Do-validation for Kernel Density Estimation

Enno Mammen

Universität Mannheim, Abteilung Volkswirtschaftslehre

L7, 3-5, 68131-Mannheim

María Dolores Martínez Miranda

Universidad de Granada, Departamento de Estadística e I.O.

Campus Fuentenueva, E - 18071 Granada

Jens Perch Nielsen

Cass Business School, City University

106 Bunhill Row, UK - London EC1Y 8TZ

Stefan Sperlich

Université de Genève, Département des sciences économiques

Bd du Pont d'Arve 40, CH - 1211 Genève 4

December 16, 2010

Abstract

Bandwidth selection in kernel density estimation is one of the fundamental model selection problems of mathematical statistics. The study of this problem took major steps forward with the papers of Hall and Marron (1987) and Hall and Johnstone (1992). Since then focus seems to have been on various versions of implementing the so called plug-in method aimed at estimating the minimum mean integrated squared error (MISE). The most successful of these efforts still seems to be the plug-in method of Sheather and Jones (1991) or Park and Marron (1990) that we also use as a benchmark in this paper. In this paper we derive a new theorem deriving the asymptotic theory for linear combinations of bandwidths obtained from different selectors as e.g. direct and indirect cross-validation and plug-in, where we take advantage of recent advances in the study of indirect cross-validation; see Hart and Yi (1998), Hart and Lee (2005) and Savchuk, Hart and Sheather (2010a,b). We conclude that the slow convergence of data-driven bandwidths implies that once asymptotic theory is close to that of plug-in then it is the practical implementation that counts. This insight led us to a bandwidth selector search with the symmetrized version of onesided cross-validation as a clear winner. [1]

Keywords: bandwidth choice, cross-validation, plug-in, nonparametric estimation.

[1]We acknowledge very helpful comments from two anonymous referees and the editors. This research was financially supported by knowledge company Festina Lente, the Spanish Ministerio de Educacion y Ciencia, projects MTM2008-03010/MTM, and the Deutsche Forschungsgemeinschaft (DFG), Projects FOR916, and MA1026/10-2.

1 Introduction

Standard (least-squares) cross-validation was proposed by Rudemo (1982) and Bowman (1984). The simplicity of its implementation and its intuitively appealing interpretation probably makes it the most popular automatic bandwidth selection method. Its practical data-driven flavor makes up for its lack of stability in the eyes of many practitioners. The lack of stability of standard cross-validation has been pointed out for both kernel density estimation and kernel regression estimation (Hall and Marron, 1987; Härdle, Hall and Marron, 1988). In Żychaluk and Patil (2008) it has been argued that instability of cross validation is often due to discretization effects. These drawbacks of standard cross-validation have motivated several studies on more stable bandwidth selectors, most of them related to the plug-in method (Biased cross-validation by Scott and Terrell, 1987; smoothed cross-validation by Hall, Marron and Park, 1989; Sheather and Jones, 1991; the stabilized bandwidth selector rule by Chiu, 1991; the kernel contrast method of Ahmad and Ran, 2004; recent bootstrap methods, see Cao, 1993; among others). Better performance of plug-in has been questioned in Loader (1999). He points out that the value of the chosen bandwidth heavily depends on the arbitrary specification of pilot bandwidths and may be strongly biased in case of misspecification. Still quite recently, SiZer became a popular alternative, see Chaudhuri and Marron (1999), Godtliebsen, Marron and Chaudhuri (2002), and Hanning and Marron (2006). Note that SiZer does not search for an optimal data-driven bandwidth, but instead highlights for each bandwidth which features (of the density or regression) get detected.

Indirect cross-validation is another class of bandwidth selectors. These selectors make use of so-called selection kernels. A bandwidth is chosen for the selection-kernel estimator by cross-validation and this bandwidth is then rescaled by an appropriate factor to be suitable for the kernel estimator at hand, see Savchuk, Hart and Sheather (2010a). It has been argued that cross-validation is in particular performing well in harder estimation problems. Indirect Cross-validation makes use of this relation by choosing selection kernels for which bandwidth choice is a more difficult estimation problem, see e.g. Section 4 in Hart and Lee (2005) for this point. Appropriate selection kernels robustify the bandwidth choice against discretization effects and data rounding. For a detailed discussion, see Savchuk, Hart and Sheather (2010b). One version of the general principle of indirect cross-validation is onesided cross-validation. Onesided cross-validation was originally proposed by Hart and Yi (1998) for local linear regression. In Hart and Lee (2005) it was shown that in contrast to classical cross validation this approach is robust against spurious and nonspurious serial correlation. Onesided cross-validation was extended to our density case by Martínez-Miranda, Nielsen and Sperlich (2009).

In this paper we derive a new theorem on combinations of kernel density bandwidths and investigate its practical potential. This new theorem discusses the difference of the bandwidth selectors to h_{ISE} and h_{MISE} where h_{ISE} is the random bandwidth that minimizes the integrated squared error (ISE) and where h_{MISE} minimizes the mean integrated squared error (MISE). The theorem allows combinations of bandwidths from both direct and indirect cross-validation and can include bandwidth selectors as well that converge to h_{MISE} with a faster rate of convergence. While we have investi-

gated a large class of combinations through simulation studies, we only present those six bandwidth combinations that are helpful for a better understanding of our overall conclusions. First we look at the asymptotically optimal combination of the standard cross-validation bandwidth and an asymptotically MISE optimal bandwidth. With the Epanechnikov kernel it turns out that the optimal combinations pick 1.21 times the plug-in bandwidth and place a negative weight of 0.21 on the cross-validation bandwidth. This is not surprising considering the well known negative correlation between the cross-validation and the plug-in bandwidth, see Hall and Marron (1987). When the combined bandwidth is compared to an asymptotically MISE optimal bandwidth it turns out that the kernel density estimator with MISE optimal bandwidth can have an increase in its expected ISE of up to 40 %, at least in the asymptotic limit. This asymptotic advantage of the combined bandwidth does not seem to describe what is going on in finite samples. Indeed, we have implemented the two bandwidth selectors with simulated data where we used a plug-in bandwidth along Sheather and Jones (1991) and in the simulations the kernel density estimator with the combined bandwidth behaved very poorly. The reason seems to be that plug-in estimators have a tendency to oversmooth in finite samples while standard cross-validation is almost unbiased. The asymptotic optimal combination of plug-in and cross-validation suffers in practice from a tendency to oversmooth even more. From this study we learn that the slow convergence of bandwidth selection asymptotics has the consequence that we can not fully rely on theory when picking our practical bandwidth selector. To illustrate this point even more, we implemented the intuitive simple average of the plug-in bandwidth and the cross-validation bandwidth. This

simple combination is appealing because of the well known practical experience of plug-in bandwidths to oversmooth, while the almost unbiased cross-validation bandwidth sometimes have very bad performance because of undersmoothing. Therefore, intuition says that a simple average should improve both, which indeed turns out to be the case. The asymptotic performance of this simple average is much better than for ordinary cross-validation and only slightly worse than for the plug-in method. In practice it outperforms both. This insight led to a search for good finite sample performance of combinations of bandwidths with an asymptotic performance close to the plug-in method. The overall winner of this search was a simple combination of the right-sided and the left-sided versions of onesided cross-validation. We call this new bandwidth selector "do-validation" (double onesided cross-validation). The conclusion of this paper therefore suggests the do-validation bandwidth as an asymptotically well performing bandwidth selector with excellent finite sample properties.

