AStA Adv Stat Anal (2013) 97:403–433
DOI 10.1007/s10182-013-0216-y

ORIGINAL PAPER

Bandwidth selection for kernel density estimation: a review of fully automatic selectors

Nils-Bastian Heidenreich · Anja Schindler · Stefan Sperlich

Received: 5 December 2011 / Accepted: 22 May 2013 / Published online: 2 June 2013

Abstract On the one hand, kernel density estimation has become a common tool for empirical studies in any research area. This goes hand in hand with the fact that this kind of estimator is now provided by many software packages. On the other hand, since about three decades the discussion on bandwidth selection has been going on. Although a good part of the discussion is about nonparametric regression, this parameter choice is by no means less problematic for density estimation. This becomes obvious when reading empirical studies in which practitioners have made use of kernel densities. New contributions typically provide simulations only to show that the own selector outperforms some of the existing methods. We review existing methods and compare them on a set of designs that exhibit few bumps and exponentially falling tails. We concentrate on small and moderate sample sizes because for large ones the differences between consistent methods are often negligible, at least for practitioners. As a byproduct we find that a mixture of simple plug-in and cross-validation methods produces bandwidths with a quite stable performance.

Keywords Bandwidth selection · Kernel density estimation

The authors thank Maria Dolores Martinez-Miranda, Lijian Yang, two anonymous referees and Göran Kauermann for helpful discussion and comments.

N.-B. Heidenreich · A. Schindler
Faculty of Economic Sciences, Georg-August Universität Göttingen, Platz der Göttinger Sieben 5,
37073 Göttingen, Germany

S. Sperlich (✉)
Departement des sciences conomiques and Research Center for Statistics, Universite de Geneve,
40 Bd du Pont d'Arve, 1211 Geneva 4, Switzerland
e-mail: stefan.sperlich@unige.ch

1 Introduction

Suppose some i.i.d. data $X_1, X_2, \dots, X_n$ from a common distribution with density $f(\cdot)$ are observed, and one aims to estimate this density using the standard Parzen–Rosenblatt kernel estimator

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

where K is a kernel and h the so-called bandwidth parameter.

The problem is then to find a reliable data-driven estimator for the optimal bandwidth. To assess the performance of $\widehat{f_h}$, generally accepted measures are the integrated squared error

$$\mathrm{ISE}(h) = \mathrm{ISE}\{\widehat{f_h}(x)\} = \int \{\widehat{f_h}(x) - f(x)\}^2 \,\mathrm{d}x \tag{2}$$

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

$$\mathrm{MISE}(h) = \mathrm{MISE}\left[\hat{f}_h(x)\right] = \int \mathrm{MSE}\left[\hat{f}_h(x)\right] \mathrm{d}x. \tag{3}$$

Let us denote the minimizers of these two criteria by $\widehat{h_0}$ and h_0 respectively. The main difference is that ISE(h) is a stochastic process indexed by $h > 0$, while MISE(h) is a deterministic function of h. Based on these criteria, we distinguish two classes of methods: the cross-validation methods trying to estimate $\widehat{h_0}$ (and, therefore, looking at the ISE), and the plug-in methods which try to minimize the MISE to find h_0. It is obvious that these criteria coincide asymptotically but not for finite samples.

Part of the community working on nonparametric statistics has accepted that there may not be a perfect procedure to select the optimal bandwidth. Nevertheless, one should be able to say which is a reasonable bandwidth selector, at least for a particular problem. The so-called SiZer method tries to indicate what is a range of reasonable bandwidths and is, therefore, quite attractive for data snooping, see Chaudhuri and Marron (1999) for an introduction, Godtlibsen et al. (2002) for an extension to the bivariate case, and Hanning and Marron (2006) for an interesting modification using extreme value theory. However, SiZer does not give back any specific data-driven bandwidth. Therefore, the development of bandwidth selectors has been going on, so that we believe that a review and comparison of existing selectors would be quite helpful to get an idea of their objective and performance.

We counted more than 30 bandwidth selectors, several of them being modifications made for particular estimation problems. So we decided to limit our study to the following restrictions. Firstly, we consider only independent observations. Secondly, we look at L_2, not L_1-based methods, see also our discussion below. Boundary problems are not discussed, because it is hard to say how they can be combined with the problem of bandwidth selection. In our simulation comparison, we concentrate on rather small and moderate sample sizes, and on quite smooth densities. The considered degree of smoothness covers a broad range of problems in any research area but excludes sharp peaks and highly oscillating functions.Notice that the latter problems should not be tackled with kernels anyway.Density problems with extreme tails are not included. It is well known that those problems should be solved by data transformation; see e.g. Wand et al. (1991) or Yang and Marron (1999) for parametric, and Ruppert and cline (1994) for nonparametric transformations.After an appropriate data transformation, the remaining estimation problem falls into the here considered smoothness class (though, may be, with boundary problems). Note that the limitation to global bandwidths is not that restrictive, and even quite common in density estimation. Moreover, when the covariates X were transformed such that a similar smoothness can be assumed over the whole (transformed) support,using a global bandwidth is a quite reasonable choice. Finally, we have limited our study to already published methods.

The idea of cross-validation methods goes back to Rudemo (1982) and Bowman (1984), but we should also mention in this context the so-called pseudo-likelihood CV methods invented by Habbema et al. (1974) and by Duin (1976). Due to the lack of stability of this method, see e.g. Wand and Jones (1995), different modifications have been proposed like the stabilized bandwidth selector of Chiu (1991a,1991b, 1992), the smoothed CV proposed by Hall et al. (1992), the modified CV (MCV) of Stute (1992) or the one of Feluch and Koronacki (1992), and most recently the one-sided CV of Martinez-Miranda et al. (2009), and the indirect CV by Savchuk et al. (2010). The biased CV (BCV) of Scott and Terrell (1987) is minimizing the asymptotic MISE, like plug-in methods do, but uses a jack-knife procedure (therefore called CV) to avoid the use of prior information. Methods that mingle different selectors or density estimators were proposed by Ahmad and Ran (2004), calling it kernel contrast method, and by Mammen et al. (2011), proposing the do-validation method.

Compared to CV, the so-called plug-in methods do minimize a different objective function, namely the MISE instead of the ISE; they are less volatile but not entirely data adaptive as they require some pilot information. In contrast, CV allows to choose the bandwidth without making assumptions about the smoothness class (or the like) to which the unknown density belongs. Plug-in methods have a faster convergence rate compared to CV. Unfortunately, they can heavily depend on the choice of pilots; but if we have excellent pilot estimators, then the performance of plug-in methods is pretty good. Among these selectors, Silverman (1986) rule-of-thumb method is probably the most popular one. Various refinements were introduced, like for example by Park and Marron (1990), Sheather and Jones (1991), or by Hall et al. (1991). The bootstrap methods of Taylor (1989) and all its modifications (cf. Cao 1993, or Chacon et al. 2008) are counted into the plug-in methods as they aim to minimize the MISE.

There are already several papers dealing with a comparison of different automatic data-driven bandwidth selection methods. But they are actually older than 10 years. In the 1970s and the early 1980s, survey papers about density estimation were published by Wegman (1972), Tartar and Kronmal (1976), Fryer (1977), Wertz and Schneider (1979), Bean and Tsokos (1980), etc. An introduction and comparison to various methods of smoothing parameter selection was released by Marron (1988a) and by Park and Marron (1990). Then, extensive simulation studies were published by Park and Turlach (1992), Cao et al. (1994) and Chiu (1996). A brief survey was provided by Jones et al. (1996a) with a comprehensive simulation study published in Jones et al. (1996b). Somewhat later, also Loader (1999) published a comparison paper, partly as a reply to Jones et al. (1996b). To our knowledge, only Chacon et al. (2008), published a comparison study in more recent years. However, they concentrated on Bootstrap methods and only compared them with classical CV and the plug-in version of Sheather and Jones (1991). While all these focused on the L_2 norm, for the L_1 view we refer to Devroye and Gyorfi (1985) for density estimation, to Devroye and Lugosi (1996) for an optimal bandwidth choice, and to Devroye (1997) for a comprehensive comparison study.

The general criticism against the two classes of selection methods can be summarized as follows: CV leads to undersmoothing and is known to hardly stabilize for large data sets (they often just choose the smallest possible value among all bandwidths), whereas plug-in depends on prior information and typically works badly for small samples.

To make some statements about asymptotic theory, we use the following assumptions on the kernel and on the density.

- (A1) The kernel K is a compactly supported density function on $\mathbb{R}$, symmetric around zero, and has a Holder-continuous derivative K'.
- (A2) It holds $\mu_2(K) < \infty$, where $\mu_l(K) = \int u^l K(u) \mathrm{d}u$.
- (A3) The density f is bounded and twice differentiable; f' and f'' are bounded and integrable, and f'' is uniformly continuous.

For some methods, we will have to modify these conditions.

In our simulation study, we restrict to selection methods not using higher-order kernels. Recall that the main motivation for the application of higher-order kernels is their theoretical advantage of faster asymptotic convergence rates. However, their substantial drawback is a loss in the practical interpretability as they involve negative weights and might, therefore, give negative density estimates, see also Marron (1994).

