as the (normalized) sum of squared residuals, the options considered in the literature are those in Ruppert et al. (1997) where some bandwidth selector (e.g. the one from Fan and Gijbels, 1995) is used to estimate both

the mean function and the variance. Fan and Yao (1998) also suggested several methods; however, this was not the focus of the paper. The fairly limited literature on the difference sequence-type variance estimators (e.g. Müller and Stadtmüller, 1993) has not considered the bandwidth selection so far.

3. Plug-in approach to bandwidth selection

Let us assume that $g(x) \in C^1[0, 1]$ while $f(x) \in C^2[0, 1]$. The kernel function $K(x) : [-1, 1] \rightarrow \mathcal{R}$ is a second-order kernel function; in other words, $\int K(u) du = 1$, $\int u K(u) du = 0$ and $\int u^2 K(u) du \equiv \sigma_K^2 \neq 0$. Also, $R_K \equiv \int K^2(u) du$. From Levine (2003), we know that the exact optimal bandwidth is

$$h = n^{-1/5} \left[\frac{C R_K \int f^2(x) dx}{2\sigma_K^4 \int [f''(x)]^2 dx} \right]^{1/5}, \quad (7)$$

where

$$C = 2 \left(2 \sum_{k=1}^r \left(\sum_{j=0}^{r-k} d_j d_{j+k} \right)^2 + 1 \right)$$

is the constant that depends on the chosen difference sequence $\{d_i\}$. (see Levine (2003)). Expression (7) contains quadratic functionals of the unknown variance function, namely $\int f^2(x) dx$ and $\int [f''(x)]^2 dx$. A common notation for them is

$$R(f) \doteq \int f^2(x) dx \quad (8)$$

and

$$R(f'') \doteq \int [f''(x)]^2 dx \quad (9)$$

(see, for example, Wand and Jones, 1995). Then, if some estimates of these functionals $\hat{R}(f)$ and $\hat{R}(f'')$ are available, the plug-in estimator of the optimal bandwidth h is

$$\hat{h} = n^{-1/5} \left[\frac{C R_K \hat{R}(f)}{2\sigma_K^4 \hat{R}(f'')} \right]^{1/5}. \quad (10)$$

This suggests using (10) to derive a possible bandwidth selection method. However, we have not found this method to be practicable.

Remark 7. To help the reader to understand why this method cannot be applied easily note that already at this stage in order to estimate one unknown quantity (bandwidth h) we have to estimate two (functionals (8) and (9)).

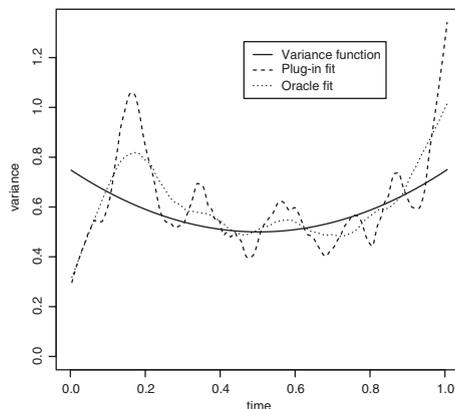


Fig. 1. Comparison of the plug-in fit and the oracle fit based on 500 data points.

As an example, we consider the variance function (2a) and the mean function $g(x) \equiv 0.5$ for $x \in [0, 1]$. Note that, since the mean function is constant, the variance function is the only source of bias in the mean squared error of the variance function estimator $\hat{f}(x)$. We use the bandwidth estimator (10) as an oracle. In order to estimate bandwidth (10), we need the estimators of the two quadratic functionals (8) and (9). The following is a brief description of the more extended presentation in Levine (2003); many ideas are parallel to those expounded in Efromovich (1999). Let us denote the j th element of the standard Fourier basis on $[0, 1]$ by $\phi_j(x)$; clearly, the j th Fourier coefficient of $f(x)$ $\alpha_j = \int_0^1 f(x)\phi_j(x) dx$ can be reasonably estimated by $\hat{\alpha}_j = n^{-1} \sum_{i=1}^n z_i^2 \phi_j(x_i)$. Based on Plancherel–Parseval inequality, we estimate (8) by

$$\hat{R}(f) = \sum_{j=0}^{J_1} \left(\hat{\alpha}_j^2 - \frac{\hat{C}_j(f)}{n} \right)$$

while (9) can be estimated by

$$\hat{R}(f'') = 2\pi^4 \sum_{j=0}^{J_2} j^4 \left(\hat{\alpha}_j^2 - \frac{\hat{C}_j(f)}{n} \right).$$

In the above, J_1 and J_2 are the cutoffs that depend on the smoothness of the function $f(x)$ and $\hat{C}_j(f)$ is some preliminary (pilot) estimate of the functional $C_j(f) = 2 \int_0^1 f^2(x)\phi_j^2(x) dx$. It is also easy to find that the first cutoff should be $J_1 = O(n)$, while the second has to be $J_2 = O(n^{-1/9})$ to ensure that the estimators $\hat{R}(f)$ and $\hat{R}(f'')$ are consistent.

As before, we use the 100 Monte-Carlo simulations and the sample size is 500 (Fig. 1). We only consider one sample size since the performance of this estimator is very poor and does not improve significantly with the increased sample size. Note that even the choice of the oracle bandwidth gives a very poor fit of the variance function. Such a bad performance is related to the fact that the plug-in bandwidth relies on the first-order asymptotics only.

In particular, the asymptotic expansion of the bias only takes into account the first two derivatives of the variance function $f(x)$ while the variance expansion relies only on the value of the function itself. Since the function (2a) has only two nonzero derivatives, the bias expansion is exact; however, the variance is underestimated and the selected bandwidth is too large, resulting in large bias. The same features are even more pronounced in the oracle fit, due to the fact that there is an additional variance resulting from estimating functionals (8) and (9).

It seems that estimating bandwidth by a plug-in method is a more complicated problem than the original problem of variance estimation itself. That makes the alternative approach that we consider in the next section all the more attractive.

4. Crossvalidation approach

The basic idea of any crossvalidation-type method is to estimate the true MISE. By definition, MISE depends on the unknown function $f(x)$; thus, we want to have a data-driven estimator that mimics its behavior.

In order to make it easier to follow we will use the notation $\hat{f}_h(x)$ to stress the fact that the variance function estimate depends on the bandwidth h . Remember that the integrated squared error is a global measure of risk defined as

$$MISE = \int E\left(\hat{f}_h(x) - f(x)\right)^2 dx. \tag{11}$$

We will also introduce its discrete counterpart that we are going to call discrete mean squared error (DMSE)

$$DMSE = n^{-1} \sum_{i=1}^n E\left(\hat{f}_h(x_i) - f(x_i)\right)^2. \tag{12}$$

Both (11) and (12) depend on the unknown variance function $f(x)$ and so we need to estimate them.

One possible way of estimating (11) can be briefly described as follows. Partition the data $\{x_i, y_i\}$ at random into K approximately equal and disjoint subsets. Each of these subsets consists of about k_j pairs where $\sum k_j = n$. Let the $\{\tilde{x}_i^j, \tilde{y}_i^j\}$, $i = 1, \dots, k_j$ denote the pairs in the j th subset with the values of \tilde{x}_i^j arranged in ascending order; that is, $\tilde{x}_i^j \leq \tilde{x}_{i+1}^j$, $i = 1, \dots, k_j - 1$. Similarly, let $\{x_i^j, y_i^j\}$, $i = 1, \dots, n - k_j$ denote the pairs in the complement of the j th subset, again with the x_i^j arranged in ascending order. Using the terminology that originates in machine learning we call the set $\{x_i^j\}$ the training dataset and the set $\{\tilde{x}_i^j\}$ a validation dataset. Now, let $\Delta_i^j, \tilde{\Delta}_i^j$ denote the pseudoresiduals formed from $\{x_i^j, y_i^j\}$ and $\{\tilde{x}_i^j, \tilde{y}_i^j\}$, respectively. Then, let \hat{f}_h^j denote the estimator derived from the j th subset $\{x_i^j, y_i^j\}$ when using bandwidth h and squared pseudoresiduals $(\Delta_i^j)^2$. Unless stated

otherwise, the superscript j will be omitted for the values of the argument x to simplify the notation.

We estimate the MISE as follows. First, define

$$CV^j(h) = \sum_{l=1}^{k_j} \left(\left(\tilde{A}_l^j \right)^2 - \hat{f}_h^j(\tilde{x}_l) \right)^2. \quad (13)$$

Then, the crossvalidation criterion is

$$CV(h) = \frac{1}{n} \sum_{j=1}^K CV^j(h). \quad (14)$$

As a final step of the algorithm, we choose as optimal bandwidth

$$h_{CV} = \underset{h \in H}{\operatorname{argmin}} CV(h), \quad (15)$$

where $H \in [0, 1]$ is the finite grid that we use for simulation purposes.