The paper is organized as follows. In Section 2 we define a new class of combined cross-validation bandwidth selectors. In Section 3 theoretical properties of linear combinations of cross-validated bandwidths and a plug-in bandwidth are derived. In Sections 4 and 5 we consider six bandwidth selectors: the standard cross-validation $\widehat{h}_{CV}$, left and right onesided cross-validation $\widehat{h}_{L,OSCV}$ and $\widehat{h}_{R,OSCV}$, do-validation $\widehat{h}_{DO}$, a plug-in bandwidth $\widehat{h}_{PI}$ as an example of a bandwidth that is asymptotically equivalent to the MISE minimizing bandwidth h_{MISE}, an asymptotically optimal combination $\widehat{h}_{mix1}$ of $\widehat{h}_{CV}$ and $\widehat{h}_{PI}$, and finally the simple average $\widehat{h}_{mix2}$ of $\widehat{h}_{CV}$ and $\widehat{h}_{PI}$. For all our six considered bandwidths we consider their difference with the ISE optimal bandwidth h_{ISE}. The asymptotic variances of all these differences are sums of

two terms. All bandwidths have the same asymptotic first component. The second component varies with the different methods. While it is quite large for the cross-validation method, it is less than one third of this value for all other methods. The exact asymptotic values become less important than the practical performance of the implementation at hand. This is one conclusion of Section 5 where we present results of a finite sample study where do-validation $\widehat{h}_{DO}$ comes out as the best of the considered methods. Another conclusion is that combinations of certain bandwidths can improve a lot in ISE compared to their individual performance.

2 A general class of data-driven bandwidth estimators

In this paper we consider a class of bandwidth selectors $\widehat{h}$ that are constructed as weighted averages of cross-validation bandwidths $\widehat{h}_j$. The aim is to get a bandwidth with a small Integrated Squared Error (ISE) for the kernel density estimator

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

The bandwidths $\widehat{h}_j$ are based on the inspection of kernel density estimators $\widehat{f}_{h,L_j}(x), 1 \leq j \leq J$, for kernels L_j that fulfill $L_j(0) = 0$. For $1 \leq j \leq J$ we define $\widehat{h}_j$ by the cross-validation method

$$\widehat{h}_j = \arg\min_h \int \widehat{f}_{h,L_j}(x)^2 dx - 2n^{-1} \sum_{i=1}^{n} \widehat{f}_{h,L_j}(X_i). \tag{1}$$

Note that because of $L_j(0) = 0$ we do not need to use a leave-one-out version of $\widehat{f}_{h,L_j}$ in the sum on the right hand side. For some weights w_j (not necessarily positive)

with $\sum_{j=1}^{J} w_j = 1$, a new bandwidth selector $\widehat{h}$ is defined by

$$\widehat{h} = \sum_{j=1}^{J} w_j \left(\frac{R(K)}{\mu_2^2(K)} \frac{\mu_2^2(L_j)}{R(L_j)} \right)^{1/5} \widehat{h}_j \tag{2}$$

where $R(g) = \int g^2(x)dx$, $\mu_l(g) = \int x^l g(x)dx$ for functions g and integers $l \geq 0$. The bandwidth $\widehat{h}_j$ is a selector for the kernel L_j. After multiplying with the factor $(R(K)\mu_2^2(L_j))^{1/5}(\mu_2^2(K)\ R(L_j))^{-1/5}$ it becomes a selector for the kernel K. This follows from classical smoothing theory and has been used at many places in the discussion of bandwidth selectors.

Our class of bandwidth selectors contains the classical cross-validation bandwidth selector as one example with $J = 1$ and $L_1(u) = K(u)\mathbf{1}(u \neq 0)$. If one uses a leave-one-out version of (1), one would replace $\widehat{f}_{h,L_j}(X_i)$ by $n(n-1)^{-1}\widehat{f}_{h,L_j}(X_i)$ in the second term on the right hand side of (1). This change is asymptotically negligible and does not lead to obvious changes in the finite sample performance. Furthermore, there is a difference if two observations have the same value. Under our assumptions this happens only with zero probability. Thus, our discussion carries over without changes to leave-one-out versions of our proposal.

Our main proposal from the general class (2) is do-validation. It is based on the combination of left and rightsided cross-validation. Onesided cross-validation (OSCV hereafter) has been proposed by Martínez-Miranda, Nielsen and Sperlich (2009) for kernel density estimation. In their implementation they make use of the local linear kernel density estimator (Jones, 1993; and Cheng 1997a, 1997b). For a kernel density estimator $\widehat{f}_{h,M}$ with kernel M the local linear kernel density estimator can be defined

as kernel density estimator $\widehat{f}_{h,M^*}$ with "equivalent kernel" M^* given by

$$M^*(u) = \frac{\mu_2(M) - \mu_1(M)u}{\mu_0(M)\mu_2(M) - \mu_1^2(M)} M(u). \tag{3}$$

In onesided cross-validation the basic kernel $M(u)$ is chosen as $2K(u)\mathbf{1}_{(-\infty,0)}$ (leftsided cross-validation) and $2K(u)\mathbf{1}_{(0,\infty)}$ (rightsided cross-validation). This results in the following equivalent kernels

$$K_L(u) = \frac{\mu_2(K) + u\mu_1^*(K)}{\mu_2(K) - (\mu_1^*(K))^2} 2K(u)\mathbf{1}_{(-\infty,0)}, \tag{4}$$

$$K_R(u) = \frac{\mu_2(K) - u\mu_1^*(K)}{\mu_2(K) - (\mu_1^*(K))^2} 2K(u)\mathbf{1}_{(0,\infty)}, \tag{5}$$

with $\mu_1^*(K) = \int_0^\infty uK(u)du$. Here we have assumed that the kernel K is symmetric. The left-OSCV criterion (OSCV_L) is defined by

$$\mathrm{OSCV}_L(h) = \int \widehat{f}^2_{h,K_L}(x)dx - 2n^{-1}\sum_{i=1}^n \widehat{f}_{h,K_L}(X_i), \tag{6}$$

with $\widehat{h}_L$ as its minimizer; and the left-OSCV bandwidth is calculated from $\widehat{h}_L$ by

$$\widehat{h}_{L,\mathrm{OSCV}} = \mathrm{C}\widehat{h}_L,$$

where

$$\mathrm{C} = \left(\frac{R(K)}{\mu_2^2(K)} \frac{\mu_2^2(K_L)}{R(K_L)}\right)^{1/5}, \tag{7}$$

see Martínez-Miranda, Nielsen and Sperlich (2009).

In exactly the same way we define the right-OSCV criterion, OSCV_R, except that $\widehat{f}_{h,K_L}$ in (6) is replaced by $\widehat{f}_{h,K_R}$. The right-OSCV bandwidth is calculated by $\widehat{h}_{R,\mathrm{OSCV}} = \mathrm{C}\widehat{h}_R$, where C is the same as in (7) and $\widehat{h}_R$ is the minimizer of OSCV_R.

Onesided cross-validation does not work in a local constant version, i.e. with the choice $K_L(u) = 2K(u)\mathbf{1}_{(-\infty,0)}$ or $K_R(u) = 2K(u)\mathbf{1}_{(0,\infty)}$. This holds because of the

inferior rate of convergence of the onesided local constant kernel density estimator. This was the reason why Martínez-Miranda, Nielsen and Sperlich (2009) suggested the use of local linear density estimation.

The do-validation selector $\widehat{h}_{DO}$ is given by

$$\widehat{h}_{DO} = \frac{1}{2}(\widehat{h}_{L,OSCV} + \widehat{h}_{R,OSCV}).$$

Left-onesided cross-validation and right-onesided cross-validation are not identical in the local linear case because of differences in the boundary. However, asymptotically they are equivalent. As we will see in our simulations do-validation delivers a good stable compromise. It has the same asymptotic theory as each of the two onesided alternatives and a better overall finite sample performance.

3 Asymptotic theory

In this section we state an asymptotic result on the difference between a combined bandwidth selector $\widehat{h}$, defined in (2), with the MISE-optimal bandwidth h_{MISE} and the ISE-optimal bandwidth h_{ISE},

$$h_{MISE} = \arg\min_h \mathrm{E}\left[\int \left(\widehat{f}_{h,K}(x) - f(x)\right)^2 dx\right],$$
$$h_{ISE} = \arg\min_h \left[\int \left(\widehat{f}_{h,K}(x) - f(x)\right)^2 dx\right].$$

Under our assumptions, see below, it holds that $h_{MISE} = \left(\frac{R(K)}{\mu_2^2(K)R(f'')}\right)^{1/5} n^{-1/5} + o\left(n^{-3/10}\right)$.