In the context of asymptotic properties of bandwidth selectors, there is a trade-off between the classical plug-in method and standard cross-validation. The plug-in has always a smaller (asymptotic) variance compared to cross-validation (see Hall and Marron 1987a) but often a larger bias in practice. To our knowledge, no other bandwidth selector has so far outperformed the asymptotic properties of the sophisticated plug-in methods. Although Hall and Jonstone (1992) stated that such methods must theoretically exist, they could not give any practical example.

2 A brief review of previous reviews

The study of Park and Turlach (1992) comprised least square cross-validation, the biased cross-validation (BCV) of Scott and Terrell (1987), the plug-in method (SJPI) of Sheather and Jones (1991), the plug-in method (PM) of Park and Marron (1990), the smooth cross-validation (SCV) of Hall et al. (1992), and a modified version (bandwidth factorized SCV by Jones et al. (1991). The algorithms were discussed in the appendix but there was no discussion about motivation, derivation or statistical properties. They considered the estimation performance of mixtures of uni-, bi-and tri-modal normal densities along three criteria: the mean integrated squared error, the mean integrated absolute error, and mode detection. To our knowledge, that study has been published only as working paper.

Jones et al. (1996a,b) gave in their first paper a brief survey including rule-of-thumb (ROT), CV, BCV, SJPI, and finally the smooth bootstrap method. They mentioned other selectors like that of Chiu (1992) or Hall et al. (1992) without giving further details. Their findings mainly coincided with those of Cao et al. (1994) who considered less selectors, and only a few, quite smooth densities but also some qualitative measures like the so-called IP- or the double kernel method. Jones et al. (1996b) took samples of sizes $n = 100$ and $n = 1{,}000$ to estimate 15 different mixtures of normal densities. Along the quality measures they chose, their own SJPI bandwidth selector performed best.

Chiu (1996) extended the set of bandwidth selectors considered in Park and Turlach (1992) by his various stabilized methods, but he used just one specific bias-related criterion to show the superiority of his methods. His article is actually neither a review nor a general (simulation) or comparison study.

Loader (1999) replied to the then often claimed superiority of plug-in methods on several fronts. He compared them with CV methods for density estimation and regression, looking into the sources of differences. He argued that plug-in methods were heavily dependent on arbitrary specifications of pilot bandwidths and failed when this specification was wrong. He considered the likelihood based CV together with its approximation by an Akaike-style criterion, the classical CV, SJPI, BCV, and the fixed point iterations (GKK) approach of Gasser, Kneip and Köhler (1991). A detailed simulation study comparing them all was not performed. Instead, he compared some selectors along real data, and the methods CV, BCV, SJPI and GKK by some particular simulations. Half of the paper was dedicated to regression.

Sheather (2004) gave a practical description of kernel density estimation revising some estimation and bandwidth selection methods which he considered to be the most popular at that time, together with software advise, a new modifications (data sharpening), and a real data application. Simulations or comparison studies were not provided.

Devroye (1997) presented the doubtless largest and most extensive comparison study with discussion. The three main differences to all the other studies (including ours) were that: first,he looked at the L_1 measures when studying the asymptotic properties. Second, he considered different kernel density estimators. Third, for the estimation of about half of the densities from which he draw the samples in his simulations one faces serious boundary problems or problems with jumps. Consequently, he considered partly 'quite' and partly 'slightly' different bandwidth selectors. For these four reasons, all other studies are hardly comparable with that one.

Nonetheless, our choice of considered densities has partly been guided by his sample even if we show only the results for a tiny subsample. Furthermore, as it has turned out that for large samples, most of the selectors behave pretty well with diminishing differences, we have concentrated in our simulation study on small ($n \geq 25$) to at most moderate ($n \leq 200$) samples sizes. We also tried with $n = 500$ and $n = 1{,}000$; but the only new findings were that the indirect CV methods reveal their (asymptotic) superiority, whereas the leave-one-out cross-validation can easily recommend bandwidths close to zero (depending on the real underlying density). This is also the moment when the modified cross-validation (MCV) of Stute (1992) starts to become attractive. See also our comment below on data rounding.

3 Cross-validation methods in density estimation

Recall the performance measure 'integrated squared error' (ISE)

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

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

Ordinary least squares cross-validation

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

$$\min_h\ \mathrm{CV}(h) = \int \widehat{f}_h^2(x)\,\mathrm{d}x - 2\,\frac{1}{n}\sum_{i=1}^{n}\widehat{f}_{h,-i}(X_i).$$

Stone (1984) showed that under the assumptions (A1)–(A3), the minimizing argument $\widehat{h}_{\mathrm{CV}}$ fulfills

$$\mathrm{ISE}(\widehat{h}_{\mathrm{CV}})/\mathrm{ISE}(\hat{h}_0) \xrightarrow{a.s.} 1.$$

However, Hall and Marron (1987a) stated that this happened at a very slow rate; specifically

$$\begin{aligned} &n^{3/10}(\widehat{h}_{\mathrm{CV}} - \widehat{h}_0) \longrightarrow N(0, \sigma^2 c^{-2}) \qquad \text{and} \\ &n\left(\mathrm{ISE}(\widehat{h}_{\mathrm{CV}}) - \mathrm{ISE}(\hat{h}_0)\right) \longrightarrow \frac{1}{2}\sigma^2 c^{-1}\chi_1^2 \end{aligned} \tag{4}$$

under assumptions (A1)–(A3), and with terms σ and c depending only on f and K. Many practitioners use this classical CV method nonetheless because of its intuitive definition and simple implementation.

Recall the criticism saying that this classical CV lacks stability (even) when the sample size increases. Silverman (1986) and others gave explanations based on considerations of what happens if the distances $|x_i - x_j|$ become very small for many observations $j \neq i$. Chiu (1991a) studied the problem occurring with data rounding such that one obtains many ties ($x_j = x_i$ for $i \neq j$). Based on these considerations, the following stabilized and modified CV versions emerged.

Stabilized bandwidth selection

Based on characteristic functions Chiu (1991a,b, 1992) gave an expression for $wide\hat{h}_{\mathrm{CV}}$ which revealed the source of variation. Note that the CV criterion is approximately equal to the expression

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

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

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

For the minimizer $\widehat{h}_{\mathrm{ST}}$ it can be shown that $\widehat{h}_{\mathrm{ST}} \xrightarrow{a.s.} \widehat{h}_0$, and that it converges to h_0 even at the optimal $\sqrt{n}$-rate. Note that in the literature this procedure is often counted among the plug-in methods as it rather minimizes the MISE than the ISE. In our implementation, when calculating Λ we came across with the computation of square roots of negative terms in our simulations. To avoid complex numbers, we calculated the absolute value of the radicand.

Modified cross-validation

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

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

yielding the criterion

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

It can be shown that under assumptions (A1),

- (A2$'$) K is three times differentiable, with $\int t^4|K(t)|\,\mathrm{d}t < \infty$, $\int t^4|K''(t)|\,\mathrm{d}t < \infty$, $\int t^4[K'(t)]^2\,\mathrm{d}t < \infty$, and $\int t^2[K'''(t)]^2\,\mathrm{d}t < \infty$,
- (A3$'$) f four times continuously differentiable, the derivatives being bounded and integrable,

one gets the consistency result

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

One-sided cross-validation

Marron (1986) made the point that ”the harder the estimation problem the better CV works”. Based on this idea, Martinez-Miranda et al. (2009) proposed to first apply CV to a harder estimation problem, and to afterward calculate the corresponding bandwidth for the underlying ‘real’ estimation problem. To make the estimation problem harder, they used a worse estimator, still (1) but with a local linear version of a one-sided kernel,

$$L(u) = \frac{\mu_2(K) - u\left(2\int_{-\infty}^0 tK(t)\,\mathrm{d}t\right)}{\mu_2(K) - \left(2\int_{-\infty}^0 tK(t)\,\mathrm{d}t\right)^2} 2K(u)\mathbf{1}_{\{u<0\}}.$$

This modification goes back to Hart and Yi (1998) who did this for regression. One defines the one-sided versions of ISE and MISE with their minimizers $\widehat{b}_0$ and b_0, and the criterion becomes

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

where $\widehat{f}_{\mathrm{left},b}$ is the one-sided (to the left) kernel density estimator. The corresponding bandwidth for the ‘real’ estimation problem is then given by

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

Note that C is deterministic, depending only on kernel K because of

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

This gives, for example $C \approx 0.537$ for the Epanechnikov kernel. The theoretical justification for the stability of one-sided CV is based on the result of Hall and Marron (1987a), recall Eq. (4). That result allows to calculate the variance reduction of OSCV compared to CV by $\{C\bar{\sigma}c/(\bar{c}\sigma)\}^2$ where $\bar{c}, \bar{\sigma}$ are just as c, σ but with kernel L instead of K. Note that L can also be constructed as a one-sided kernel to the right.