The reasons we chose a version of K -fold crossvalidation can be described as follows. In general, the classical “leave-one-out” crossvalidation ($K = n$) is not popular as it is very computationally expensive. As a practical measure, the K -fold crossvalidation provides a useful alternative to the hold-out estimate of the generalization error (the empirical error on the test set) without reducing the cardinality of the training data set. This intuition seems to have some basis in reality; for example, Blum et al. (1999) established that for any nontrivial learning algorithm the K -fold estimate of generalization error is strictly more accurate than a single hold-out estimate on $1/K$ of the original data set for any $2 < K < n$. For some earlier work on the same subject, see e.g. Anthony and Holden (1998). However, all of the above handle only uncorrelated data. The question as to what happens to crossvalidation when the data is correlated has been only partly investigated before. Chu and Marron found (see Chu and Marron, 1991) that, in the case of “leave-one-out” crossvalidation and positively correlated data, the crossvalidation will produce very small bandwidths resulting in undersmoothed estimates; on the other hand, if the observations are negatively correlated, then crossvalidation will produce very large bandwidths that result in oversmoothed estimates.

Our method is essentially a K -fold crossvalidation for the correlated data. As far as we know, the properties of K -fold crossvalidation in this case have not been investigated so far. Thus, our method may have finite sample properties that are somewhat different from the standard K -fold crossvalidation. It was established in Levine (2003) that the estimator class defined by (6) has the variance that decays rapidly as r rises and, therefore, for many potential applications only small values of r , such as $r = 2$ or 3 , are needed. We conjecture that for small values of r (which implies a relatively low-depth correlation between

pseudoresiduals z_i^2) the performance of K -fold crossvalidation will be similar to that for independent data.

The following heuristics should help to understand why the algorithm described by (13)–(14) works. Note that the CV criterion as defined in (14) gives

$$\begin{aligned} & \frac{1}{n} \sum_{j=1}^K \sum_{l=1}^{k_j} \left[\left(\tilde{\Delta}_l^j \right)^2 - \hat{f}_h^j(\tilde{x}_l) \right]^2 \\ &= \frac{1}{n} \sum_{j=1}^K \sum_{l=1}^{k_j} \left[\left(\tilde{\Delta}_l^j \right)^2 - f(\tilde{x}_l) \right]^2 \\ & \quad + \frac{1}{n} \sum_{j=1}^K \sum_{l=1}^{k_j} \left[f(\tilde{x}_l) - \hat{f}_h^j(\tilde{x}_l) \right]^2 \\ & \quad + \frac{2}{n} \sum_{j=1}^K \sum_{l=1}^{k_j} \left[\left(\tilde{\Delta}_l^j \right)^2 - f(\tilde{x}_l) \right] \left[f(\tilde{x}_l) - \hat{f}_h^j(\tilde{x}_l) \right]. \end{aligned} \tag{16}$$

Looking at the terms one by one, we note that the first term in (16) does not depend on the bandwidth h . More precisely, it is easy to check that the first term on the right-hand side of (16) behaves asymptotically as

$$2 \int f^2(u) du + O(n^{-2}) \tag{17}$$

when $n \rightarrow \infty$. Thus, the first term is approximately a constant for large n .

To check this, note that since $\left(\tilde{\Delta}_l^j \right)^2$ can be viewed as an estimator of the variance function $f(x)$ at the point \tilde{x}_l the expected value of the first term in (16)

$$E \left[\left(\tilde{\Delta}_l^j \right)^2 - \hat{f}_h^j(\tilde{x}_l) \right]^2 \tag{18}$$

is the mean squared error of $\tilde{\Delta}_l^j$ as an estimator of $\hat{f}_h^j(\tilde{x}_l)$. Let us approximate $\left[\left(\tilde{\Delta}_l^j \right)^2 - f(\tilde{x}_l) \right]^2$ by its expectation

$$\left[\left(\tilde{\Delta}_l^j \right)^2 - f(\tilde{x}_l) \right]^2 \approx Bias^2 \left(\tilde{\Delta}_l^j \right)^2 + Var \left(\tilde{\Delta}_l^j \right)^2. \tag{19}$$

The notation $\bar{y}_i \doteq y_i - g(x_i)$ (where i is the generic index) will be used for centered observations. Next, using the first-order Taylor formulas for both $f(x)$ and $g(x)$ we

find

$$\begin{aligned}
 E\left(\tilde{\Delta}_l^j\right)^2 &= E\left[\sum_{j=0}^{r-1} d_j g\left(\tilde{x}_{j+l}\right) + \sum_{j=0}^{r-1} d_j \bar{y}_{j+l}\right]^2 \\
 &= \left[\sum_{j=0}^{r-1} d_j\left(g\left(\tilde{x}_l\right) + g'\left(\tilde{x}_l + \delta_1\right) \frac{j}{n}\right)\right]^2 \\
 &\quad + f\left(\tilde{x}_l\right) + n^{-1} \sum_{j=0}^{r-1} j d_j^2 f'\left(\tilde{x}_l + \delta_2\right) \\
 &\leq f\left(\tilde{x}_l\right) + \frac{C_1}{n^2}\left(\sum_{j=1}^{r-1} j d_j^2\right)^2 + n^{-1} C_2 \sum_{j=1}^{r-1} j d_j^2,
 \end{aligned} \tag{20}$$

where $0 < \delta_i < 1$, $i = 1, 2$, $C_1 > 0$ and $C_2 > 0$ are constants that depends on the choice of $g(x)$ and $f(x)$ only. Thus, the squared bias of $\left(\tilde{\Delta}_l^j\right)^2$ is of the order $O\left(n^{-2}\right)$ as $n \rightarrow \infty$.

Next, note that $\left(\tilde{\Delta}_l^j\right)^2$ is a scaled $\chi_1^2(0)$ where the scaling factor is equal to $f\left(\tilde{x}_l\right)$ plus the terms of the order $o(1)$. As a result, it is easy to check that the variance of the squared pseudoresidual $\left(\tilde{\Delta}_l^j\right)^2$ is

$$\text{Var}\left(\tilde{\Delta}_l^j\right)^2 = 2f^2\left(\tilde{x}_l\right) + O\left(n^{-2}\right). \tag{21}$$

Eq. (21) and the asymptotic behavior of the bias of $\left(\tilde{\Delta}_l^j\right)^2$ mean that (17) is true.

The second term in (16) is an approximation for the average integrated mean squared error (MISE) of the estimator $\hat{f}_h(x)$. If we can argue that the expected value of the third term goes to zero at the rate faster than the second, the minimizer of the left-hand side in (16) (the crossvalidation criterion as we defined it) can be expected to mimic the behavior of the minimizer of the mean squared error of $\hat{f}_h(x)$.

To analyze the third term, note that the estimator $\hat{f}_h^j(x)$ is the same as the generic variance estimator $\hat{f}_h(x)$ except that it is defined using not the entire data set of size n but only $n - k_j$ points $\left\{x_i^j\right\}$, $i = 1, \dots, n - k_j$.

Thus, at any point \tilde{x}_l from the validation data set we have

$$\hat{f}_h^j\left(\tilde{x}_l\right) = \frac{\sum_{m=1}^{n-k_j}\left(\Delta_m^j\right)^2 K\left(\frac{\tilde{x}_l - x_m}{h}\right)}{\sum_{m=1}^{n-k_j} K\left(\frac{\tilde{x}_l - x_m}{h}\right)}. \tag{22}$$

Using the notation first introduced in Härdle (1990), we denote

$$W(\tilde{x}_l - x_m) = \frac{K\left(\frac{\tilde{x}_l - x_m}{h}\right)}{\sum_{m=1}^{n-k_j} K\left(\frac{\tilde{x}_l - x_m}{h}\right)} \tag{23}$$

thus enabling us to write estimator (22) in the form

$$\hat{f}_h^j(\tilde{x}_l) = \sum_{m=1}^{n-k_j} \left(\Delta_m^j\right)^2 W(\tilde{x}_l - x_m). \tag{24}$$

Remember that the pseudoresiduals Δ_m^j form an r -dependent sequence; this means that the number of those Δ_m^j that are correlated with $\tilde{\Delta}_l^j$ is $O(r)$. Let us represent (24) as

$$\hat{f}_h^j(\tilde{x}_l) = \hat{f}_{h,r}^j(\tilde{x}_l) + \hat{f}_{h,-r}^j(\tilde{x}_l), \tag{25}$$

where

$$\hat{f}_{h,r}^j(\tilde{x}_l) = \sum_{m \in J_1} \left(\Delta_m^j\right)^2 W(\tilde{x}_l - x_m) \tag{26}$$

and

$$\hat{f}_{h,-r}^j(\tilde{x}_l) = \sum_{m \in J_2} \left(\Delta_m^j\right)^2 W(\tilde{x}_l - x_m). \tag{27}$$