We are also interested in $\widehat{h}^*$, with $\widehat{h}^*$ being a combination with an asymptotical MISE-optimal bandwidth $\widehat{h}_{MISE}$ defined by

$$\widehat{h}^* = \sum_{j=2}^{J} w_j \left(\frac{R(K)}{\mu_2^2(K)} \frac{\mu_2^2(L_j)}{R(L_j)} \right)^{1/5} \widehat{h}_j + w_1 \widehat{h}_{MISE}, \tag{8}$$

where $\sum_{j=1}^{J} w_j = 1$ and where $\widehat{h}_{MISE}$ is a bandwidth selector with $\widehat{h}_{MISE} = h_{MISE} + o_P(n^{-3/10})$. In the simulations we will choose $\widehat{h}_{MISE} = \widehat{h}_{PI}$ where $\widehat{h}_{PI}$ is refers to a plug-in selector along Sheather and Jones (1991) and Park and Marron (1990), respectively.

We now state a theorem about the asymptotic distribution of $\widehat{h} - h_{MISE}$, $\widehat{h} - h_{ISE}$ and $\widehat{h}^* - h_{ISE}$. For this result we need the following assumptions:

(A1) K and L_j $(j = 1, \ldots, J)$ are compactly supported. The kernels are continuous on $\mathbb{R} - \{0\}$ and have one-sided derivatives that are Hölder continuous on $\mathbb{R}$ i.e. there exist constants $c, \delta > 0$ such that $|g(x) - g(y)| \leq c|x - y|^\delta$ for $x, y < 0$ or $x, y > 0$ with g equal to K' or L_j' $(j = 1, \ldots, J)$. The left- and right-sided derivatives differ at most on a finite set. For $j = 1, \ldots, J$, $L_j(0) = 0$, $\int uL_j(u)du = 0$ and $\int uK(u)du = 0$.

(A2) The density f is bounded and twice differentiable. The derivatives f' and f'' are bounded and integrable. The second derivative is Hölder continuous with exponent $\delta > \frac{1}{2}$.

For do-validation condition (A1) is satisfied if the kernel K satisfies the required smoothness conditions. In particular, Assumption A1 allows kernels that are smooth inside their support but are not differentiable at the boundary of the support.

Theorem 1. ***Combination of bandwidths.*** *Under A1, A2 the bandwidth selector* $\widehat{h}$ *in (2) satisfies*

$$n^{3/10}(\widehat{h} - h_{ISE}) \ \rightarrow N(0, \sigma_1^2) \quad \textit{in distribution}, \tag{9}$$

$$n^{3/10}(\widehat{h} - h_{MISE}) \ \rightarrow N(0, \sigma_2^2) \quad \textit{in distribution}, \tag{10}$$

where

$$\begin{aligned} \sigma_k^2 = \ & \frac{4}{25} R(K)^{-2/5} \mu_2^{-6/5}(K) R(f'')^{-8/5} \, \mathrm{V}(f'') \delta_k + \frac{1}{50} R(K)^{-7/5} \mu_2^{-6/5}(K) \\ & \times R(f'')^{-3/5} R(f) \int \left[\delta_k H(u) - \sum_{j=1}^{J} w_j \left(\frac{R(K)}{R(L_j)} \right) H_j(u) \right]^2 du, \end{aligned} \tag{11}$$

with

$$\mathrm{V}(f'') = \int f''^2(x) f(x) dx - \left(\int f''(x) f(x) dx \right)^2,$$

$$\begin{aligned} H(u) = \ & 2 \int K(u+v) K(v) dv + 2 \int K(-u+v) K(v) dv \\ & + 2 \int K(u+v) v K'(v) dv + 2 \int K(-u+v) v K'(v) dv, \end{aligned}$$

$$\begin{aligned} H_j(d_j u) = \ & 2 \int L_j(u+v) L_j(v) dv + 2 \int L_j(-u+v) L_j(v) dv \\ & + 2 \int L_j(u+v) v L_j'(v) dv + 2 \int L_j(-u+v) v L_j'(v) dv \\ & - 2 \left[L_j(u) + u L_j'(u) + L_j(-u) - u L_j'(-u) \right], \end{aligned}$$

$$d_j = \left(\frac{R(K)}{R(L_j)} \frac{\mu_2^2(L_j)}{\mu_2^2(K)} \right)^{-1/5}, \qquad \delta_k = \begin{cases} 1 & \textit{for } k = 1, \\ 0 & \textit{for } k = 2. \end{cases}$$

For a choice of $\widehat{h}_{MISE}$ *with* $\widehat{h}_{MISE} = h_{MISE} + o_P(n^{-3/10})$*, the combination* $\widehat{h}^*$ *in (8)*

satisfies

$$n^{3/10}(\widehat{h}^* - h_{ISE}) \quad \to N(0, \sigma_1^2) \quad \textit{in distribution,} \tag{12}$$

with σ_1^2 *as in (11) but with* $H_1 = 0$.

For the special choice $J = 1, w_1 = 1$, the asymptotic expansions in (9), (10) and (12) reduce to the classical results in Hall and Marron (1987), see their Theorem 2.1 and discussions in Section 2.3. Then, equation (12) implies an expansion for the difference between the ISE optimal bandwidth h_{ISE} and the MISE optimal bandwidth h_{MISE} and equations (9) and (10) compare the classical cross-validation bandwidth selector with these two target bandwidths.

Under additional smoothness assumptions on f, Hall and Johnstone (1992) discussed efficient estimation of the ISE-optimal bandwidth h_{ISE}. They showed that estimation of h_{ISE} is asymptotically equivalent to the estimation of $R(f') = \int f'(x)^2 dx$. Using an efficient estimator of $R(f')$, one gets an estimator $\widehat{h}$ of h_{ISE} such that $n^{3/10}(\widehat{h} - h_{ISE})$ has asymptotic variance $\frac{4}{25} R(K)^{-2/5} \mu_2^{-6/5}(K) R(f'')^{-8/5} \, \mathrm{V}(f'')$. Thus, in our class of bandwidth selectors a bandwidth would achieve the optimality bound if

$$\int \left[H(u) - \sum_{j=1}^{J} w_j \left(\frac{R(K)}{R(L_j)} \right) H_j(u) \right]^2 du \quad = \quad 0. \tag{13}$$

We do not know if this can be achieved by appropriate choice of kernels L_j and weights w_j. In the next section we will discuss the size of the left hand side of (13) for some choices of the kernels L_j and weights w_j.

4 Six combinations of bandwidth selectors

For given kernels K and L_j $(j = 1, \ldots, J)$ and a density f, we might explicitly calculate the components in the asymptotic variance in (11), for the just introduced class of bandwidth selectors (2), and also for the combinations (8). We look for good selectors, in the optimality sense of Hall and Johnstone (1992), i.e. bandwidths which, holding constant the first variance term in (11), have a small second term. Hall and Johnstone (1992) showed that an asymptotically achievable bandwidth exists where the second term is zero but they never pursued the issue any further and did not provide practical examples of such bandwidth selectors. In our search for good bandwidth selectors, we do for the first time provide a bandwidth selector with better asymptotic theory than the plug-in method, namely the optimal combination of plug-in (with weight w_1) and classical cross-validation. Figure 1 shows up to a factor the graphs of the resultant second variance term in (11) against the weight for the Epanechnikov kernel. We plot the noisy term in a wide range including negative weights to get optimality (as we argued in Section 1 because of the known negative correlation between them). And indeed the optimum is achieved by weighting the plug-in bandwidth with $w_1 = 1.21$ and cross-validation with $w_2 = 1 - w_1 = -0.21$. Such an optimal combination yields a second term in (11) of $0.51C_{f,K}$ with $C_{f,K} = \frac{1}{25}R(K)^{-7/5}\mu_2^{-6/5}(K)R(f'')^{-3/5}R(f)$. With $0.72C_{f,K}$ an asymptotically MISE optimal bandwidth $\widehat{h}_{MISE}$ has a second term that is about 40 per cent above the term of the combined bandwidth. Thus, asymptotically, the expected ISE can increase up to 40 per cent. For the quartic kernel we get weights $w_1 = 1.37$ and $w_2 = -0.37$. Here

the factor of the second term increases from 0.44 to 0.83. This gives an increase of expected ISE of up to 90 per cent. However, in the next section we will see that despite the excellent asymptotic properties these combined bandwidth selectors can show poor performance in finite samples.