Indirect cross-validation

Based on the same idea, Savchuk et al. (2010) proposed to use

$$L(u) = (1+\alpha)\phi(u) - \alpha\varsigma^{-1}\phi(u\varsigma^{-1}),$$

where ϕ is the Gaussian kernel, and $\alpha > 0, \varsigma > 0$ have to be chosen appropriately. They demonstrated the excellent theoretical properties of such an 'indirect method', and discussed the robustness of the indirect methods to data rounding (see above or Density estimation 1986). For the two additional prior parameter (α, ς) they made several proposals derived from their asymptotic theory. Specifically, they first recommended for $100 \le n \le 500{,}000$ to take the values

$$\alpha = 10^{3.39-1.093\log_{10}(n)+0.025\log_{10}^3(n)-0.00004\log_{10}^6(n)},$$
$$\varsigma = 10^{-0.58+0.386\log_{10}(n)-0.012\log_{10}^2(n)}.$$

But based on asymptotic and practical considerations, the following rule is proposed:

$$(\alpha, \varsigma) = (2.42; \max(5.06, 0.149n^{3/8})),$$

where the max function chooses always 5.06 unless $n > 12,094$. For our implementation with Epanechnikov kernels, their method worked well only for pretty large samples whatever proposal for choosing (α, ς) we tried.

Further cross-validation methods

Feluch and Koronacki (1992) proposed to cut out not only X_i when estimating $f(X_i)$ but rather dropping also the $m < n$ nearest neighbors with $m \to \infty$ such that $m/n \to 0$. The idea is similar to the CV selection for time series data, cf. Hardle and Vieu (1992). Like Stute (1992), they called this version *modified CV*. Unfortunately, it turned out that the quality of this method crucially depended on the choice of m. As we could not find any recommendation for its choice, this method cannot be classified as one being *automatic* or *data driven*, and would not be considered further.

Scott and Terrell (1987) introduced the B(iased)CV. As they worried about unreliable small-sample results, i.e. the high variability of CV, they directly focused on minimizing the asymptotic MISE. The unknown term $||f''(x)||_2^2$ was estimated via jack-knife methods. Already in their own paper they admitted a poor performance for small samples and mixtures of densities, see also Chiu (1996). Jones et al. (1996b) underlined in their simulation study its deficient performance (‘quite good’ to ‘very poor’) even when referring to situations where it still seemed to be a reasonable selector.

The S(moothed)CV was evolved by Hall et al. (1992). The general idea was a kind of presmoothing of the data before applying the CV criterion. This procedure of presmoothing results in smaller sample variability but enlarges the bias. Therefore, the resulting bandwidth is often oversmoothing and cuts off some important features of the underlying density. With this method, it is possible to achieve the optimal $\sqrt{n}$ rate of convergence—but only when using a kernel of order ≥ 6. So it seems to be appropriate for huge samples only.Jones et al. (1996b) showed that without such a higher-order kernel, there exists an $n^{-1/10}$ convergent version of SCV that is identical to Taylor’s bootstrap method (Taylor 1989), and is closely related to the bootstrap method of Cao (1993). These methods do not belong to the cross-validation methods, and hence, will be discussed later. Additionally, with a special choice of pilot bandwidth (necessary in all these methods), the SCV results in an $n^{-5/14}$ convergent version that is similar to the so-called diagonal-in selector of Park and Marron (1990). In conclusion, we have not implemented the SCV, because its similarity to other methods and because we did not want to use higher-order kernels for samples with $n < 500$.

The P(artitioned)CV was suggested by Marron (1988b). He modified the CV criterion by splitting the sample of size n into m subsamples. The PCV is calculated by minimizing the average of the score functions of the CV-score for all subsamples. In a final step, the resulting bandwidth needs to be rescaled. The number of subsamples affects the trade-off between variance and bias. Hence, the choice of a pilot m is the crucial problem in this case, and as Park and Marron (1990) noticed: “this method ... is not quite fully objective”. It further requires a large sample size to get subsamples of reasonable sizes.

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

Wegkamp (1999) suggested a method being very much related to the CV technique providing a quasi-universal bandwidth selector for bounded densities. This was based on a optimality concept of Devroye and Lugosi (1996) but translated to the L_2-norm context. Among other problems in practice, the procedure requires sample splitting which can be quite problematic for small and moderate sample sizes, see above. His paper stayed on a rather technical level without providing any algorithm or how to do for example the sample splitting in practice.

4 Plug-in methods in density estimation

Under (A1)–(A3) the MISE can be written as

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

for $n \to \infty$, $h \to 0$, such that the asymptotically optimal bandwidth is then equal to

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

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

Park and Marron's plug-in

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

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

where g is a prior bandwidth.

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

$$\widehat{||f''||_2^2} = ||\widehat{f}_g''||_2^2 - \frac{1}{ng^5} ||K''||_2^2, \tag{6}$$

which is used in (5) to obtain

$$\hat{h} = \left(\frac{||K||_2^2}{\widehat{||f''||_2^2} \mu_2^2(K) n} \right)^{1/5}. \tag{7}$$

The question which arises is how to obtain a proper prior bandwidth g. In Park and Marron (1990), g was the minimizer for the asymptotic mean squared error of $\widehat{||f''||_2^2}$. With (7), one gets a prior bandwidth g in terms of $\hat{h}$ (using the notation in the original paper)

$$g = C_3(K) C_4(f) \hat{h}^{10/13}, \tag{8}$$

where $C_3(K)$ contains the fourth derivative and convolutions of K, and $C_4(f)$ contains the second and third derivatives of f. The optimal (g, h_{PM}) can be obtained by numerical solution of the Eqs. (7) and (8). The relative rate of convergence to h_0 is of order $O_p(n^{-4/13})$, which is suboptimal compared to the optimal $\sqrt{n}$-rate, cf. Hall and Marron (1991).

Sheather and Jones' plug-in

An often cited method is the so-called Sheather and Jones (1991) bandwidth, see also Jones and Sheather (1991). They used the same idea like Park and Marron (1990) but replaced the 'diagonal-out' estimator of $||f''||_2^2$ by their 'diagonal-in' version to avoid the problem that the estimator $\widehat{||f''||_2^2}$ (see (6)) may give negative results. They stated that the non-stochastic term in (6) was subducted because of its positive effect on the bias in estimating $||f''||_2^2$. The idea was to choose the prior bandwidth g such that the negative bias due to the smoothing compensates the impact of the diagonal-in terms. As a result they estimated $||f''||_2^2$ by $||\widehat{f}_g''||_2^2$ which is always positive, and obtained

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

where $C(K, L)$ depends on L, the kernel used to estimate $||f''||_2^2$. As usual, K indicates the kernel of the original estimation. Then, $||f''||_2^2$ and $||f'''||_2^2$ were estimated using $||\widehat{f}_a''||_2^2$ and $||\widehat{f}_b'''||_2^2$, where a and b were set equal to the rule-of-thumb bandwidths of Silverman. Sheather and Jones (1991) showed that their optimal bandwidth had a convergence rate of $n^{-5/14}$ which is slightly better than that of Park and Marron (1990). Using real data, Jones et al. (1996b) found that $\widehat{h}_{\mathrm{SJ}}$ was pretty close to Park and Marron's bandwidth in practice. Hence, without beating that one in practical performance, having only a slightly better convergence rate, but being computationally much more expensive, we suppressed $\widehat{h}_{\mathrm{SJ}}$ in favor of the (simplified) Jones et al. (1991) bandwidth.

Implemented refined plug-in

For small samples and small (optimal) bandwidths, the above estimator $\widehat{||f''||_2^2}$ can easily fail in practice. Also, to find a numerical solution for (g, h_{PM}) may become quite difficult in practice; the final result depend on stopping rules and there might exist multiple local maxima for the finite-sample two-dimensional problem. To avoid these inconveniences, and to offer a quick and easy solution, we propose to first take Silverman's rule-of-thumb bandwidth for Gaussian kernels, i.e. $h_S = 1.06 \min\{1.34^{-1}\mathrm{IR}, s_n\}n^{-1/5}$ with interquartile range (IR) of X, and s_n the sample standard deviation. Then adjust h_S for Quartic kernels along the idea of canonical kernels and equivalence bandwidths, see Hardle et al. (2004). The Quartic kernel is pretty similar to the Epanechnikov kernel but allows for the estimation of second derivatives. Then, adjusting for the slower optimal rate for second derivative estimation gives as a prior for (6)

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

This bandwidth leads to very reasonable estimates of the second derivative of f, and hence of $||f''||_2^2$. A further advantage is that this prior g is rather easily obtained. For the rest we follow Park and Marron (1990) and call the resulting bandwidth $\widehat{h}_{PM}$ because this is a simplified version of their ideas.