Here the index set J_1 includes those pseudoresiduals Δ_m^j that are correlated with $\tilde{\Delta}_l^j$. Clearly, the cardinality of this set is at most $O(r)$. The index set J_2 is its complement. Of course, $|J_1| + |J_2| = n - k_j$. The crossproduct term

$$C_{2n} = \frac{2}{n} \sum_{j=1}^K \sum_{l=1}^{k_j} \left[\left(\tilde{\Delta}_l^j\right)^2 - f(\tilde{x}_l) \right] \left[f(\tilde{x}_l) - \hat{f}_h^j(\tilde{x}_l) \right] \tag{28}$$

can be now split into two terms:

$$\begin{aligned} C_{2n} &= \frac{2}{n} \sum_{j=1}^K \sum_{l=1}^{k_j} \left[\left(\left(\tilde{\Delta}_l^j\right)^2 - f(\tilde{x}_l) \right) \left(f(\tilde{x}_l) - \hat{f}_{h,-r}^j(\tilde{x}_l) \right) \right] \\ &\quad - \frac{2}{n} \sum_{j=1}^K \sum_{l=1}^{k_j} \left[\left(\left(\tilde{\Delta}_l^j\right)^2 - f(\tilde{x}_l) \right) \hat{f}_{h,r}^j(\tilde{x}_l) \right]. \end{aligned} \tag{29}$$

It is easy to analyze the first term. The factors that comprise it are independent; also, as was pointed out before in (20), the bias of $\left(\tilde{\Delta}_l^j\right)^2$ as an estimator of $f(\tilde{x}_l)$ is of the order $O(n^{-1})$. At the same time, note that the second factor can be viewed as the bias of $\hat{f}_{h,-r}^j$

used as an estimator of $f(\tilde{x}_l)$. Since the cardinality of the set J_1 is $O(r)$, we can easily conclude that $|J_2| = O(n)$. Hence, this bias is asymptotically the same as the order of the bias of the regular estimator $\hat{f}_h(\tilde{x}_l)$: $E[f(\tilde{x}_l) - \hat{f}_{h,-r}^j(\tilde{x}_l)] = O(h^2)$. Consequently, choosing the usual optimal bandwidth $h = O(n^{-1/5})$ we find that

$$\begin{aligned} & E \left[\left(\tilde{\Delta}_l^j \right)^2 - f(\tilde{x}_l) \right] \left[f(\tilde{x}_l) - \hat{f}_{h,-r}(\tilde{x}_l) \right] \\ &= E \left[\left(\tilde{\Delta}_l^j \right)^2 - f(\tilde{x}_l) \right] E \left[f(\tilde{x}_l) - \hat{f}_{h,-r}(\tilde{x}_l) \right] \\ &= O\left(\frac{1}{n}\right) O(h^2) = O\left(\frac{h^2}{n}\right) = O\left(n^{-7/5}\right). \end{aligned}$$

Thus, the first term in (16) goes to zero much faster than $O(n^{-4/5})$.

Analyzing the second term in (29), it is important to remember that $\hat{f}_{h,r}^j(\tilde{x}_l)$ is based on a finite number of terms only. Without loss of generality, it is possible to assume that the mean $g(x) \equiv 0$ for the purpose of this analysis. Then, using representation (26) for $\hat{f}_{h,r}(\tilde{x}_l)$, we find that

$$\begin{aligned} & \sum_{l=1}^{k_j} \left(\tilde{\Delta}_l^j \right)^2 \hat{f}_{h,r}^j(\tilde{x}_l) \\ &= \sum_{l=1}^{k_j} \left(\tilde{\Delta}_l^j \right)^2 \sum_{k \in J_1} \left(\Delta_k^j \right)^2 W(\tilde{x}_l - x_k) \\ &= \sum_{l=1}^{k_j} \sum_{k \in J_1} \left(\sum_{p,q=0}^{r-1} d_p d_q \tilde{y}_{p+l} y_{q+k} \right)^2 W(\tilde{x}_l - x_k). \end{aligned}$$

In order to estimate the expected value of the above, we will use the upper bound for the absolute s th moment of quadratic forms in independent random variables due to Whittle (1960). It allows us to conclude that

$$\begin{aligned} & E \left(\sum_{p,q=0}^{r-1} d_p d_q \tilde{y}_{p+l} y_{q+k} \right)^2 \\ &\leq C \sum_{p,q=0}^{r-1} d_p^2 d_q^2 \left(E \tilde{y}_{p+l}^4 \right)^{1/4} \left(E y_{q+k}^4 \right)^{1/4} \\ &= C \sum_{p,q=0}^{r-1} d_p^2 d_q^2 \sqrt{f(x_{p+l}) f(x_{q+k})}, \end{aligned} \tag{30}$$

where the constant C depends on r and the choice of the difference sequence $\{d_i\}$. As the function $f(x)$ is bounded expression (30) is bounded as well. This allows us to conclude that

$$\sum_{l=1}^{k_j} \left(\tilde{\Delta}_l^j\right)^2 \hat{f}_{h,r}^j(\tilde{x}_l) \leq C \sum_{l=1}^{k_j} \sum_{k \in J_l} W(\tilde{x}_l - x_k) \tag{31}$$

for some constant C . By definition, the coefficients $W(\tilde{x}_l - x_k)$ are normalized and add up to one:

$$\sum_{k=1}^{n-k_j} W(\tilde{x}_l - x_k) = 1.$$

Since the kernel function $K(\cdot)$ is assumed bounded, it is easy to see that asymptotically, as $n \rightarrow \infty$, each coefficient $W(\tilde{x}_l - x_k)$ behaves as $O\left(\frac{1}{n}\right)$. The number of observations in J_l is finite (at most $O(r)$), and thus

$$\frac{1}{n} E \sum_{j=1}^K \sum_{l=1}^{k_j} \left(\tilde{\Delta}_l^j\right)^2 \hat{f}_{h,r}^j(\tilde{x}_l) = O(1). \tag{32}$$

It was shown in Levine (2003) that the estimator $\hat{f}_{h,r}^j(x_l)$ is asymptotically consistent and, in particular,

$$\hat{f}_{h,r}^j(x) = f(x) + \frac{1}{\sqrt{n}} O_p(1). \tag{33}$$

As the variance function $f(x)$ is bounded, (33) means that the expected value of the product $f(x_l) \hat{f}_{h,r}^j(x_l)$ is bounded as well. This fact and (32) mean that the expectation of the second term in (29) is of the order $O\left(\frac{1}{n}\right)$. As the first term goes to zero much faster, the expectation of the entire crossproduct (16) is also of the order $O\left(\frac{1}{n}\right)$.

We know that the first term in (16) behaves as a constant asymptotically, the second term imitates MISE of the variance estimator $\hat{f}_h(x)$ and the third one tends to zero at the rate of $O\left(\frac{1}{n}\right)$. Thus, minimizing the crossvalidation criterion $CV(h)$ mimics equivalent to minimizing the MISE of the variance estimator $\hat{f}_h(x)$ and the minimizer of the crossvalidation criterion (14) mimics the behavior of the minimizer of the MISE of the variance estimator $\hat{f}_h(x)$.

5. The finite sample performance of the method and the choice of K : the first case study

It is known from Levine (2003) (and earlier literature cited there) that the shape of the mean function is not important asymptotically as long as the minimal smoothness condition

(the existence and continuity of the first derivative $g'(x)$) is satisfied. However, the mean function can become rather important in finite samples. The mean-function related bias term can be easily estimated for any sample size n . It was shown in Levine (2003) that asymptotically this term becomes approximately equal to

$$\frac{C(\{d_i\}, r) [g'(x)]^2}{n^2}, \quad (34)$$

where $C(\{d_i\}, r)$ is the constant that depends on the choice of the difference sequence $\{d_i\}$ and its order r . It is possible to show that, given the order r , the difference sequence can be uniquely selected in a way that minimizes the variance of estimator (6) under constraints (2) and (3). Such a sequence cannot be written down in the closed form but it is possible to calculate its elements for any order r ; for details, see Levine, 2003 and Hall et al., 1990. If this optimal difference sequence is chosen, the constant $C(\{d_i\}, r)$ becomes $\frac{(2r+1)(r+1)}{12}$ (see, for example Dette et al., 1998).

Example. Suppose our variance function is constant: $f(x) \equiv 1$ while the mean function is such that it has large (in absolute value) first derivative; for example, let us define

$$g(x) = \begin{cases} 100x & \text{if } 0 \leq x < 1, \\ 0 & \text{otherwise.} \end{cases} \quad (35)$$

Then, it is easy to check that for $n \approx 112$ term (34) is of the same magnitude as the variance function (parameter to be estimated) itself; its absolute value is approximately equal to 1. Clearly, in this example the bias-related mean term is anything but negligible.