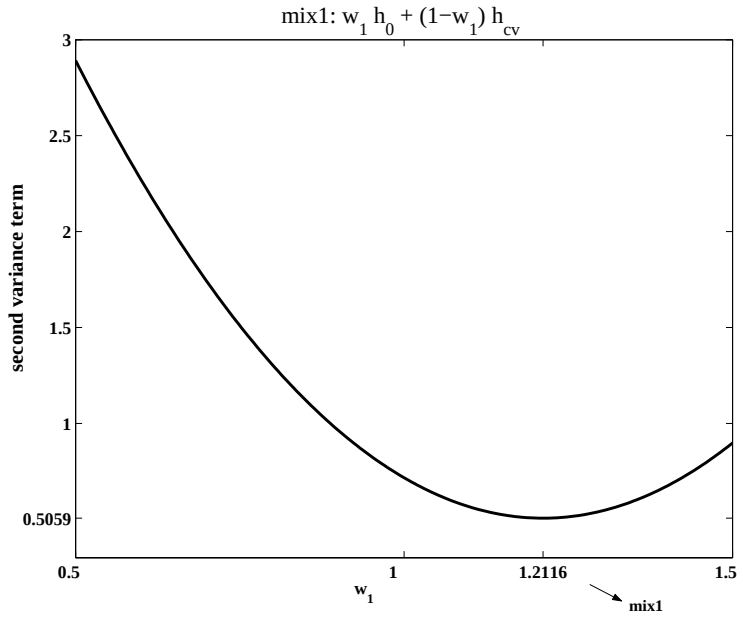

Figure 1: *The factor* $\frac{1}{2}\int\left[H(u)-\sum_{j=1}^{J}w_j\left(\frac{R(K)}{R(L_j)}\right)H_j(u)\right]^2 du$ *of the second component of the asymptotic variance in (11). The factor is plotted for the combination of an asymptotical MISE optimal bandwidth with standard cross-validation. The density estimator is calculated with an Epanechnikov kernel. For the combined bandwidth the factor is equal to* $\frac{1}{2}\int\left[w_1H(u)-(1-w_1)4(K(u)+uK'(u))\right]^2 du$. *The optimal value is achieved by weighting the plug-in bandwidth with* $w_1 = 1.21$.

In the following display we show the asymptotic variances of $n^{3/10}(\widehat{h}-h_{ISE})$ or $n^{3/10}(\widehat{h}^*-h_{ISE})$, respectively (as given by expression (11)), for each of the following six bandwidths: onesided cross-validation $\widehat{h}_{OSCV}$, do-validation method $\widehat{h}_{DO}$, the

standard cross-validation bandwidth $\widehat{h}_{CV}$, an asymptotical MISE optimal bandwidth $\widehat{h}_{MISE}$, and two combinations of $\widehat{h}_{CV}$ and $\widehat{h}_{MISE}$. The first combination $\widehat{h}_{mix1}$ is the optimal combination $\widehat{h}_{mix1} = 1.2116\widehat{h}_{MISE} - 0.2116\widehat{h}_{CV}$. The second one is the pragmatic average $\widehat{h}_{mix2} = 0.5\widehat{h}_{MISE} + 0.5\widehat{h}_{CV}$. The asymptotic variances of these bandwidths are given for the Epanechnikov kernel by:

$$\begin{aligned}
\sigma^2_{\mathrm{OSCV}} &= C_{f,K}\left\{4R(K)\frac{\mathrm{V}(f'')}{R(f'')R(f)} + 2.19\right\} \\
\sigma^2_{\mathrm{DO}} &= C_{f,K}\left\{4R(K)\frac{\mathrm{V}(f'')}{R(f'')R(f)} + 2.19\right\} \\
\sigma^2_{\mathrm{CV}} &= C_{f,K}\left\{4R(K)\frac{\mathrm{V}(f'')}{R(f'')R(f)} + 7.42\right\} \\
\sigma^2_{MISE} &= C_{f,K}\left\{4R(K)\frac{\mathrm{V}(f'')}{R(f'')R(f)} + \ 0.72\right\} \\
\sigma^2_{\mathrm{mix1}} &= C_{f,K}\left\{4R(K)\frac{\mathrm{V}(f'')}{R(f'')R(f)} + \ 0.51\right\} \\
\sigma^2_{\mathrm{mix2}} &= C_{f,K}\left\{4R(K)\frac{\mathrm{V}(f'')}{R(f'')R(f)} + \ 2.89\right\}
\end{aligned}$$

with $C_{f,K}$ as above. For the quartic kernel we get the same expressions with the constants 2.19, 2.19, 7.42, 0.72, 0.51, 2.89 replaced by 1.46, 1.46, 5.87, 0.83, 0.44, 2.63.

There are two components which inflate all the variances, but only the second term differs between selectors. We can observe a clear reduction in this second variance term in both onesided cross-validation and do-validation, compared with standard cross-validation. The asymptotic variance of asymptotical MISE optimal methods is lower than for all competitors with one exception: it is beaten by the optimal combination of itself with classical cross-validation. But as we will see in the next section, the asymptotic properties of all our considered bandwidths - except for classical cross-validation - are that good and at the same time that close to each other that

in practice these differences become irrelevant. Now it is the numerical performance that matters.

5 Finite Sample Performance

The purpose of this section is to study the performance of the six bandwidths defined in Section 4 in finite samples, sometimes just of moderate size. As asymptotical MISE optimal bandwidth we use in the simulations a plug-in bandwidth $\widehat{h}_{PI}$ as proposed by Sheather and Jones (1991); see also Park and Marron (1990). This plug-in bandwidthis calculated from the asymptotic expression of the MISE-optimal bandwidth, $h_{MISE} \approx \left(\frac{R(K)}{\mu_2^2(K) R(f'')} \right)^{1/5} n^{-1/5}$, where $R(K)$ and $\mu_2(K)$ are known, whereas $R(f'')$ has to be estimated. We now describe the estimator $\widehat{R}_{f''}$ that we used. In a first step we calculate a kernel density estimator of f'' with bandwidth g_p. For the choice of g_p we take Silverman's rule of thumb bandwidth for Gaussian kernels, see Silverman (1986, page 48). In our implementation the standard deviation of X is estimated by the minimum of two methods: the empirical standard deviation s_n and the interquartile range IR_X divided by 1.34, i.e. $g_S = 1.06 \ \min\{IR_X 1.34^{-1}, s_n\} n^{-1/5}$. As the quartic kernel K_Q comes close to the Epanechnikov but allows for estimating the second derivative, we normalize g_S by the factors of the canonical kernel (Gaussian to quartic) and adjust for the slower rate ($n^{-1/9}$) needed to estimate second derivatives, i.e.

$$g_p \ = \ g_S \frac{2.0362}{0.7764} \ n^{1/5-1/9} \ .$$

Next,

$$\widehat{R}_{f''} \;=\; \int \widehat{f}''^2 - \frac{1}{n g_p^5} \int K_Q''^2$$

to correct for the bias inherited by

$$\widehat{f}''(x) \;=\; \frac{1}{n g_p^3} \sum_{i=1}^{n} K_Q'' \left(\frac{X_i - x}{g_p} \right) \;.$$

In simulation studies not shown here this prior choice turned out to perform better than any of the many other plug-in estimators we tried, at least for the densities considered in our simulations.