Bootstrap methods

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

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

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

$$\begin{aligned} \mathrm{ISE}^*(h) &:= \int \left(\widehat{f}_h^*(x) - \widehat{f}_g(x)\right)^2 \mathrm{d}x, \\ \mathrm{MISE}^*(h) &:= E_*\left[\int \left(\widehat{f}_h^*(x) - \widehat{f}_g(x)\right)^2 \mathrm{d}x\right]. \end{aligned} \tag{9}$$

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

$$V^*(h) = \frac{1}{nh} \cdot ||K||_2^2 + \frac{1}{n} \cdot \int \left(\int K(u) \cdot \widehat{f}_g(x - hu)\,\mathrm{d}u\right)^2 \mathrm{d}x \tag{10}$$

and the integrated squared bias

$$SB^*(h) = \int \left(\int K(u) \cdot (\widehat{f}_g(x - hu) - \widehat{f}_g(x))\,\mathrm{d}u\right)^2 \mathrm{d}x \tag{11}$$

one obtains

$$V^*(h) = \frac{1}{nh}||K||_2^2 + \frac{1}{n^3} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[(K_h \star K_g) \star (K_h \star K_g)\right](X_i - X_j), \tag{12}$$

where $\star$ denotes convolution, and

$$SB^*(h) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[(K_h \star K_g - K_g) \star (K_h \star K_g - K_g)\right](X_i - X_j). \quad (13)$$

In practice, it is hard to get explicit formulae for these integrals when the kernel has a bounded support. However, using the Gaussian kernel in (12) and (13), we can directly calculate the optimal bandwidth as the minimizer of

$$\text{MISE} * (h) = \frac{1}{2nh\sqrt{\pi}} + \frac{1}{\sqrt{2\pi}}\left[\frac{\sum_{i,j}\left(\exp\left(-\frac{1}{2}\left(\frac{X_i - X_j}{g\sqrt{2}}\right)^2\right)\right)}{\sqrt{2g^2}\cdot n^2} - \frac{2\cdot\sum_{i,j}\left(\exp\left(-\frac{1}{2}\left(\frac{X_i - X_j}{\sqrt{h^2+2g^2}}\right)^2\right)\right)}{\sqrt{h^2+2g^2}\cdot n^2} + \frac{(n+1)\sum_{i,j}\left(\exp\left(-\frac{1}{2}\left(\frac{X_i - X_j}{\sqrt{2(h^2+g^2)}}\right)^2\right)\right)}{\sqrt{2(h^2+g^2)}\cdot n^3}\right]. \quad (14)$$

The equivalent bandwidth for any other kernel can be obtained as described in Marron and Nolan (1988) or Hardle et al. (2004).

The bootstrap approach in kernel density estimation was first presented by Taylor (1989). However, many modified versions were published later on, see for example Faraway and Jhun (1990), Hall (1990) or Cao (1993). The crucial differences between these versions are how they choose the pilot bandwidth g, and they generate the bootstrap samples.

Taylor (1989) suggested to take $g = h$ and used a Gaussian kernel. Several authors pointed out that this procedure had no finite minimum and hence would choose a local minimum or the upper limit of the bandwidths grid as its optimum. Marron (1992) showed that this led to an inappropriate choice and a serious positive bias. Differing from this approach, Faraway and Jhun (1990) proposed a least-square cross-validation estimate to find g. Hall (1990) recommended to use the empirical distribution to draw bootstrap samples of size $m < n$, proposed $m \simeq n^{1/2}$, $h = g(m/n)^{1/5}$, and minimized MISE* with respect to g. Cao et al. (1994) demonstrated that the bootstrap version of Hall was quite unstable and showed a bad performance, especially for mixtures of normal distributions. They found also that the methods of Faraway and Jhun (1990) as well as the one of Hall (1990) were outperformed by the method of Cao (1993) which we introduce below.

A bias corrected bootstrap estimate was developed by Grund and Polzehl (1997). They obtained an root-n convergent bandwidth estimate which attained very good results for larger sample sizes, but less so for moderate and small samples. Moreover, to derive their asymptotic theory they had to use extraordinarily strong assumptions compared to the other methods. In their simulation study, Grund and Polzehl showed that the performance heavily depended on the choice of g. They stated that using their oversmoothing bandwidth (that guaranteed root-n convergence) seemed to be far from optimal for smaller sample sizes. In contrast, using $g = h$ would achieve better performance in practical applications. However, setting $g = h$ results in a convergence rate of order $n^{-1/10}$. Summing up, they remarked that faster rates of convergence do not result in better practical performance.

In the smoothed bootstrap version of Cao (1993), the pilot bandwidth g is estimated by asymptotic expressions of the minimizer of the dominant part of the mean squared error. For further details see Cao (1993). He noticed that in (13) for $i = j$ this term would inflate the bias artificially. He, therefore, proposed to use a modified bootstrap integrated squared bias, namely

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

Concerning the convergence rate, he showed for his bandwidth, say h_0^*,

$$\frac{\mathrm{MISE}(h_0^*) - \mathrm{MISE}(h_0)}{\mathrm{MISE}(h_0)} = O_P(n^{-5/7})$$
$$\frac{\mathrm{MISE}(h_{0_M}^*) - \mathrm{MISE}(h_0)}{\mathrm{MISE}(h_0)} = O_P(n^{-8/13}).$$

Note that the convergence rate for the original bootstrap version was slightly faster.

Recently, Chacon et al. (2008), published a bootstrap version quite similar to Cao (1993). They showed that the asymptotic expressions of his bandwidth estimates might be inadequate and defined an expression $g(h)$ for fixed h. They proposed estimators for g, and allowed for different kernels L and K for the bandwidths g and h. They stated that their approach was a good compromise between classical cross-validation and plug-in. However, its performance depended seriously on the reference density. Exploring the asymptotics, they achieved root-n convergence only under the use of higher-order kernels.

In sum, in our simulation study, we concentrate on just one representative of the class of bootstrap estimates, going back to Cao (1993). He proved that the pilot bandwidth g as the minimizer of (9) coincides with the minimizer of the dominant part of the mean squared error. Specifically, it is given by

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

This formula is used for the pilot bandwidth g when calculating (14). In our simulations, we additionally ran the bootstrap for the Epanechnikov kernel calculating formulae (10) and (11) numerically. As this was much slower and gave uniformly worse results, we will neglect that approach for the rest of the paper.

Further plug-in methods

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

Hart et al. (1991) introduced a plug-in method giving back a bandwidth $\widehat{h_{\text{HSJM}}}$ which achieved the optimal $\sqrt{n}$-rate of convergence. A problem with their bandwidth $\widehat{h_{\text{HSJM}}}$ was that they used higher-order kernels to ensure the $\sqrt{n}$ convergence (actually a kernel of order 6 or higher). It is well known (see Marron and Wand 1992) that albeit their theoretical advantages, higher-order kernels have a surprisingly bad performance in practice, at least for moderate sample sizes. Furthermore, in the simulation study of Park and Turlach (1992) $\widehat{h_{\text{HSJM}}}$ was generally bad for bi- and trimodal densities.

Jones et al. (1991) developed a plug-in method based on the SCV idea, see above. They used the prior bandwidth $g = C(f)n^p h^m$, where the normal density was used as a reference distribution to calculate the unknown constant $C(f)$. The advantage of their estimator was the $\sqrt{n}$ convergence rate if $m = -2$ and $p = \frac{23}{45}$ even for kernels of order 2. In their simulation studies, Turlach (1994) and Chiu (1996) found a small variance compared to CV, but an unacceptable huge bias.

Also Kim et al. (1994) showed the existence of a $\sqrt{n}$ convergent method without the use of higher-order kernels. The main idea of obtaining asymptotically best bandwidth selectors was based on an exact MISE expansion. But the results of their paper were primarily provided for theoretical completeness; the practical performance in simulation studies was rather disappointing, which was already explicitly mentioned in their own paper and also confirmed later in a simulation study performed by Jones et al. (1996b).

For the sake of completeness, we also refer to the 'Double Kernel method' of Devroye (1989) and Jones (1998). This method has the advantage to be quite universal. Under some particular assumptions, it coincides with Taylor's bootstrap selector, respectively the BCV method, see above. As already mentioned, these two methods had several disadvantages, and also the Double Kernel method required the use of higher-order kernels. In Jones (1998), the performance of the Double Kernel method was assessed by comparing asymptotic convergence rates, but it did not exhibit the expected improvement in the estimation of h_0 (MISE optimal bandwidth) compared for example to SCV.

5 Mixing methods in density estimation

Recall that all authors have criticized that the cross-validation criterion tends to undersmooth and suffers from high sample variability. At the same time, the plug-in estimates deliver a much more stable estimate but typically oversmooth. These findings suggest to combine different bandwidths or density estimators.