Thus, it is important to investigate how our method will perform for finite samples. Another important issue is a choice of parameter K . The opinion prevailing in the literature (see, for example Hastie et al., 2001) is that $K = 5$ or 10 should be used; however, it is better to run a test to see what the experience may suggest in our case.

In order to do this, we run a fixed model with different choices of K . We use a simple choice of smooth quadratic variance function $f(x) = (x - 0.5)^2 + 0.5$ and a constant mean function $g(x) = \frac{3}{4}$. The design is equispaced and fixed; in other words, $x_i = \frac{i}{n}$, $i = 1, 2, \dots, n$. We use three different sample sizes: first, $n = 500$, then $n = 1000$ and $n = 2000$ data points on $[0, 1]$. For each of the choices of n we try three different choices of K : $K = 5, 10$ and 15 , respectively. Results are shown in Figs. 2–4. All of them employ the following convention.

On the X-axis are the values of the discrete oracle loss which is denoted

$$ODMSE = \min_{h \in H} n^{-1} \sum_{i=1}^n [\hat{f}_h(x_i) - f(x_i)]^2. \quad (36)$$

Here, ODMSE stands for Oracle Discrete Mean Squared Error. Of course, we can calculate ODMSE only because we work with the simulated data where $f(x)$ is known beforehand.

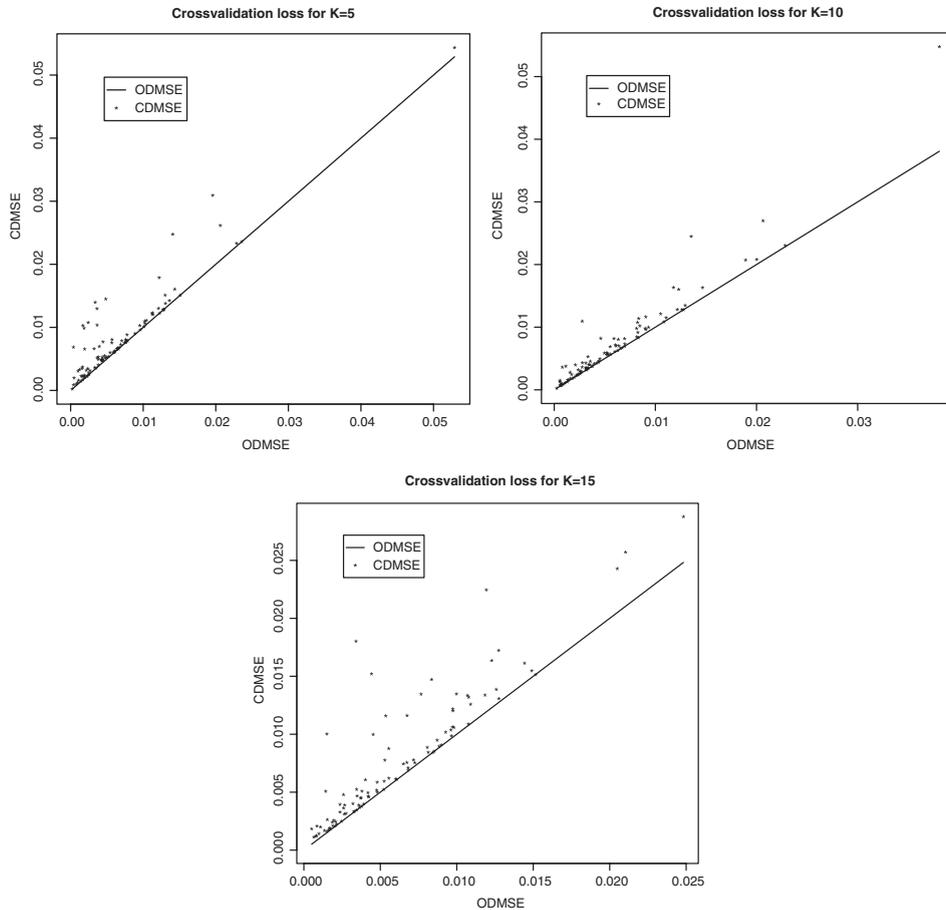


Fig. 2. Comparison of the crossvalidation risk and the oracle risk based on 500 data points for different choices of partition size K .

The Y-coordinate is a crossvalidation DMSE (CDMSE) that is defined as

$$CDMSE = n^{-1} \sum_{i=1}^n \left[\hat{f}_{h_{CV}}(x_i) - f(x_i) \right]^2 \tag{37}$$

with h_{CV} defined by (15). Clearly, $ODMSE \leq CDMSE$. The closer CDMSE is to the ODMSE, the better choice of the bandwidth has been made.

First, note that for the fairly low sample size of $n = 500$, increasing K seem to decrease slightly the variability of CDMSE values. At the same time, the method performance is also getting better (in case of $K = 15$, all values of CDMSE except for three outliers are less than 0.025 while for $K = 5$ the largest nonoutlier value is above 0.03). This is not true for larger sample sizes, though. When $n = 1000$, we have values of CDMSE that are highly

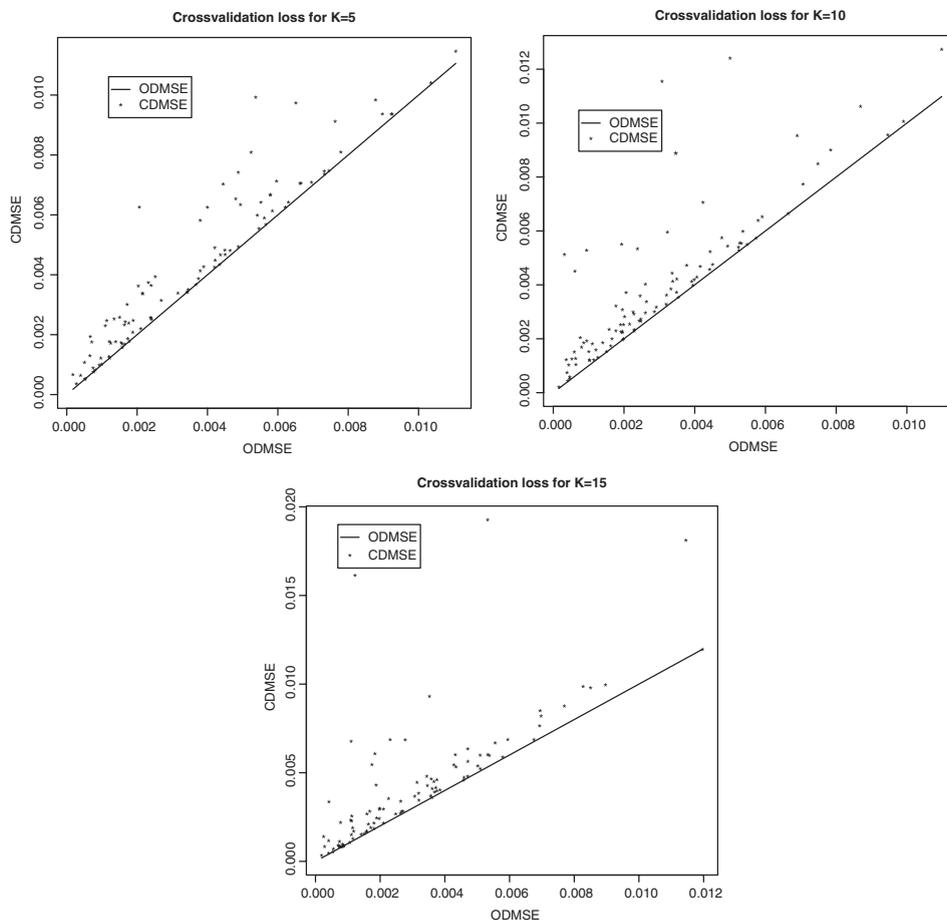


Fig. 3. Comparison of the crossvalidation risk and the oracle risk based on 500 data points for different choices of partition size K .

spread out. Increasing K does little to reduce the variability of CDMSE; only the choice of $K = 15$ seems to offer some improvement over $K = 5$, but the reduction is not great. At the same time, the method performance as measured by the difference between ODMSE and CDMSE gets worse as K increases; when $K = 15$, the largest values of CDMSE are over 0.015 while for $K = 5$ they are (except for a very few outliers) staying close to 0.010. The same trends hold when sample size is $n = 2000$. In other words, increasing K does very little, if anything, to decrease the variability while making the method performance worse.

One of the few literary sources on the subject, [Hastie et al. \(2001\)](#) suggests $K = 5$ or 10 as an optimal choice. The argument goes that the choice of K is, probably, the matter of bias-variance tradeoff for the K -fold crossvalidation criterion as an estimator of the global mean squared error. Our empirical analysis suggests that, unless the sample size is quite small, $K = 10$ is, probably, the better option.