Our selected data generating processes are the following six densities (see also Figure 2):

1. a simple normal distribution, $N(0.5, 0.2^2)$,

2. a bimodal mixture of two normals which were $N(0.35, 0.1^2)$ and $N(0.65, 0.1^2)$,

3. a mixture of three normals, namely $N(0.25, 0.075^2)$, $N(0.5, 0.075^2)$ and $N(0.75, 0.075^2)$ giving three clear modes,

4. a gamma distribution, $Gamma(a, b)$ with $b = 1.5$, $a = b^2$ applied on $5x$ with $x \in I\!R_+$, i.e.

$$f(x) = 5 \frac{b^a}{\Gamma(a)} (5x)^{a-1} e^{-5xb},$$

5. a mixture of two gamma distributions, $Gamma(a_j, b_j)$, $j = 1, 2$ with $a_j = b_j^2$, $b_1 = 1.5$, $b_2 = 3$ applied on $6x$, i.e.

$$f(x) = \frac{6}{2} \sum_{j=1}^{2} \frac{b_j^{a_j}}{\Gamma(a_j)} (6x)^{a_j - 1} e^{-6xb_j}$$

giving one mode and a plateau,

6. and a mixture of three gamma distributions, $Gamma(a_j, b_j)$, $j = 1, ..., 3$ with $a_j = b_j^2$, $b_1 = 1.5$, $b_2 = 3$, and $b_3 = 6$ applied on $8x$ giving two bumps and one plateau.

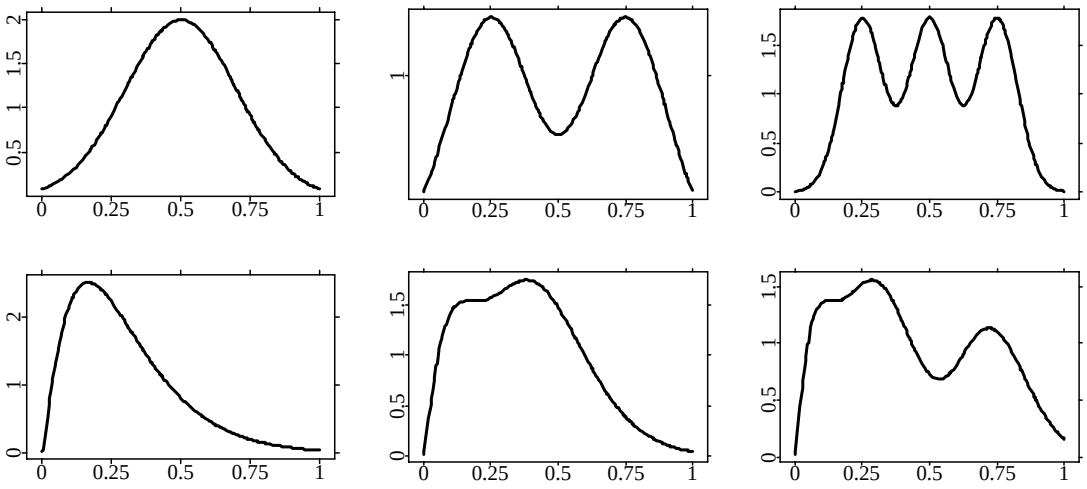

Figure 2: *The six data generating densities: Designs 1 to 6 from the upper left to the lower right.*

Our set of densities contains density functions with one, two or three modes, some being asymmetric. They all have exponentially falling tails, because otherwise one has to work with boundary correcting kernels. The main mass is always in $[0, 1]$. We use the following measures to summarize the stochastic performance of the bandwidth selectors:

$$m_1 = \text{mean}(\text{ISE}(\widehat{h})), \quad m_2 = \text{std}(\text{ISE}(\widehat{h}))$$

$$m_3 = \text{mean}(\widehat{h} - h_{ISE}), \quad m_4 = \text{std}(\widehat{h} - h_{ISE}).$$

The simulated kernel density estimators use the Epanechnikov kernel. The sample sizes are $n = 50$, $n = 100$, and 200 as examples for moderate and large samples. The

onesided cross-validation presented here is the left-onesided. The original simulation study comprised more designs, kernels, and samples sizes, but the findings were all in line with the here presented results. The results given in Tables 1 to 3 and 4 were calculated from 250 repetitions for each model and each sample size. Note that the standard deviations of the Monte Carlos means m_1 and m_3 in Tables 1-4 are simply $m_2/\sqrt{250} \approx 0.06\ m_2$ and $m_4/\sqrt{250}$, respectively.

	Design 1						Design 2					
	$\widehat{h}_{PI}$	$\widehat{h}_{CV}$	$\widehat{h}_{OSCV}$	$\widehat{h}_{DO}$	$\widehat{h}_{mix1}$	$\widehat{h}_{mix2}$	$\widehat{h}_{PI}$	$\widehat{h}_{CV}$	$\widehat{h}_{OSCV}$	$\widehat{h}_{DO}$	$\widehat{h}_{mix1}$	$\widehat{h}_{mix2}$
$n = 50$												
m_1	.047	.083	.051	.049	.049	.049	.115	.111	.106	.103	.130	.083
m_2	.033	.100	.043	.036	.033	.037	.023	.114	.033	.034	.025	.032
m_3	.047	-.011	.032	.034	.059	.018	.117	.018	.092	.090	.138	.067
m_4	.054	.095	.066	.064	.053	.069	.044	.078	.076	.072	.047	.054
$n = 100$												
m_1	.029	.049	.031	.030	.031	.030	.077	.063	.049	.049	.091	.052
m_2	.020	.059	.021	.020	.021	.020	.017	.055	.027	.026	.017	.024
m_3	.035	-.016	.019	.019	.046	.010	.098	.010	.035	.035	.116	.054
m_4	.047	.078	.057	.056	.046	.058	.024	.059	.048	.046	.022	.039
$n = 200$												
m_1	.017	.026	.018	.018	.018	.018	.049	.043	.034	.030	.059	.033
m_2	.012	.029	.014	.014	.012	.013	.013	.054	.041	.017	.012	.016
m_3	.029	-.005	.010	.011	.036	.012	.075	.000	.014	.015	.091	.037
m_4	.036	.064	.045	.044	.045	.047	.019	.044	.035	.031	.017	.029

Table 1: *Criteria m_1, m_2, m_3 and m_4 for Designs 1 and 2.*

As mentioned above the performance of cross-validation compared to plug-in can be used to classify the difficulty of data adaptive bandwidth choice. It has been argued that cross-validation performs relatively well in harder estimation problems. Using the relative performance of cross-validation to rank the difficulty of bandwidth selection in our different settings, we get Design 1 as the easiest problem, Designs

	Design 3						Design 4					
	$\widehat{h}_{PI}$	$\widehat{h}_{CV}$	$\widehat{h}_{OSCV}$	$\widehat{h}_{DO}$	$\widehat{h}_{mix1}$	$\widehat{h}_{mix2}$	$\widehat{h}_{PI}$	$\widehat{h}_{CV}$	$\widehat{h}_{OSCV}$	$\widehat{h}_{DO}$	$\widehat{h}_{mix1}$	$\widehat{h}_{mix2}$
$n = 50$												
m_1	.158	.130	.156	.156	.163	.126	.130	.138	.114	.109	.144	.105
m_2	.011	.117	.015	.016	.011	.024	.063	.124	.061	.060	.067	.058
m_3	.163	.036	.154	.154	.189	.099	.095	.007	.063	.056	.114	.051
m_4	.026	.080	.039	.037	.028	.047	.054	.074	.060	.057	.057	.057
$n = 100$												
m_1	.142	.070	.115	.115	.153	.091	.087	.078	.072	.068	.099	.066
m_2	.008	.057	.042	.036	.009	.019	.042	.054	.038	.037	.047	.035
m_3	.146	.007	.104	.106	.175	.076	.075	-.006	.037	.030	.092	.035
m_4	.012	.042	.067	.059	.012	.025	.042	.050	.049	.045	.046	.041
$n = 200$												
m_1	.120	.042	.039	.038	.136	.063	.053	.049	.048	.040	.062	.039
m_2	.006	.032	.024	.021	.008	.015	.023	.046	.054	.021	.026	.021
m_3	.127	.000	.018	.018	.154	.064	.062	-.007	.026	.019	.077	.027
m_4	.010	.030	.034	.029	.010	.019	.027	.038	.042	.033	.030	.029

Table 2: *Criteria m_1, m_2, m_3 and m_4 for Designs 3 and 4.*

2 and 4 as harder problems and Design 3 as a very hard estimation problem. For Designs 5 and 6 the difficulty depends on the sample size and it increases with sample size. We conjecture that for these two designs accurate estimation of the small wiggles of the underlying density is impossible for small sample sizes but it requires a fine tuning of the bandwidths for larger sample sizes.