Mixing the bandwidths: Do-validation

Recently, Mammen et al. (2011) took the idea of indirect cross-validation of which OSCV is a particular case, and extended it to the idea of mixing bandwidth selectors. For these mixtures they calculated the asymptotic properties and derived numerically optimal weights for particular cases. They considered

$$\widehat{h} = \sum_{j=1}^{J} w_j \left(\frac{R(K)}{\mu_2^2(K)} \frac{\mu_2^2(L_j)}{R(L_j)} \right)^{1/5} \widehat{h}_j \tag{15}$$

for some weights w_j (not necessarily positive) with $\sum_{j=1}^{J} w_j = 1$, where the $\widehat{h}_j$ were bandwidth estimates based on selector methods with selection kernel L_j, see above. After multiplying with the factor $(R(K)\mu_2^2(L_j))^{1/5}(\mu_2^2(K)\ R(L_j))^{-1/5}$ one gets a selector for a density estimator with kernel K. They further looked at

$$\widehat{h}^* = \sum_{j=2}^{J} w_j \left(\frac{R(K)}{\mu_2^2(K)} \frac{\mu_2^2(L_j)}{R(L_j)} \right)^{1/5} \widehat{h}_j + w_1 \widehat{h}_{\mathrm{MISE}}, \tag{16}$$

with an asymptotically MISE-optimal bandwidth $\widehat{h}_{\mathrm{MISE}}$. For all these selectors they showed that

$$\begin{aligned} & n^{3/10}(\widehat{h} - \hat{h}_0) \to N(0, \sigma_1^2) \quad \text{in distribution}, \\ \text{and} \quad & n^{3/10}(\widehat{h} - h_0) \to N(0, \sigma_2^2) \quad \text{in distribution}. \end{aligned}$$

Explicit expressions for σ_1 and σ_2 were given in that paper for all kind of (mixtures of) bandwidth selectors. For $J = 2$ and L_2 being the left-sided version of K, they found that the asymptotically optimal weights were $w_2 = 1 - w_1 = -0.21$ in (16), and $w_2 = 1 - w_1 = 0.5$ in (15) with L_1 being the right-sided version of K. They recommended mixing left-sided CV with right-sided CV, calling it Do-validation. Finally, they compared the asymptotics and finite sample behavior of their proposals with three standard methods.

Mixing the estimators: the contrast method

Ahmad and Ran (2004) proposed a kernel contrast method for choosing bandwidths either minimizing the ISE or alternatively the MISE. Their idea is as follows. Choose J different kernels K_j with arbitrary contrast coefficients a_j and positive weights b_j such that $\sum_j a_j = 0$, $\sum_j b_j = 1$. Then construct the contrast $\sum_j a_j \widehat{f_h}(x; K_j)$ with $\widehat{f_h}$ as in (1) but with different kernels K_j. Find $\hat{h}$ that minimizes the ISE (or the MISE, respectively) of the contrast

$$\int \left(\sum_j a_j \widehat{f_h}(x; K_j) \right)^2 \mathrm{d}x.$$

Take as a final density estimator

$$\widetilde{f}(x) = \sum_j b_j \widehat{f}_{\lambda\hat{h}}(x; K_j), \quad \text{where } \lambda^5 = \frac{\left(\sum_j a_j \sigma_j^2\right)^2 R(\varphi)}{\left(\sum_j b_j \sigma_j^2\right)^2 R(\psi)}$$

$$\varphi(u) = \sum_j a_j K_j(u), \quad \psi(u) = \sum_j b_j K_j(u), \quad \text{and} \quad \sigma_j^2 = \mu_2(K_j).$$

The evident problem is that one has to choose J and needs two series of coefficients which can have a serious impact on the performance of the method, especially for small and moderate sample sizes. We doubt that practitioners will opt for a method that even increases the number of prior parameter to be chosen—and this even by an arbitrary amount—instead of reducing it. Note that different choices lead to different outcomes. As we are not aware of any reasonable data driven method to choose them, we will not consider this bandwidth selector in the simulation study.

Further mixing methods

We are aware of the existence of other approaches which combine various density estimators by using a mixture of their smoothing parameters. In the literature, we found some papers that addressed the problem of linear and/or convex aggregation, e.g. Rigollet and Tsybakov (2007), Samarov and Tsybakov (2007) as well as Yang (2000). However, as the main focus of our review paper is not on the aggregation of different density estimators, we will not investigate this further in detail, but only study some mixtures of bandwidths which, admittedly, arise from intuition[1]. More specifically, having in mind that CV undersmoothes and PI oversmoothes, and that bandwidths are scaling parameters which should be combined on a logarithmic (i.e.multiplicative) scale, we will consider $(\widehat{h}_{\mathrm{CV}}^{\alpha} \widehat{h}_{\mathrm{PM}}^{\beta})^{1/(\alpha+\beta)}$ with $\alpha = 1,\ \beta = 2$ (mix1), $\alpha = 2,\ \beta = 1$ (mix2), and $\alpha = \beta = 1$ (mix3).

6 Finite sample performance

The small sample performance of the different cross-validation, plug-in and bootstrap methods is compared. For obvious reasons, we limit the study to data adaptive methods without boundary correction. Although we tried many different designs we summarize here the results for six densities where the estimation results expose pretty well the main findings, in particular:

1. Laplace distribution $f(x) = 4\exp(-|8(x-0.5)|)$
2. Simple Gamma distribution: Gamma(a, b) with $b = 1.5$, $a = b^2$ applied on $5x$ with $x \in \mathbf{R}$, i.e. $f(x) = 5\frac{b^a}{\Gamma(a)}(5x)^{a-1}\mathrm{e}^{-5xb}$
3. Mixture of three gamma, Gamma(a_j, b_j), $a_j = b_j^2$, $b_1 = 1.5$, $b_2 = 3$ and $b_3 = 6$ applied on $8x$ giving two bumps and one plateau

[1] We are grateful to the comments and suggestions of one of the anonymous referees.

4. Simple normal distribution, $\mathcal{N}(0.5, 0.2^2)$ with only one mode
5. Mixture of $\mathcal{N}(0.35, 0.1^2)$ and $\mathcal{N}(0.65, 0.1^2)$ giving two modes
6. A triple mode mixture of $\mathcal{N}(0.25, 0.075^2)$, $\mathcal{N}(0.5, 0.075^2)$, and $\mathcal{N}(0.75, 0.075^2)$

As can be seen in Fig. 1, all densities have the main mass in [0, 1] with exponentially decreasing tails. So that we can neglect possible boundary effects. We also simulated estimators with boundary corrections getting results very close to what we have found in the here presented study.

We have compared the performance by several measures based on the integrated squared error (ISE) of the resulting density estimate (not the bandwidth estimate), and on the distance to the numerically ISE minimizing bandwidth, say $h_{\text{opt}} \approx \widehat{h_0}$ (of each simulation run, as it is sample-dependent). The considered performance measures were

c_1: mean$(\hat{h} - h_{\text{opt}})$, bias of the estimated bandwidth

c_2: mean $\left[\text{ISE}(\hat{h})\right]$, the average (or expected) ISE

c_3: std $\left[\text{ISE}(\hat{h})\right]$, the volatility of the ISE

c_4: mean $\left(\left[\text{ISE}(\hat{h}) - \text{ISE}(h_{\text{opt}})\right]^2\right)$, squared L_2 distance of the ISEs

c_5: mean $\left[\mid \text{ISE}(\hat{h}) - \text{ISE}(h_{\text{opt}}) \mid\right]$, L_1-distance of the ISEs

We also considered other indicators of quality but will concentrate now only on these as we believe that they are the most meaningful ones. Instead of looking at

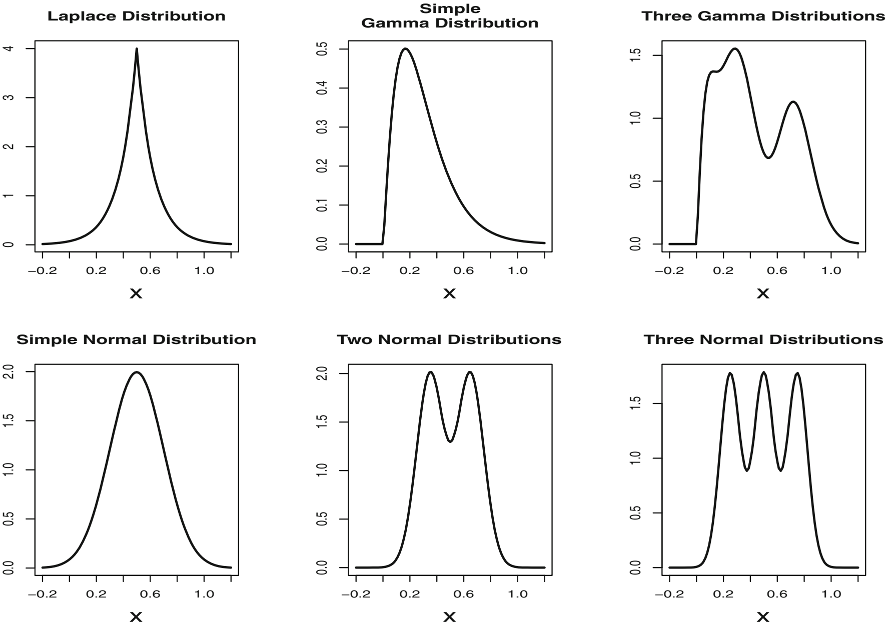

Fig. 1 The data generating densities: design 1–6 from the *upper left* to the *lower right*

(c_2, c_3), one can certainly look at the ISE distribution as a whole, for example, via box-plots.