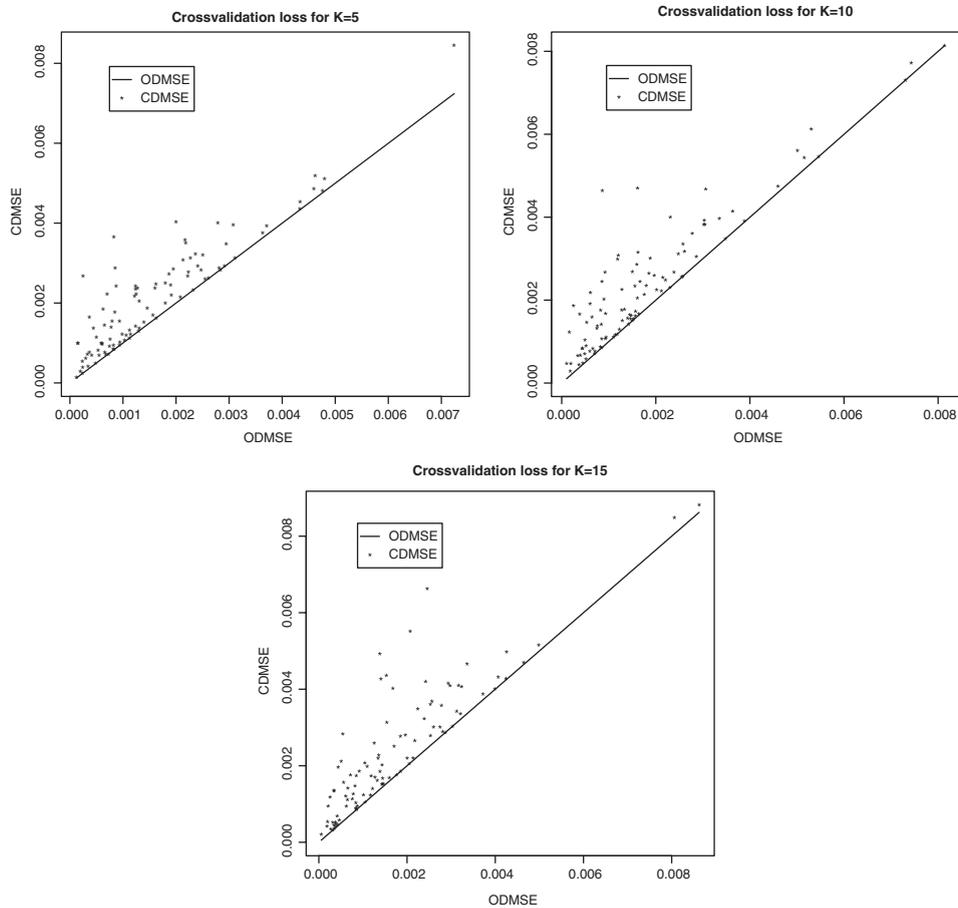


Fig. 4. Comparison of the crossvalidation risk and the oracle risk based on 2000 data points for different choices of partition size K .

6. Second case study

To test how well the method performs for different possible mean functions, the following study design was followed. We chose a smooth quadratic variance function

$$f(x) = \left(x - \frac{1}{2}\right)^2 + \frac{1}{2} \tag{38}$$

that remained fixed throughout this study. The mean function choices ranged from ones with the small mean-related bias term (34) to less smooth ones. Since this term, if measured

globally over $[0, 1]$, is asymptotically proportional to the

$$R(g') = \int [g'(x)]^2 dx \quad (39)$$

we will use the quadratic functional (39) as a measure of the mean function influence on the variance estimation. Clearly, the larger (39) is, the more distortion the mean function would be expected to bring to the process.

Next we define a sequence of mean functions we use for our simulation study. Let us define first the indicator function

$$I(A) = \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{otherwise} \end{cases} \quad (40)$$

for any set A . Then, the following choices are considered:

1. $g_1(x) = \frac{3}{4}$,
2. $g_2(x) = \left(x + \frac{\sqrt{3}}{2}\right)^2 * I(x < 0.5) + \left(\frac{2+\sqrt{3}}{2} - x\right)^2 * I(x > 0.5)$,
3. $g_3(x) = \frac{3}{4} * \sin(16\pi x) + \frac{3}{4}$.

Setup (1)–(3) is fairly logical. Function choices range from constant to the sinusoid (thus increasing the value of (39)). The mean function (1) is constant, thus all its derivatives and the functional (39) are equal to zero. Consequently, the mean-related bias term is also zero and the difference between the small sample and large sample performance of the method is, probably, not very large. This is clearly not the case for (2) and (3); it is easy to calculate the value of $R(g')$ for each of them and find out that (39) is greater for (3) than it is for (2). Consequently, the finite sample performance is likely to be the best for (1) and the worst for (3).

As before, we consider an equidistant design with $x_i = \frac{i}{n}$, $i = 1, \dots, n$ where the number of observations n is first assumed to be 500, then 1000 and, finally, 2000. We perform 100 simulations for each choice of n and use $K = 10$ (Figs. 5–7).

Definition 8. We define the oracle bandwidth as

$$h_o = \operatorname{argmin}_{h \in H} DMSE(h) \quad (41)$$

and the crossvalidation bandwidth as before in (15).

Remark 9. The ODMSE, previously defined in (36), can be also described as $R_o = DMSE(h_o)$ while CDMSE is $R_{CV} = DMSE(h_{CV})$. As before, $R_{CV} \geq R_o$. The smaller (and closer to the oracle loss, respectively) the crossvalidation loss is, the better the performance of our method.

In graphs (5)–(7), we want to compare crossvalidation risk to the oracle risk. To achieve the best visual effect, we plot the oracle risk as a function of itself (the diagonal straight line) while the values of crossvalidation risk are depicted as separate points above the oracle risk line. We show three graphs corresponding to the three possible mean function choices

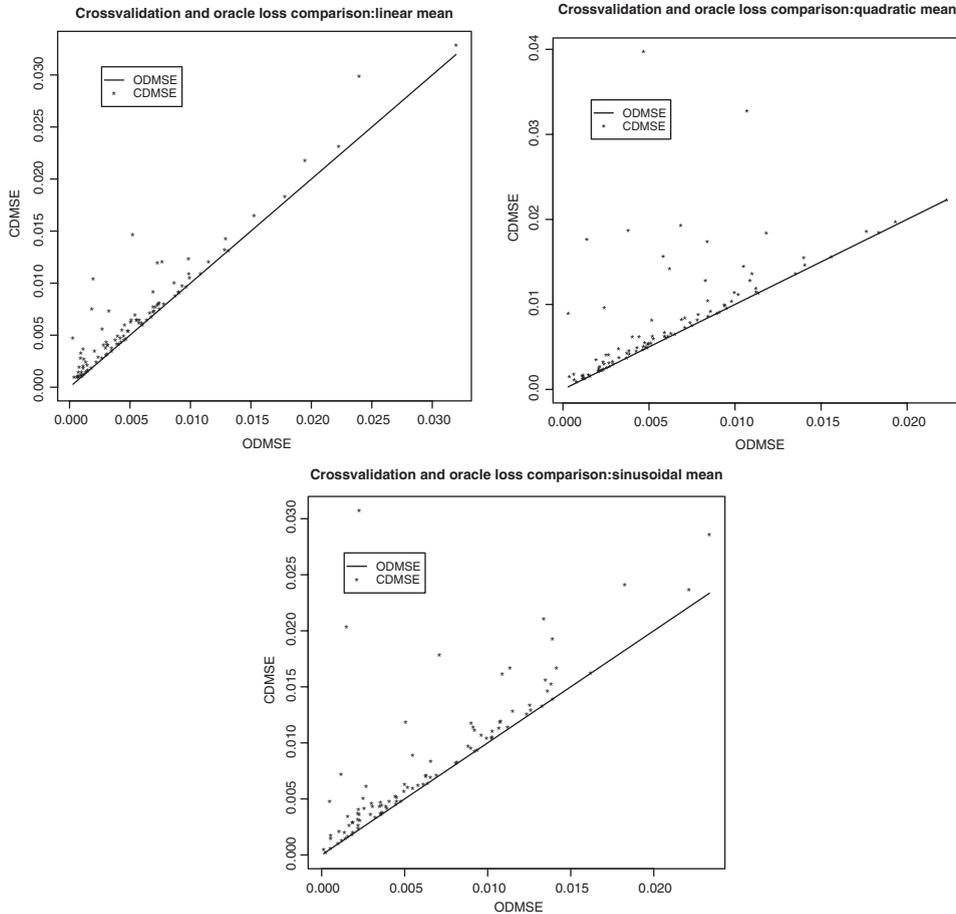


Fig. 5. Comparison of the crossvalidation risk and the oracle risk based on 500 data points for different choices of the mean function $g(x)$.

for each of the sample sizes $n = 500, 1000$ and 2000 . In each panel, we move clockwise from (1)–(3).