The cross-validation bandwidth $\widehat{h}_{CV}$ and the asymptotically optimal combination $\widehat{h}_{mix1}$ of $\widehat{h}_{CV}$ and $\widehat{h}_{PI}$ show the poorest behaviour in the simulations. They have in all designs and for all sample sizes the largest expected ISE m_1. For $\widehat{h}_{CV}$ this is mainly caused by the large variance of this bandwidth selector, see the outcomes of m_2 and m_4. This is a well known phenomen that motivated the study of asymptotical MISE optimal bandwidths which have smaller variances, at least asymptotically. To

	Design 5						Design 6					
	$\widehat{h}_{PI}$	$\widehat{h}_{CV}$	$\widehat{h}_{OSCV}$	$\widehat{h}_{DO}$	$\widehat{h}_{mix1}$	$\widehat{h}_{mix2}$	$\widehat{h}_{PI}$	$\widehat{h}_{CV}$	$\widehat{h}_{OSCV}$	$\widehat{h}_{DO}$	$\widehat{h}_{mix1}$	$\widehat{h}_{mix2}$
$n = 50$												
m_1	.063	.093	.064	.064	.066	.061	.073	.090	.070	.070	.081	.063
m_2	.023	.097	.025	.025	.024	.027	.020	.076	.027	.027	.018	.026
m_3	.078	.008	.069	.064	.092	.043	.104	.015	.074	.071	.122	.060
m_4	.057	.099	.064	.067	.057	.073	.036	.093	.065	.066	.033	.061
$n = 100$												
m_1	.046	.055	.045	.044	.049	.040	.056	.058	.049	.048	.062	.045
m_2	.013	.045	.015	.015	.014	.015	.015	.045	.018	.019	.014	.017
m_3	.074	-.002	.059	.051	.090	.036	.096	.005	.057	.053	.115	.051
m_4	.045	.072	.055	.056	.047	.052	.028	.068	.051	.052	.026	.044
$n = 200$												
m_1	.034	.033	.032	.029	.037	.027	.042	.035	.033	.031	.048	.031
m_2	.008	.020	.010	.010	.009	.010	.009	.031	.012	.012	.009	.010
m_3	.077	.000	.055	.045	.094	.039	.093	.001	.048	.041	.113	.047
m_4	.032	.057	.047	.044	.033	.040	.023	.050	.044	.041	.022	.034

Table 3: *Criteria m_1, m_2, m_3 and m_4 for designs 5 and 6.*

understand the poor behaviour of $\widehat{h}_{mix1}$ we have carried out additional simulations where we replaced $\widehat{h}_{PI}$ in the definition of $\widehat{h}_{mix1}$ by h_{MISE}. In Table 4 we compare the theoretical bandwidth $h_{mix1} = 1.2116 h_{MISE} - 0.2116 \widehat{h}_{CV}$ with the MISE minimizer h_{MISE}. When looking at Table 4 one immediately notes that all differences between the performance h_{MISE} and h_{mix1} are insignificant. Mostly, the two methods give exactly the same values in the table. We conjecture that the second term in the asymptotic variance of the plug-in method already is so small that it is irrelevant to improve more on it from a practical point of view. Finally, we also see how far the plug-in estimates are from h_{MISE}, and so is $\widehat{h}_{mix1}$ from h_{mix1}. The difference of the latter is even much larger. This may be explained by the larger weight that is given to h_{MISE} and $\widehat{h}_{PI}$ in the definitions of $\widehat{h}_{mix1}$ or h_{mix1}, respectively.

	Design 1		Design 2		Design 3		Design 4		Design 5		Design 6	
	h_{MISE}	Mix1	h_{MISE}	Mix1	h_{MISE}	Mix1	h_{MISE}	Mix1	h_{MISE}	Mix1	h_{MISE}	Mix1
$n = 50$												
m_1	.040	.039	.059	.059	.074	.075	.083	.083	.055	.054	.054	.053
m_2	.033	.032	.037	.039	.036	.039	.049	.049	.025	.025	.026	.026
m_3	.009	.014	-.007	-.012	-.001	-.008	.007	.007	.007	.007	.006	.004
m_4	.035	.030	.034	.036	.026	.028	.028	.026	.048	.045	.036	.033
$n = 100$												
m_1	.026	.025	.036	.035	.046	.046	.055	.055	.037	.037	.039	.038
m_2	.019	.019	.020	.020	.025	.025	.032	.032	.015	.015	.016	.016
m_3	.003	.007	.006	.005	-.002	-.004	.002	.004	.000	.000	-.003	-.004
m_4	.032	.029	.017	.015	.012	.012	.024	.023	.038	.039	.028	.026
$n = 200$												
m_1	.016	.015	.025	.024	.030	.029	.034	.034	.024	.024	.025	.025
m_2	.011	.011	.014	.014	.013	.013	.019	.019	.010	.010	.010	.010
m_3	.003	.005	.005	.006	-.001	-.002	.006	.008	.003	.003	-.002	-.003
m_4	.026	.023	.014	.012	.009	.008	.018	.018	.027	.027	.023	.022

Table 4: *Values of m_1, m_2, m_3 and m_4 for h_{MISE} and h_{mix1}.*

In the simulation, with respect to m_1, the winners are $\widehat{h}_{mix2}$ and $\widehat{h}_{DO}$. The pragmatic average $\widehat{h}_{mix2}$ of classical cross-validation and plug-in beats each of its two components for almost every design and sample size (with the only exemption of Design 1). Therefore, from a practical point of view this simple average is much better than its two well known alternatives. We can get a hint of what is going on when looking on the bias m_3 and the volatility measures m_2 and m_4. We see that the stability of the plug-in method comes with the cost of a clear tendency to oversmoothing and that the unbiasedness of cross-validation comes with the cost of volatility. The simple average provides a good compromise between these two very different bandwidth selection methods. The selector $\widehat{h}_{mix2}$ works very well and probably is better than most published bandwidth selectors so far. Do-validation is quite similar to $\widehat{h}_{mix2}$ and it is

only marginally better - measured by m_1 and m_2 - than one-sided cross-validation on most of the designs. But for the asymmetric density design 4 it is clearly better. For the bimodal density 2 and the trimodal density 3 $\widehat{h}_{mix2}$ outperforms $\widehat{h}_{DO}$ whereas for the trimodal density with larger sample size 200, $\widehat{h}_{mix2}$ has an increase in the expected ISE of 66 %. The simplicity of do-validation, no need of the choice of pilot bandwidth and its overall excellent performance in this simulation make do-validation a very promising bandwidth selector. To conclude again from this study, classical cross-validation is a crystal clear loser of this test. It is almost unbiased and that is good, but volatility just kills its overall performance. Therefore we suggest that practitioners leave classical cross-validation and start to use do-validation. Do-validation is just another cross-validation technique, it is relatively simple to carry out, it is well defined without ambiguities and does not need complicated pilot estimation.

6 Conclusions

In this paper we have studied the use of combined bandwidth selectors for kernel density estimation. We have compared averages of indirect cross-validation bandwidth together with and without asymptotical MISE optimal selectors. The study was led by an asymptotic theory for this class of bandwidth selectors but we also pointed out the limitations of asymptotics in the study of bandwidth selectors. We showed that there is some potential in this class for outperforming plug-in and classical cross-validation. Our practical recommendation is do-validation, a bandwidth selector that is comparable to plug-in in its asymptotic properties but that showed a much better and more stable performance in our simulation study. Do-validation is also a very

simple procedure that does not need any pilot estimation and that is to implement as simple as classical cross-validation.