We studied almost all selection methods, excluding the non-automatic ones and those having proved to perform uniformly worse than their competitors. In the presentation of the results, we concentrate on the methods which delivered the best results. Hence, some methods were dropped, e.g. the MCV sometimes provides multiple minima with a global minimum being far outside the range of reasonable bandwidths. We do neither show results for the indirect cross-validation since for small and moderate samples it just works worse than OSCV. Among the bootstrap methods, we concentrate on the presentation of the version (14) of the Smoothed Bootstrap which has achieved the best results among all bootstrap methods. For our mixed version (CV with refined plug-in), we first concentrate on mix3 when comparing it to the other selection methods, and later sketch the results comparing different mixtures.

While it is clear that one-sided CV and Do-validation give almost identical results for symmetric distributions, it is also clear that the latter will be more robust when asymmetry is present but unknown. We, therefore, skipped all results for Do-validation and refer to the paper of Mammen et al. (2011) instead. Notice that they only considered additive mixtures.

Summarizing, we present the following methods: CV (cross validation), OSCV-l (one-sided CV to the left), OSCV-r (oscv to the right), STAB (stabilized), RPI (refined plug-in), SBG (smooth bootstrap with Gaussian kernel—the results refer to the equivalent bandwidth for the Epanechnikov kernel), Mix 1/2 (our mix3), and as a benchmark the ISE (infeasible ISE minimizing h_{opt}). For all methods, the bandwidth search is performed on the same bandwidth grid of 25 bandwidths on a logarithmic scale from n^{-1} to 1. We give only results referring to sample sizes $n = 25, 50, 100$, and $n = 200$.

6.1 Simulation results

To summarize and compare the different bandwidth selectors, we first consider the selected bandwidths and the corresponding biases for each method separately. Afterward, we compare the methods by various performance measures. All results are based on 250 simulation runs.

Comparison of the bias for the different bandwidths

In Fig. 2, we illustrate the Bias (c_1) for the different methods for different sample sizes and distributions.

Let us consider the cross-validation method (CV). Many authors have mentioned the lack of stability of the CV criterion and its tendency to undersmooth. In Fig. 2, we see that CV has the smallest bias for all sample sizes and densities (except for the simple normal for which the mix3 is competitive). This is simply due to the fact that CV chooses always a smaller bandwidth than the other selectors. When the ISE optimal bandwidth is indeed very small, CV does, therefore, very well. However, CV clearly undersmoothes in the case of the simple normal distribution as id does for all rather smooth densities.

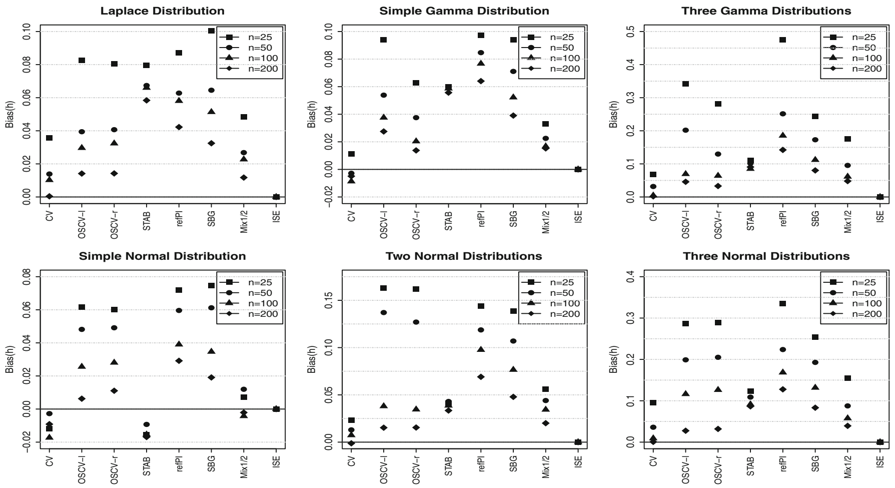

Fig. 2 Comparison of the BIAS for different sample sizes and different densities

The one-sided versions (OSCV) are quite stable. Regarding the bias, they are neither the best nor the worst over all sample sizes and models. As already stated by the original authors, the OSCV tends to overestimate the bandwidth a little bit. While for $n = 25$, the OSCV is outperformed by almost all other methods, the bias problem disappears rapidly for increasing n. For $n = 100$ and 200 we see that their biases become much smaller than for the other methods except CV (and STAB in the simple normal case). Moreover, they never fail dramatically when $n > 25$. This feature is an intuitive benefit of this method when in practice the underlying density is completely unknown. For the densities studied, the differences between the left-(OSCV-l) and the right-sided (OSCV-r) versions are negligible except for the gamma distributions because of its asymmetry.

The stabilized procedure of Chiu (STAB) is excellent for the simple normal case but it falls short when estimating other densities confirming that ”when the true density is not smooth enough, the stabilized procedure is more biased toward oversmoothing than CV”, see (Chiu 1991a,b). This fact can be seen well in Fig. 2 where STAB has increasing difficulties, compared to CV, with an increasing number of bumps. Even though this method demonstrates here a reasonable performance, the results should be interpreted with care, since in the derivation of Λ one has to deal with complex numbers. This problem we solved in favor of this method for the context of our simulations such that all presented results are clearly biased in favor of STAB.

The refined plug-in (refPI) and the smoothed bootstrap SBG show a similar behavior, though the bias of the SBG is mostly smaller than for refPI. Both are worse than STAB for small samples but outperform it for increasing n. Not surprisingly, in general, the bias for the MISE minimizing methods is larger than for all others. This partly results from the fact that we constructed our prior bandwidth on second and third derivatives that result from a simple normal distribution.

The mixture of CV and plug-in is a compromise giving biases lying between the ISE and the MISE minimizing methods. It is interesting to see that this yields such a stable performance. We should mention that there were only minor differences between the three versions of mixtures (not shown here). Clearly, the larger the share of the respective method, the bigger their impact on the final estimate.

Comparison of the ISE values

Next, we compare the ISE values of the density estimates based on the different bandwidth selectors. The results are given in terms of boxplots plus the mean (linked squares) displaying this way the distribution of the ISEs over 250 simulation runs. In Fig. 3, we consider the mixture of three normal distributions (model 6) and compare different sample sizes, whereas in Fig. 4 the sample size is fixed to $n = 100$ while the true underlying distribution varies.

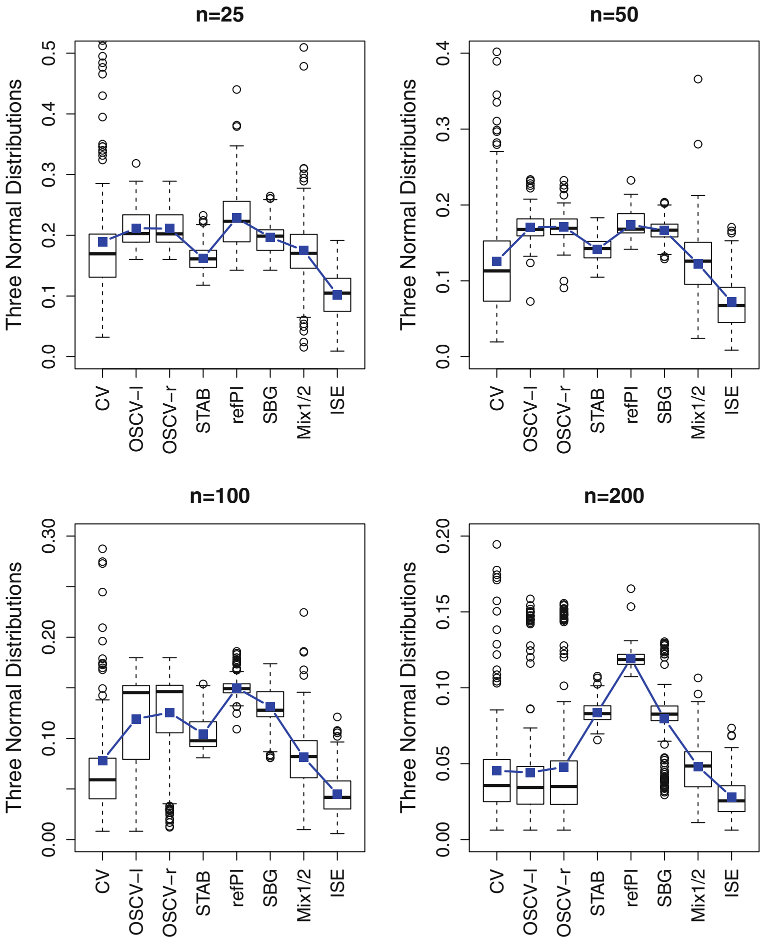

Fig. 3 Box-plots and means (*filled square*) of the ISE values for the mixture of three normal densities with different sample sizes