Tables 1 and 2 summarize values of $R(g')$ for each of the choices (1)–(3). It also gives the respective median oracle loss and median crossvalidation loss for every combination of the mean function (1)–(3) and the sample size. We report both ODMSE and CDMSE in the form of approximate 95% binomial confidence intervals. Their boundaries are m_1 st and m_2 d (ordered) values of ODMSE (CDMSE) where $m_i = Np_i$ and the percentiles $p_i = 0.5 \pm \frac{1.96}{2\sqrt{N}}$, $i = 1, 2$, N being the number of simulations performed (which in our case is equal to 100).

Note that for different choices of the mean function the differences in ODMSE do not look statistically significant as the confidence intervals overlap strongly. This seems to be

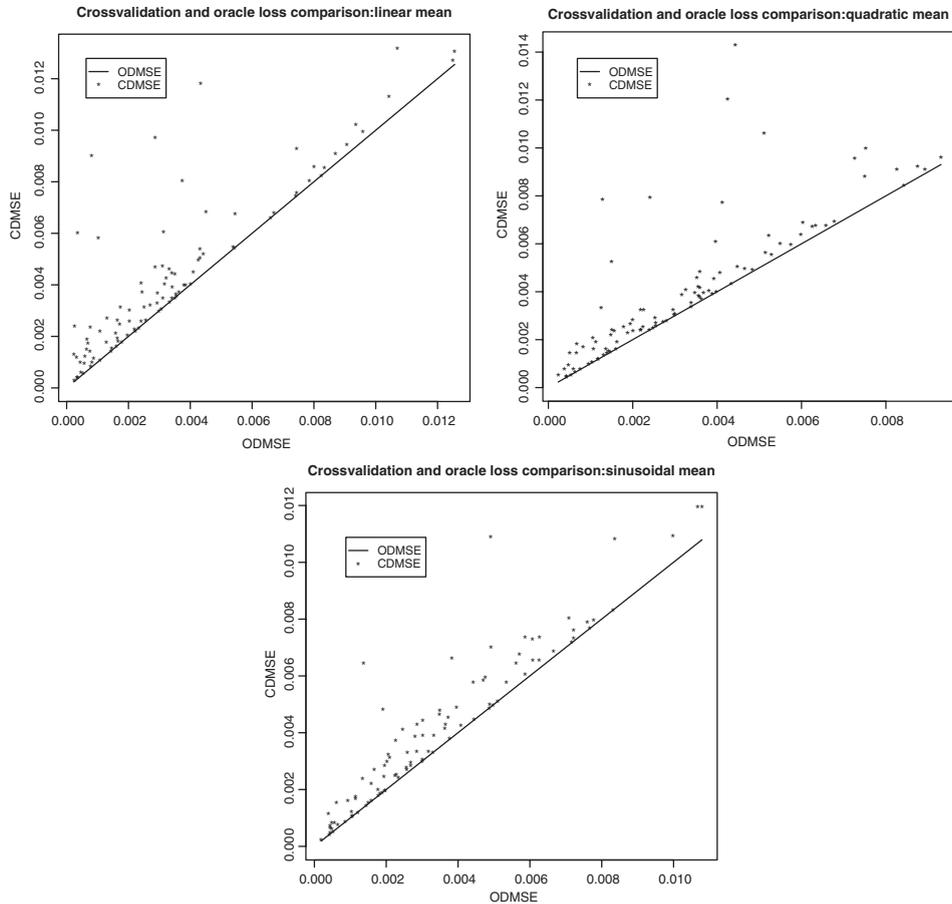


Fig. 6. Comparison of the crossvalidation risk and the oracle risk based on 1000 data points and different choices of the mean function $g(x)$.

true for each of the three sample sizes considered. In other words, the oracle loss changes very little, if at all, with the introduction of a more complicated mean function.

It is not entirely clear at this stage what is the influence of the particular choice of the mean on the performance of the bandwidth selection method. There seems to be a tendency towards increased crossvalidation loss as the mean function curvature increases; however, the effect is not very pronounced. Tables 1 and 2 show that overlap between the confidence intervals for CDMSE and ODMSE is still present for all choices of the mean function at any given sample size, although it becomes very small for quadratic and sinusoid mean functions for large $n = 1000$ and 2000 . More extensive simulations with wider choice of the mean function are necessary to clarify this issue better.

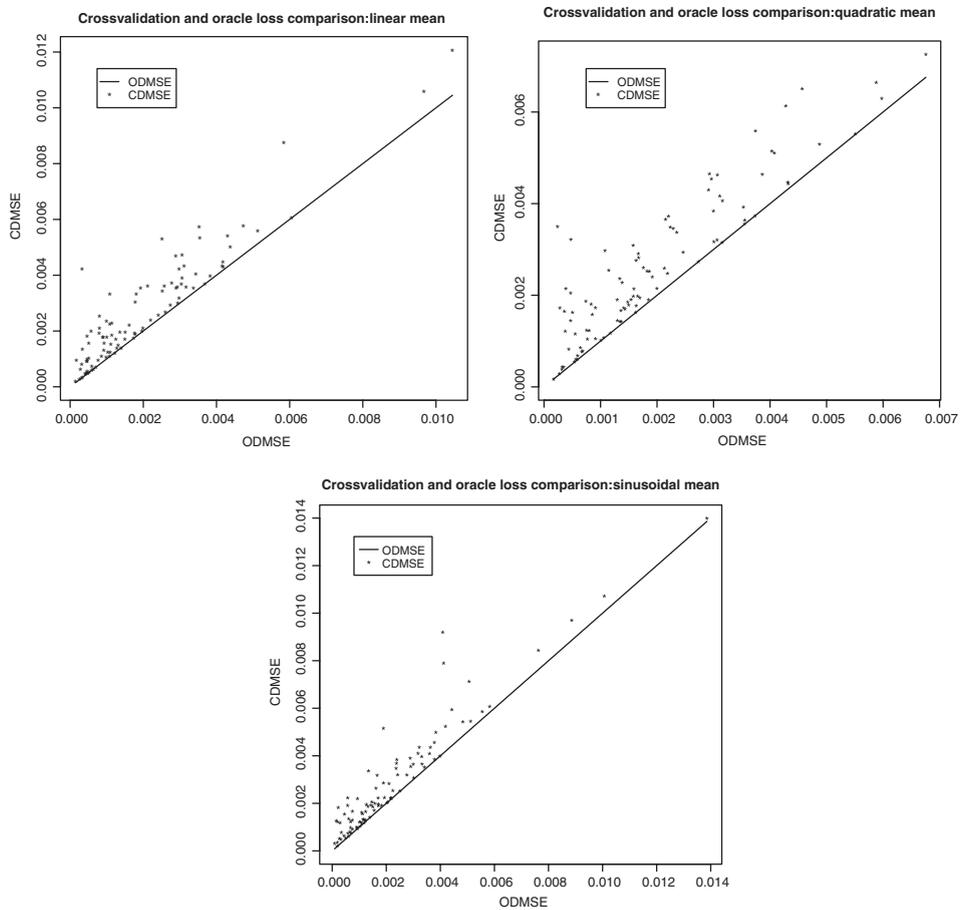


Fig. 7. Comparison of the crossvalidation risk and the oracle risk based on 2000 data points and different choices of the mean function $g(x)$.

Table 1
Performance under the changing curvature of the mean function

Mean function	$R(g')$	95% Confidence Intervals for Median ODMSE		
		$n = 500$	$n = 1000$	$n = 2000$
Constant	0	0.00390 0.00609	0.00215 0.00375	0.00124 0.00206
Quadratic	5.07	0.00395 0.00663	0.00226 0.00329	0.00124 0.00216
Sinusoid	$72\pi^2 \approx 710.61$	0.00357 0.00503	0.00204 0.00314	0.00107 0.00177

Table 2
Performance under the changing curvature of the mean function

Mean function	$R(g')$	95% Confidence Intervals for Median CDMSE		
		$n = 500$	$n = 1000$	$n = 2000$
Constant	0	0.00512 0.00780	0.00292 0.00464	0.00168 0.00284
Quadratic	5.07	0.00482 0.00814	0.00318 0.00442	0.00203 0.00263
Sinusoid	$72\pi^2 \approx 710.61$	0.00405 0.00568	0.00299 0.00441	0.00162 0.00235

Table 3
Performance under the changing curvature of the variance function

Variance function	$R(f'')$	95% Confidence Intervals for Median ODMSE		
		$n = 500$	$n = 1000$	$n = 2000$
(2a)	4	0.00423 0.00695	0.00223 0.00342	0.00113 0.00175
(2b)	16	0.00579 0.01073	0.00337 0.00496	0.00209 0.00321
(2c)	64	0.01330 0.02189	0.00765 0.01117	0.00438 0.00645

7. Third case study

Of course, the most important issue is whether the method performs well for variance functions of different smoothness. As a smoothness measure of the variance function the following quadratic functional is used:

$$R(f'') = \int [f''(x)]^2 dx. \quad (42)$$

The reason for this choice is that the asymptotically optimal bandwidth for estimator (6) is inversely proportional to (42). The larger (42) is, the smaller the optimal bandwidth has to be chosen to pick up ever smaller details of the original function $f(x)$.