We think that there is some further need for research on the use of indirect cross-validation, in kernel density estimation and in other smoothing problems, maybe even in other smoothing parameter and model choice problems. For kernel density estimation there are some more interesting results on indirect cross-validation in the recent papers of Savchuk, Hart and Sheather (2010a,b). In their asymptotic approach they compare bandwidths with the MISE optimal bandwidth h_{MISE}. This differs from our asymptotic study where we compare bandwidths with the ISE optimal bandwidth h_{ISE}. They showed that one can use indirect cross-validation to get a bandwidth selector that is asymptotically equivalent to the MISE optimal bandwidth h_{MISE}. Thus their estimator competes with plug-in estimators as the Sheather and Jones (1991) plug-in bandwidth, but it does not need any pilot estimation. The key idea of their paper is to do indirect cross-validation using an selection kernel depending on two parameters, one of which is a function of n. While this pilot-free MISE optimal estimation does not need any pilot density, it does need to determine somehow a good trade off between these two parameters. The paper Savchuk, Hart and Sheather (2010a) does contain some practical proposals but we think the choice of these parameters and the comparison with other selection-kernels in the indirect cross-validation needs some further research. We think that the asymptotic result of Savchuk, Hart and Sheather (2010a) is also very interesting from a purely theoretic point of view. Indeed, if one plugs their bandwidth selector into our combined formula for $\widehat{h}_{mix1}$ one gets a bandwidth selector that is purely based on cross-validation

principles and beats plug-in asymptotically.

We consider pilot-free MISE optimal estimation and pilot-free MISE near optimal estimation an important area of future research in kernel density bandwidth selection. The key research element of pilot-free MISE optimal estimation seems to be to determine a practical solution to the trade off between the two entering components in the indirect kernel. The discussion on pilot-free MISE optimal estimation also gives us some intuition to why onesided cross-validation and do-validation work so well. While both of these two latter methods are practical and easy to implement, they are a first step towards a pilot-free MISE optimal estimator. In the first step it does blow up variance to some extent while keeping bias relatively stable. However, it does not bother to optimize this idea asymptotically by blowing up the variance indefinitely in the indirect step with all the practical problems this implies. One-sided cross-validation and do-validation are practical and pragmatic first steps towards pilot free MISE optimal estimation, but with easy stable implementations that work extremely well in practice.

Appendix

Proof of Theorem 1.

For $L = K$ and $L = L_j$ $(j = 1, \ldots, J)$ we use the following notation. Define

$$\begin{aligned}
\Delta_L(h) &= \int \left(\widehat{f}_{L,h}(x) - f(x)\right)^2 dx \quad \text{(ISE)}, \\
\mathrm{M}_L(h) &= \mathrm{E}\left[\Delta_L(h)\right] \quad \text{(MISE)}, \\
\mathrm{D}_L(h) &= \Delta_L(h) - \mathrm{M}_L(h), \\
\delta_L(h) &= 2\int f(x)\widehat{f}_{L,h}(x)dx - 2n^{-1}\sum_{i=1}^{n} \widehat{f}_{L,h}(X_i).
\end{aligned}$$

Here $\widehat{f}_{L,h}(x)$ denotes the kernel density estimator with bandwidth h and kernel L. Define

$$\begin{aligned}
h_{L,0} &= \arg\min_h \mathrm{M}_L(h), \\
\widehat{h}_{L,0} &= \arg\min_h \Delta_L(h), \\
\widehat{h}_{L,c} &= \arg\min_h \left(\Delta_L(h) + \delta_L(h) - \int f(x)^2 dx\right) \quad \text{(CV-bandwidths)}.
\end{aligned}$$

Under our conditions it holds that

$$h_{L,0} = \left(\frac{R(L)}{\mu_2^2(L)R(f'')}\right)^{1/5} n^{-1/5} + o\left(n^{-3/10}\right).$$

Proceeding as in Hall and Marron (1987) one can show that for $L = K, L_1, \ldots, L_J$:

$$\widehat{h}_{L,0} - h_{L,0} = -\mathrm{M}_L''(h_{L,0})^{-1}\,\mathrm{D}_L'(h_{L,0}) + o_p\left(n^{-3/10}\right), \tag{14}$$

$$\widehat{h}_{L,c} - h_{L,0} = -\mathrm{M}_L''(h_{L,0})^{-1}\left(\mathrm{D}_L'(h_{L,0}) + \delta_L'(h_{L,0})\right) + o_p\left(n^{-3/10}\right). \tag{15}$$

For the proof of these statements it can be checked that it is not needed that L is symmetric and continuous at the point 0. Furthermore, we will argue now that one can allow that the kernels are only piecewise differentiable, as assumed in Assumption (A1). The first step for getting the expansions (14)-(15) for differentiable kernels is

to note that:

$$\begin{aligned} \mathrm{M}_L'(h_{L,0}) &= 0, && (16) \\ \Delta_L'(\widehat{h}_{L,0}) &= 0, && (17) \\ \Delta_L'(\widehat{h}_{L,c}) + \delta_L'(\widehat{h}_{L,c}) &= 0. && (18) \end{aligned}$$

Equation (16) still holds under Assumption (A1). We now argue that equations (17)-(18) remain to hold under Assumption (A1) if the right hand sides of the equations are replaced by $O_P(n^{-2}h^{-2}) = O_P(n^{-8/5})$. To see this one has to note that for a finite set $T \subset I\!R$ the following event has probability zero: there exists an $h > 0$ such that for two pairs (i, j) and (i', j') it holds that $h^{-1}(X_i - X_j) \in T$ and $h^{-1}(X_{i'} - X_{j'}) \in T$. This statement follows easily because the observations have a density. Thus we have that with probability equal to one for all $h > 0$ at most for one pair (X_i, X_j) it holds that $h^{-1}(X_i - X_j)$ lies on a point where the kernel L is not differentiable. Consider now $\Delta_L(\widehat{h}_{L,0})$ and $\Delta_L(\widehat{h}_{L,c}) + \delta_L(\widehat{h}_{L,c})$. These are double sums over indices (i, j). Using our considerations we get that with probability one the left- and right-sided derivatives of the summands differ only for at most one double index (i, j). Now one uses a simple bound for the one-sided derivatives of the summands that is of the order $O(n^{-2}h^{-2})$. Thus we have that (17)-(18) hold with the right hand sides replaced by $O_P(n^{-8/5})$. One can check that this change does not affect the lines of proof used in Hall and Marron (1987).

One can show that for $L = K, L_1, \ldots, L_J$ with $h = h_{L,0}$:

$$\mathrm{D}_L'(h) = n^{-2} \sum_{i<j} W_{L,i,j} + n^{-1} \sum_i W_{L,i} + o_p\left(n^{-7/10}\right) \qquad (19)$$

with $W^*_{L,i,j} = -h^{-2} H_L\left(\frac{X_i - X_j}{h}\right)$,

$$H_L(u) = 2\int L(u+v)L(v)dv + 2\int L(-u+v)L(v)dv + 2\int L(u+v)vL'(v)dv + 2\int L(-u+v)vL'(v)dv,$$

$$\begin{aligned} W_{L,i,j} &= W^*_{L,i,j} - \mathrm{E}\left[W^*_{L,i,j}|X_i\right] - \mathrm{E}\left[W^*_{L,i,j}|X_j\right] + \mathrm{E}\left[W^*_{L,i,j}\right], \\ W^*_{L,i} &= 2h\sqrt{\mu_2^2(L)}f''(X_i), \\ W_{L,i} &= W^*_{L,i} - \mathrm{E}\left[W^*_{L,i}\right]. \end{aligned}$$

For a proof of (19) note that $\mathrm{D}'_L(h)$ is a U-statistic with quadratic terms $W^*_{L,i,j}$. We replace $W^*_{L,i,j}$ by $W_{L,i,j}$ to have that $\mathrm{E}\left[W_{L,i,j}|X_i\right] = \mathrm{E}\left[W_{L,i,j}|X_j\right] = 0$, a.s. The remaining terms are sums of independent mean zero variables. By standard smoothing theory expansions it can be shown that these summands are asymptotically equivalent to $W_{L,i}$.

Furthermore, we have for $h = h_{L,0}$ and $L = L_1, \ldots, L_J$

$$\delta'_L(h) = n^{-2}\sum_{i<j} V_{L,i,j} - n^{-1}\sum_i W_{L,i} + o_p\left(n^{-7/10}\right)$$

with $V^*_{L,i,j} = h^{-2} G_L\left(\frac{X_i - X_j}{h}\right)$, $G_L(u) = 2\left[L(u) + uL'(u) + L(-u) - uL'(-u)\right]$.