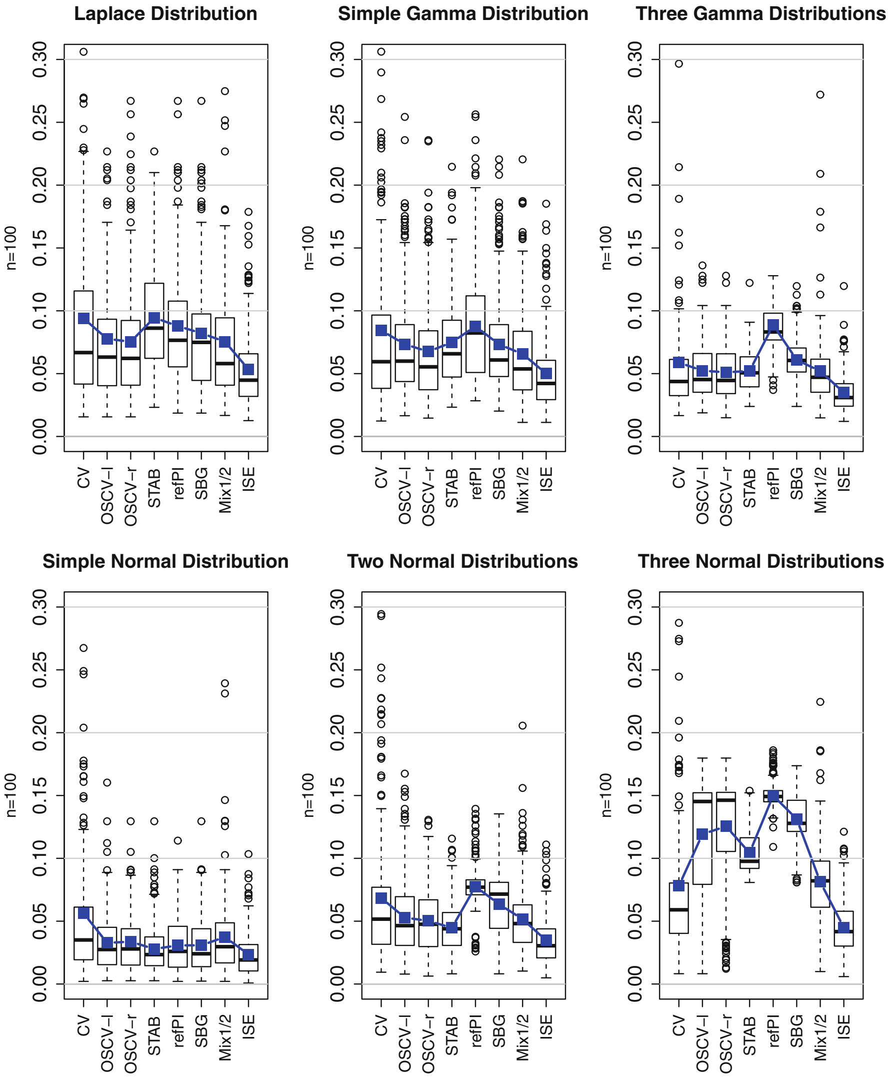

Fig. 4 Box-plots and means (*filled square*) of the ISE values for different distributions with sample size $n = 100$

Certainly, for all methods the ISE values decrease with the sample size and increase with the complexity of the estimation problem. As expected, the classical CV shows a high variation for all cases (upper extreme values are not shown for the sake of presentation). The one-sided CV and the STAB versions do much better, while the least variation is achieved by the MISE minimizing methods (STAB, refPI and SBG). However, the drawback of these three methods becomes obvious when looking at the size of its ISE values; they are clearly smaller for the CV-based methods when $n \geq 25$. Moreover, for increasing sample size the ISE values decrease very slowly for the MISE-based methods, whereas for the CV methods they come close to the smallest achievable ISE. Note that regarding the ISE minimization, the one-sided CV methods show the best performance, except for the triple mode normal mixture. They do not vary as much as the classical CV selector, and their mean value is almost always smaller than for the other methods, see Fig. 4.

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

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

To get an even better idea of the distance between the ISE values achieved by the selectors and the ISE optimal (i.e. achievable) values, let us have a closer look at c_5 and c_6, say the L_1 and L_2 distances. In our opinion, these measures are probably the most interesting ones for practitioners. Figures 5 and 6 show these L1- and the L2-distances for the different sample sizes and models.

The pictures reveal that for CV bandwidths, the c_5 are really big, even if the underlying density is not wiggly at all. This obviously is due to the high variability of the selected bandwidths. This effect does especially apply for small sample sizes; but notice that for large samples like $n = 500$ the classical CV still does not work well (not shown). Regarding the L_1 measure (c_5), we see that the CV delivers often pretty small values for samples of size $n > 50$.

While both OSCV methods have problems with particularly small sample sizes, they else easily compete with all the other selectors. One may say that for the normal

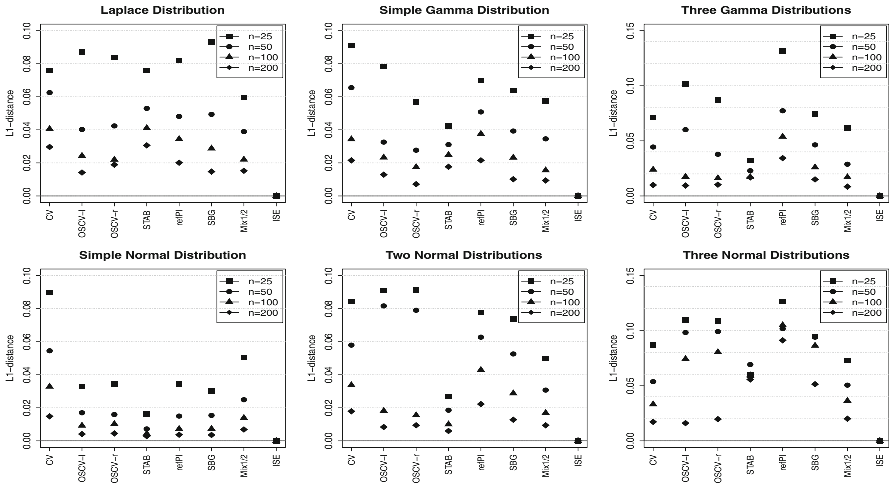

Fig. 5 L1-distances to ISE(h_{opt}) for different sample sizes of the six underlying densities

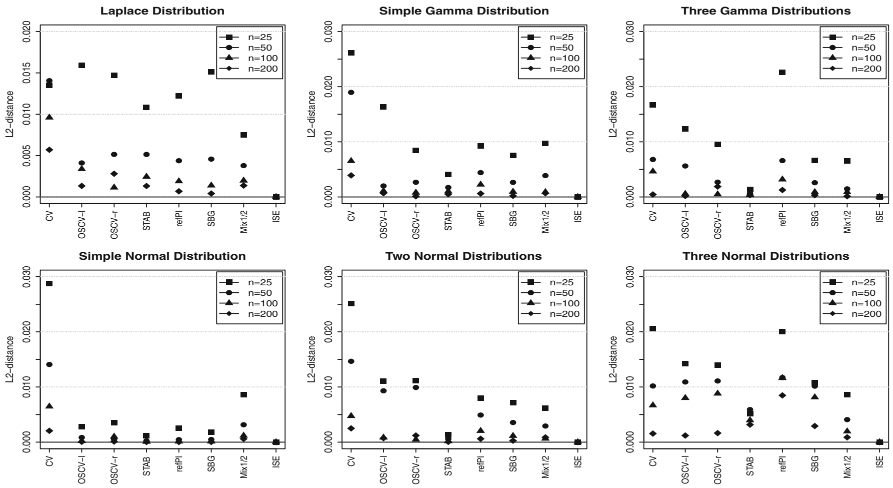

Fig. 6 L2-distances to ISE(h_{opt}) for different sample sizes of the six underlying densities

densities the OSCV methods are neither the best nor the worst methods, but always within the grasp of the best method. This corroborates our statement from above that an OSCV selector should be used if we do not know anything about the underlying density. Another conspicuous finding in Fig. 5 is the difference between the two one-sided versions for the gamma distribution(s). Because of missing boundary correction on the left, the OSCV-l behaves very badly especially for small sample sizes. We actually get similar results for $n = 25$ when looking at the L2-distances, see below.

The three MISE minimizing methods do very well for the simple normal and simple gamma distribution, but else we observe a worse performance which can be traced back to the prior selection problem already described above. Even for bigger sample sizes, all three methods deliver a relative big L1-distance for the mixture models. They do further not benefit as much from an increasing n as other methods do. Within this MISE minimizing group, the STAB shows a smaller L1-distance for more complex densities. Actually, for the mixture of the three Gamma distributions, we can see that his L1-distances to the optimal ISE are always very small, except for the refPI and SBG with $n = 25$.

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

We obtain not exactly the same but similar results when looking at the L2-distance to the optimal ISE, plotted in Fig. 6.CV obtains very large values for small sample sizes. The one-sided versions show an important improvement. The three MISE minimizing methods perform excellent for the simple normal (not surprisingly) and the simple gamma. Among them, the STAB shows the smallest L2 distance to ISE(h_{opt}). For sample sizes $n > 50$ the one-sided CV versions outperform the others in most cases. Large differences between the left and the right one-sided version can be observed where we have asymmetric densities.