We use the following study design:

1. We assume that the mean function is constant; for example, $g(x) = \frac{1}{2}$ for $x \in [0, 1]$. Our main purpose here is to keep the asymptotic mean-related bias term equal to zero so that only the variance function $f(x)$ contributes to the mean squared error asymptotic expansion of the estimator (6).
2. We will consider several choices of the variance function $f(x)$. They are arranged in a “ladder” of possible candidates from the ones having small (42) to those with a larger value of this functional. The choices are:
 - (a) $f(x) = (x - 0.5)^2 + 0.5$,
 - (b) $f(x) = 2(x - 0.5)^2 + 0.5$,
 - (c) $f(x) = 4(x - 0.5)^2 + 0.5$.

Table 4
Performance under the changing curvature of the variance function

Variance function	$R(f'')$	95% Confidence Intervals for Median CDMSE		
		$n = 500$	$n = 1000$	$n = 2000$
(2a)	4	0.00532 0.00818	0.00315 0.00430	0.00165 0.00235
(2b)	16	0.00775 0.01396	0.00505 0.00676	0.00344 0.00516
(2c)	64	0.02220 0.02967	0.01308 0.01750	0.00775 0.01014

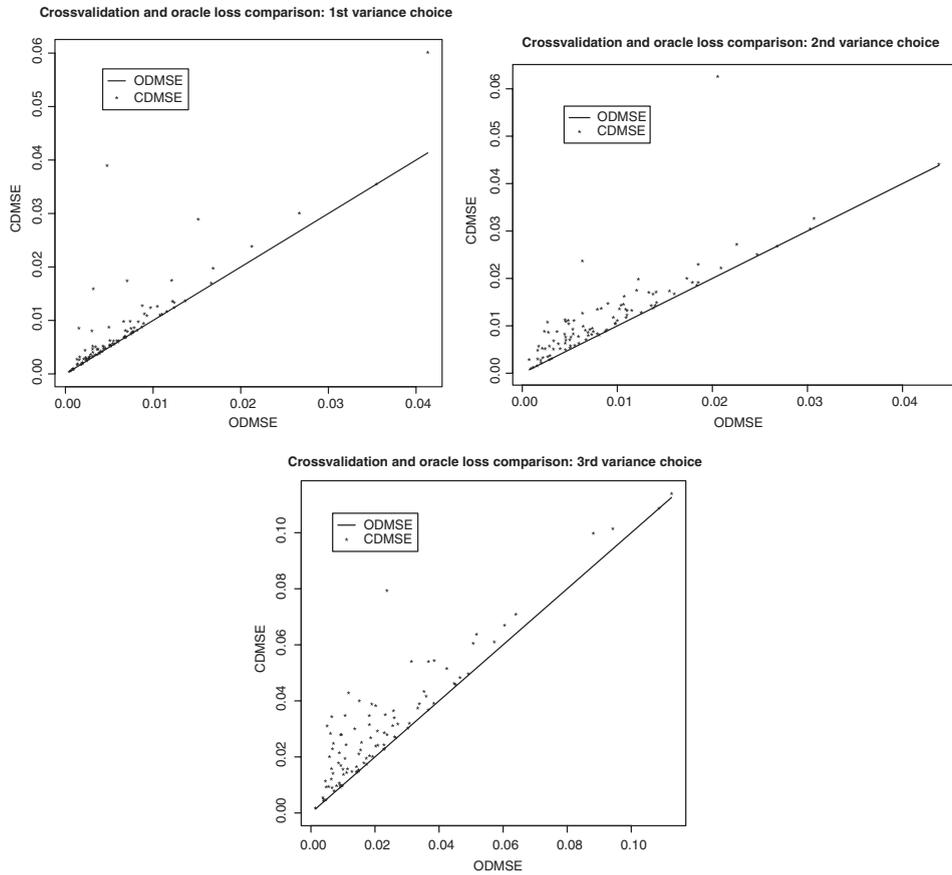


Fig. 8. Comparison of the crossvalidation risk and the oracle risk based on 500 data points and different choices of the variance function $f(x)$.

As we move from (2a) to (2c), the value of (42) increases; as a consequence, we expect the bandwidth selection to become increasingly more difficult.

As in the second study, we specify the crossvalidation parameter as $K = 10$. The sample size is first $n = 500$, then $n = 1000$ and eventually $n = 2000$. The number of Monte-

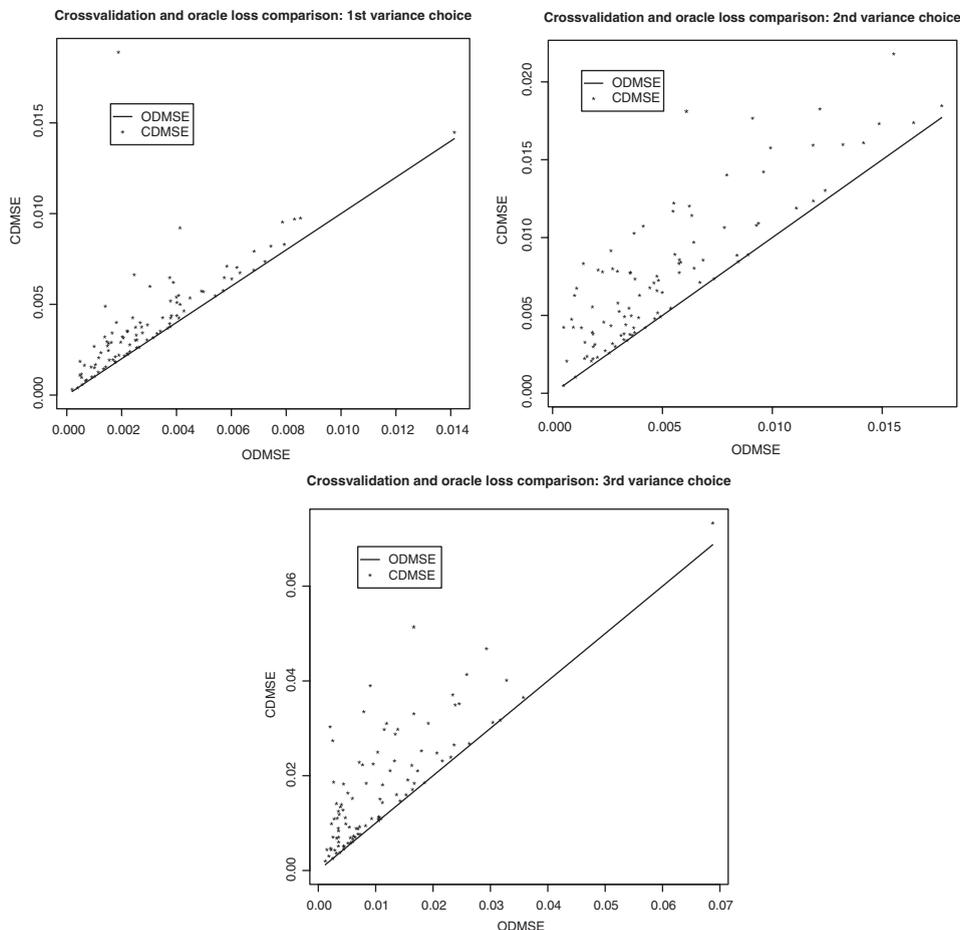


Fig. 9. Comparison of the crossvalidation risk and the oracle risk based on 1000 data points and different choices of the variance function $f(x)$.

Carlo simulations for each possible combination of the sample size and variance function is 100. Tables 3 and 4 give some indication of the performance of the method, together with Figs. 8–10.

The first conclusion we can make is that, indeed, the median oracle loss (ODMSE) seems to grow as the variance function becomes more difficult to handle. For each of the three sample sizes, ODMSE grows as we move from (2a) to the (2c) options for the variance functions. Note also that the median crossvalidation loss (CDMSE) exhibits the same tendency, in effect mimicking the oracle loss.