We now use

$$\begin{aligned} h_{K,0} &= \left(\frac{R(K)}{\mu_2^2(K)}\frac{\mu_2^2(L_j)}{R(L_j)}\right)^{1/5} h_{L_j,0} + o\left(n^{-3/10}\right) \\ &= \left(\frac{R(K)}{\mu_2^2(K)}\frac{\mu_2^2(L_j)}{R(L_j)}\right)^{1/5} h_{L_j,0} + o\left(n^{-3/10}\right). \end{aligned}$$

This gives together with the above expansions:

$$\begin{aligned}\widehat{h} - h_{ISE} &= \sum_{j=1}^{J} w_j \left(\frac{R(K)}{\mu_2^2(K)} \frac{\mu_2^2(L_j)}{R(L_j)}\right)^{1/5} (\widehat{h}_{L_j,c} - h_{L_j,0}) - (h_{ISE} - h_{K,0}) + o\left(n^{-3/10}\right) \\ &= \sum_{j=1}^{J} w_j M''_{L_j}(h_{L_j,0})^{-1} \left(\frac{R(K)}{\mu_2^2(K)} \frac{\mu_2^2(L_j)}{R(L_j)}\right)^{1/5} \left[-n^{-2} \sum_{i<k} \left(W_{L_j,i,k} + V_{L_j,i,k}\right)\right] \\ &\quad -M''_K(h_{MISE})^{-1} \left[-n^{-2} \sum_{i<k} W_{K,i,k} - n^{-1} \sum_i W_{K,i}\right] + o_p\left(n^{-3/10}\right). \quad (20)\end{aligned}$$

Now

$$\begin{aligned}M''_L(h_{L,0}) &= 2\frac{R(L)}{n h_{L,0}^3} + 3\mu_2^2(L) R(f'') h_{L,0}^2 + o\left(n^{-2/5}\right) \\ &= n^{-2/5} 5 R(L)^{2/5} R(f'')^{3/5} \mu_2^{6/5}(L) + o\left(n^{-3/5}\right)\end{aligned}$$

With the above expansion this gives

$$\begin{aligned}\widehat{h} - h_{ISE} &= M''_K(h_{MISE})^{-1} n^{-1} \sum_i W_{K,i} \\ &\quad + M''_K(h_{MISE})^{-1} n^{-2} \sum_{i<k} Z_{ik} + o_p\left(n^{-3/10}\right)\end{aligned}$$

with $Z_{ik} = Z^*_{ik} - \mathrm{E}\left[Z^*_{ik}|X_i\right] - \mathrm{E}\left[Z^*_{ik}|X_j\right] + \mathrm{E}\left[Z^*_{ik}\right]$,

$$\begin{aligned}Z^*_{ik} &= -h_{0,K}^{-2} \left[H_K\left(\frac{X_i - X_k}{h_{MISE}}\right)\right. \\ &\quad \left. - \sum_{j=1}^{J} w_j \left(\frac{R(K)}{R(L_j)}\right) \left(H_{L_j}\left(\frac{X_i - X_k}{h_{0,L_j}}\right) + G_{L_j}\left(\frac{X_i - X_k}{h_{0,L_j}}\right)\right)\right]\end{aligned}$$

Note that we collect in the definition of Z^*_{ik} all quadratic terms in the right hand side of (20). These are the terms: $-n^{-2} w_j M''_{L_j}(h_{L_j,0})^{-1} \left(\frac{R(K)}{\mu_2^2(K)} \frac{\mu_2^2(L_j)}{R(L_j)}\right)^{1/5} W_{L_j,i,k}$, $-n^{-2} w_j$ $M''_{L_j}(h_{L_j,0})^{-1} \left(\frac{R(K)}{\mu_2^2(K)} \frac{\mu_2^2(L_j)}{R(L_j)}\right)^{1/5} V_{L_j,i,k}$ and $n^{-2} M''_K(h_{MISE})^{-1} W_{K,i,k}$.

The variance of the asymptotic expansion of $\widehat{h} - h_{ISE}$ can be easily calculated. Furthermore, using a central limit theorem for U-statistics (e.g. Hall, 1984) one gets the asymptotic result for $\widehat{h} - h_{ISE}$ in our theorem. The second statement of the theorem can be proved similarly.

References

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

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

Cao, R., 1993, Bootstrapping the Mean Integrated Squared Error, *Journal of Multivariate Analysis*, **45**, 137–160.

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

Cheng, M.Y., 1997a, Boundary-aware estimators of integrated squared density derivatives. *Journal of the Royal Statistical Society Ser. B*, **50**, 191–203.

Cheng, M.Y., 1997b, A bandwidth selector for local linear density estimators. *The Annals of Statistics*, **25**, 1001–1013.

Chiu, S.T., 1991, Bandwidth selection for kernel density estimation. *The Annals of Statistics*, **19**, 1883–1905.

Godtliebsen, F.; Marron, J.S. and Chaudhuri, P., 2002, Significance in Scale Space for Bivariate Density Estimation. *Journal of Computational and Graphical Statistics*, **11**, 1–21.

Härdle, W., Hall, P. and Marron, J.S., 1988, How far are automatically chosen regression smoothing parameters from their optimum?. *Journal of the American Statistical Association*, **83**, 86–99.

Hall, P., 1984, Central Limit Theorem for Integrated Square Error of Multivariate Nonparametric Density Estimators. *Journal of the Multivariate Analysis*, **14**, 1–16.

Hall, P. and Johnstone, I., 1992, Empirical Functionals and Efficient Smoothing Parameter Selection. *Journal of the Royal Statistical Society B*, **54** (2), 475–530.

Hall, P. and Marron, J.S., 1987, Extent to which Least-Squares Cross-Validation Minimises Integrated Square Error in Nonparametric Density Estimation. *Probability Theory and Related Fields*, **74**, 567–581.

Hall, P., Marron, J.S. and Park, B., 1992, Smoothed cross-validation. *Probability Theory and Related Fields*, **92**, 1–20.

Hanning, J. and Marron, J.S., 2006, Advanced Distribution Theory for SiZer. *Journal of the American Statistical Association*, **101**, 484–499.

Hart, J.D., and Lee, C.-L., 2005, Robustness of one-sided cross-validation to autocorrelation. *Journal of Multivariate Statistics*, **92**, 77–96.

Hart, J.D. and Yi, S., 1998, One-Sided Cross-Validation. *Journal of the American Statistical Association*, **93**, 620–631.

Jones, M.C., 1993, Simple boundary correction in kernel density estimation. *Statistics and Computing*, **3**, 135–146.

Loader, C.R., 1999, Bandwidth selection: classical or plug-in. *The Annals of Statistics*, **27**, 415–438.

Martínez-Miranda, M.D., Nielsen, J. and Sperlich, S., 2009, One sided cross-validation for density estimation with an application to operational risk. *In "Operational Risk Towards Basel III: Best Practices and Issues in Modelling. Management and Regulation, ed. G.N. Gregoriou; John Wiley and Sons, Hoboken, New Jersey.*

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

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

Savchuk, O.Y., Hart, J.D., and Sheather S.J., 2010a, Indirect cross-validation for Density Estimation. Submitted to *Journal of the American Statistical Association*, **105**, 415–423.

Savchuk, O.Y., Hart, J.D., and Sheather S.J., 2010b, An empirical study of indirect cross-validation for Density Estimation. *IMS Lecture Notes - Festschrift for Tom Hettmansperger.*

Scott, D.W. and Terrell, G.R., 1987, Biased and unbiased cross-validation in density estimation. *Journal of the American Statistical Association*, **82**, 1131–1146.

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

Silverman, B. W., 1986, Density Estimation. *London: Chapman and Hall.*

Żychaluk, K. and Patil, P.N., 2008, A cross-validation method for data with ties in kernel density estimation. *Annals of the Institute of Statistical Mathematics*, **60**, 21–44.