It is fairly clear just from the visual inspection of graphs (8)–(10) that the distance between ODMSE and CDMSE is larger for more complicated choices of the variance function. For example, for $n = 1000$ and the simplest choice of the variance function (2a) the ODMSE and CDMSE confidence intervals still have a small overlap of 0.00027. However, for (2b)

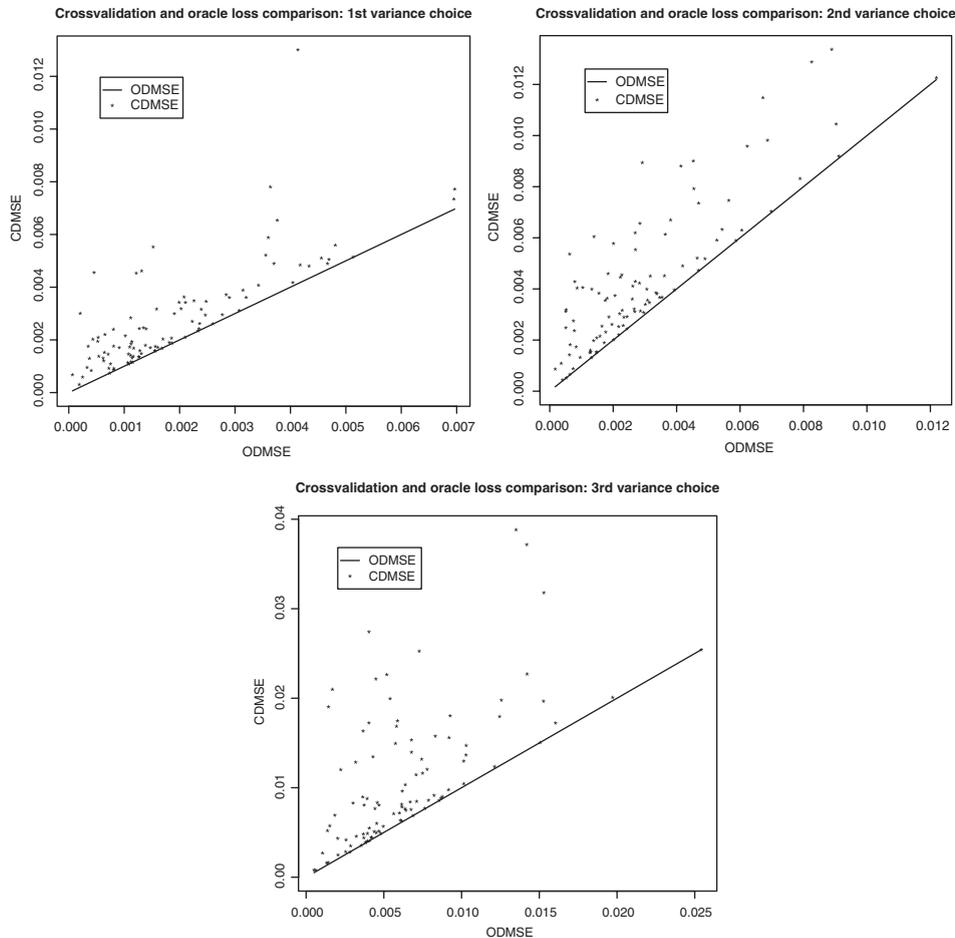


Fig. 10. Comparison of the crossvalidation risk and the oracle risk based on 2000 data points and different choices of the variance function $f(x)$.

they do not overlap at all and the same is true for (2c). The situation is exactly the same for $n = 2000$. These results suggest that CDMSE becomes an increasingly poor estimator of ODMSE as the degree of the variance function curvature increases.

Visually, the performance of the method seems to improve for larger sample sizes. This improvement looks to be rather slow for the fixed choice of the variance function $f(x)$ which seems to indicate similarity between our method and the classical “leave-one-out” crossvalidation. It is known that the relative rate of convergence of the crossvalidation-selected bandwidth estimator to the oracle bandwidth is $O(n^{-1/10})$ (for details, see Härdle et al., 1988). We may conjecture that the similar phenomenon is at work here as well. More extensive simulations with the wider choice of possible variance functions $f(x)$ are probably needed here.

8. Conclusions

At this point, we have a workable method that allows us to select the bandwidth for the general variance function estimator (6). The method can be easily implemented using any statistical software and is intuitively appealing. The method is reasonably insensitive to fluctuations in the mean function $g(x)$; however, it does tend to break down for highly variable variance functions, as shown in Figs. 8–10.

There are a lot of questions that remain unanswered for now and that need a careful investigation. A number of theoretical questions are raised by this research. To begin with, it is unclear what is the rate at which the bandwidth estimated by our method converges to the minimizer of the (11). There is very little research done so far on the K -fold crossvalidation (except when $K = N$ —the standard “leave-one-out” crossvalidation) even under the ordinary assumption of independent data, let alone in a case like ours that deals with correlated data. We know that in the case of “leave-one-out” “crossvalidation” having positively correlated data results in choosing a bandwidth that is smaller than the minimizer of MISE while the opposite is true for the negatively correlated data (see Chu and Marron, 1991). It is sensible to assume that having correlated data also influences K -fold crossvalidation to a degree; however, a precise theoretical investigation is very much in order here. Another interesting question that merits further discussion is whether there possibly exists an optimal value of K such that it minimizes the MISE. In much of our development here we have used $K = 10$ on the basis of simulations shown in graphs (2)–(4).

Acknowledgements

The author would like to thank his advisor, Prof. Lawrence D. Brown for his invaluable help while preparing this article.

References

- Anthony, M., Holden, S.B., 1998. Cross-validation for binary classification by real-valued functions: theoretical analysis. in: Proceedings of the International Conference on Computational Learning Theory, pp. 218–229.
- Blum, A., Kalai, A., Langford, J., 1999. Beating the hold-out: bounds for K -fold and progressive cross-validation. Proceedings of the 12th Annual Conference on Computational Learning Theory, pp. 203–208.
- Carroll, R., Ruppert, D., 1988. Transformation and Weighting in Regression. Chapman & Hall, New York.
- Chu, C.-K., Marron, J.S., 1991. Comparison of two bandwidth selectors with dependent errors. The Ann. Statist. 19, 1906–1918.
- Dette, H., 2002. A consistent test for heteroscedasticity in nonparametric regression based on the kernel method. J. Statist. Plann. Inference 103 (1–2), 311–329.
- Dette, H., Munk, A., Wagner, T., 1998. Estimating the variance in nonparametric regression—what is a reasonable choice? J. Roy. Statist. Soc. Ser. B 60 (Part 4), 751–764.
- Efromovich, S., 1999. Nonparametric Curve Estimation. Springer, Berlin.
- Fan, J., Gijbels, I., 1995. Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. J. Roy. Statist. Soc. Ser. B 57, 371–394.
- Fan, J., Yao, Q., 1998. Efficient estimation of conditional variance functions in stochastic regression. Biometrika 85, 645–660.

- Gasser, T., Müller, H.G., 1979. Kernel estimation of regression functions. in: Gasser, T., Rosenblatt, M. (Eds.), *Smoothing Techniques for Curve Estimation*, Lecture Notes in Mathematics, vol. 757, Springer, Heidelberg, pp. 23–68.
- Gasser, U., Sroka, L., Jennen-Steinmetz, C., 1986. Residual variances and residual pattern in the nonlinear regression. *Biometrika* 73, 625–633.
- Hall, P., Kay, J.W., Titterton, D.M., 1990. Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika* 77, 521–528.
- Härdle, W., Hall, P., Marron, J.S., 1988. How far are automatically chosen regression smoothing parameters from their optimum? (with discussion). *J. Amer. Statist. Assoc.* 83, 86–101.
- Härdle, W., Tsybakov, A., 1997. Local polynomial estimators of the volatility function in nonparametric autoregression. *J. Econom.* 81, 223–242.
- Härdle, W., 1990. *Applied Nonparametric Regression*. Cambridge University Press, Cambridge.
- Hastie, T., Tibshirani, R., Friedman, J., 2001. *The Elements of Statistical Learning*. Springer, Berlin.
- Levine, M., 2003. Variance estimation for nonparametric regression and its applications. Ph.D. Dissertation, University of Pennsylvania.
- Müller, H.-G., Stadtmüller, U., 1993. On variance function estimation with quadratic forms. *J. Statist. Plann. Inference* 35, 213–231.
- Neumann, M., 1994. Fully data-driven nonparametric variance estimators. *Statistics* 25, 189–212.
- Parzen, E., 1962. On estimation of a probability density function and mode. *Ann. Math. Statist.* 33, 1065–1076.
- Rice, J., 1984. Bandwidth choice for nonparametric kernel regression. *Ann. Statist.* 12, 1215–1230.
- Rosenblatt, M., 1956. Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.* 27, 832–837.
- Ruppert, D., Sheather, S.J., Wand, M.P., 1995. An effective bandwidth selector for local least squares regression. *J. Amer. Statist. Assoc.* 90, 1257–1270.
- Ruppert, D., Wand, M.P., Holst, U., Hössjer, O., 1997. Local polynomial variance-function estimation. *Technometrics* 39 (3), 262–273.
- Wand, M.P., Jones, M.C., 1995. *Kernel Smoothing*. Chapman & Hall, New York.
- Whittle, P., 1960. Bounds for the moments of linear and quadratic forms in independent variables. *Theory Probab. Appl.* 5, 302–305.