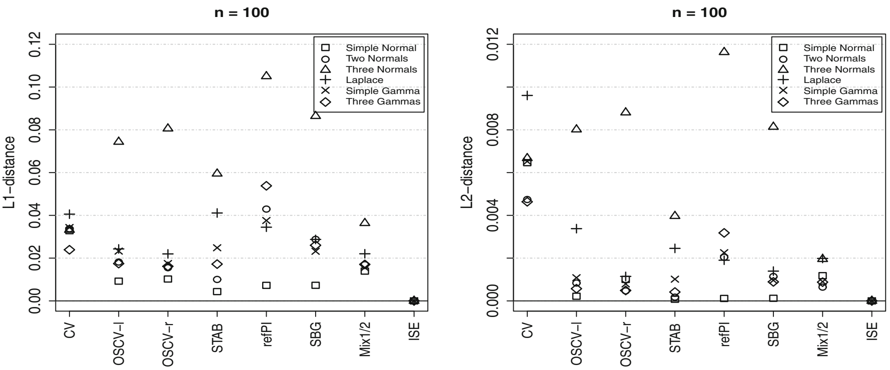

Fig. 7 L1- and L2-distances to ISEh_{opt} for different underlying densities when $n = 100$

A comparison of the L1- and the L2-distance for a sample size fixed at $n = 100$ but varying the distributions is shown in Fig. 7. As can be seen in both pictures, the performance of all measures (without CV) for the simple normal distribution and the mixture of the three gamma distributions is pretty good. Also for the mixture of two normals, most of the methods deliver good results; only the values for CV, refPI and the SBG become much larger. For more complex densities, the pictures show that the MISE minimizing measures deliver worse results, because of the large biases. STAB shows a pretty good performance with respect to the L_1 measure what is not surprising when recalling its construction. The most stable versions are the OSCV and the Mix, except for the triple mode normal mixture. For smaller sample sizes (not shown), the pictures are quite similar, but the tendencies are strengthened and only the Mix version delivers stable results for all distributions.

Comparison of the mixed methods

Finally, we have a closer look to the quite promising results obtained by mixing CV with refPI. We have done this mingling using the different proportions described above. In Tables 1 and 2 we have tabulated the different performance measures looking at the bias of the chosen bandwidth, the average ISE as well as the L1- and L2-distances to the optimal ISE for all of the six densities.

For all smooth densities, we observe that the values of the different measures are pretty close to each other. The main differences occur for small sample sizes and wiggly densities. It is hard to say which mixture is the best as sometimes mix2 is the best and sometimes mix1 whereas mix3 lies certainly always in between. The reason seems to be obvious, either refPI is more appropriate than CV or vice verse. But this is a conclusion one may draw from the means while at the same time they reduce a lot the variance. We, therefore, see the potential gain of the methods is best when looking at the L_2 distances. Recall also our results from the last subsections and compare mix3 with CV and refPI looking at that measure c_4. We should also now give a special emphasis on this performance measure c_4.

Table 1 Laplace, simple gamma, and mixture of three gamma distributions

n	Crit.	Design 1			Design 2			Design 3		
		mix1	mix2	mix3	mix1	mix2	mix3	mix1	mix2	mix3
25	c_1	.0620	.0454	.0483	.0535	.0264	.0331	.2460	.1354	.1756
	c_3	.1180	.1185	.1133	.0999	.1218	.1076	.0503	.0746	.0596
	c_4	.0085	.0094	.0075	.0074	.0148	.0097	.0076	.0078	.0066
	c_5	.0647	.0659	.0597	.0545	.0687	.0574	.0728	.0601	.0618
50	c_1	.0404	.0241	.0268	.0431	.0140	.0225	.1388	.0741	.0954
	c_3	.0810	.0917	.0812	.0634	.0997	.0759	.0297	.0367	.0298
	c_4	.0031	.0065	.0038	.0024	.0086	.0039	.0021	.0020	.0015
	c_5	.0391	.0467	.0389	.0340	.0455	.0346	.0380	.0305	.0288
100	c_1	.0366	.0197	.0227	.0382	.0088	.0167	.0988	.0414	.0615
	c_3	.0460	.0756	.0536	.0380	.0560	.0409	.0203	.0374	.0277
	c_4	.0013	.0050	.0020	7e−04	.0026	9e−04	9e−04	.0015	9e−04
	c_5	.0246	.0288	.0220	.0184	.0213	.0155	.0244	.0175	.0171
200	c_1	.0245	.0083	.0117	.0330	.0097	.0152	.0782	.0326	.0480
	c_3	.0292	.0533	.0385	.0223	.0400	.0295	.0107	.0123	.0105
	c_4	7e−04	.0030	.0014	3e−04	.0015	7e−04	3e−04	1e−04	1e−04
	c_5	.0154	.0203	.0152	.0109	.0135	.0094	.0144	.0080	.0084

Table 2 Simple normal distribution, mixture of two and three normal distributions

n	Crit.	Design 4			Design 5			Design 6		
		mix1	mix2	mix3	mix1	mix2	mix3	mix1	mix2	mix3
25	c_1	.0290	.0030	.0073	.0818	.0447	.0561	.1979	.1338	.1545
	c_3	.0749	.1126	.0926	.0567	.1011	.0740	.0475	.0739	.0571
	c_4	.0045	.0143	.0086	.0043	.0119	.0061	.0096	.0104	.0086
	c_5	.0405	.0640	.0504	.0504	.0602	.0498	.0819	.0731	.0729
50	c_1	.0281	.0097	.0120	.0663	.0337	.0440	.1216	.0670	.0876
	c_3	.0435	.0799	.0581	.0396	.0695	.0510	.0342	.0526	.0420
	c_4	.0014	.0068	.0032	.0023	.0059	.0029	.0057	.0043	.0041
	c_5	.0190	.0354	.0249	.0343	.039	.0307	.0657	.0445	.0506
100	c_1	.0151	−.0074	−.0042	.0571	.0259	.0344	.0897	.0366	.0577
	c_3	.0248	.0490	.0358	.0234	.0379	.0273	.026	.0456	.0302
	c_4	4e−04	.0025	.0012	8e−04	.0015	6e−04	.004	.0023	.0020
	c_5	.0096	.0203	.0139	.0218	.0225	.0169	.0572	.0276	.0364
200	c_1	.0134	−.0022	−.0021	.0382	.0153	.0200	.0684	.0230	.0393
	c_3	.0314	.0258	.0134	.0223	.0354	.0305	.0225	.0285	.0235
	c_4	1e−04	9e−04	6e−04	5e−04	.0012	9e−04	.0021	9e−04	9e−04
	c_5	.0047	.0093	.0069	.0119	.0122	.0094	.0406	.0154	.0201

We first see that over the different sample sizes the considered performance measures converge toward zero as expected for increasing sample size. Notice, however, that depending on the smoothness of the underlying density they do so at seemingly different rates. A similar observation can be made if comparing the development over an decrease of smoothness: compare, for example, design 4–6 (from a simple to a triple mode normal mixture). Since we could not identify a clear winner between refPI and CV, it may be not surprising that the best compromise seems indeed to be mix3. In total, the main conclusion is that the considered bandwidth mixtures produce very stable results and are attractive competitor to the other bandwidth selection methods.

7 Conclusions

A first finding is that it definitely makes a difference which bandwidth selector is chosen; not only in numerical terms but also for the quality of density estimation. We can identify clear differences in quality, and we can say in which situation what kind of selector is preferable.As is well known, the CV leads to a small bias but a large variance. It works well for rather wiggly densities and a moderate sample size. However, it neither performs well for rather small nor for rather large samples. The quality is unfortunately dominated by its variability. An also fully automatic alternative is the one-sided version. In contrast to the classical CV, the OSCV methods exhibit much less variation without increasing too much in bias. For very small samples these methods have their numerical problems, what is caused by their construction. They may be not uniformly but quite often the best, and never the worst. Depending on the skewness, either the left- or the right-sided CV performs better. This disadvantage is no longer present for the alternative Do-validation or the indirect SHS bandwidth selector. Unfortunately, for a reasonable working of the SHS selector, a sample size of $n > 100$ is strongly recommended. Further, it also depends on two prior parameters for which some recommendations exist for $n > 100$. Note that also all the following statements are conditioned on our prior choices, and may be just the selection of densities and sample sizes. We are aware that for large samples and quite wiggly densities our findings and conclusions might change.

The refPI and the SBG show a similarly stable behavior due to the fact that they are minimizing the MISE, and depend on prior information. It is generally accepted that the need of prior knowledge is the main disadvantage of these methods, and—as explained above—typically requires a smooth underlying density. We have to admit that larger samples would allow for more complex plug-in methods but these often require more prior knowledge.

The STAB method is quite stable as suggested by its name. Although the full name refers to cross-validation, it actually minimizes the MISE like refPI and SBG do. Consequently, it performs particularly well for the estimation of rather smooth densities but else does not. It certainly pays for the stability with some bias increase. It is, therefore, hard to say to what extend it is an improvement to CV, but it seems to be an improvement compared to refPI when looking at the ISE of the density estimator.

While the mix methods (combining CV and plug-in) do very well, one cannot really identify a 'best mix' in advance. A further evident computational disadvantage is that we first have to apply two other methods (CV and refPI) to achieve good results. Therefore, we have studied here only the combination of the simplest plug-in with the simplest CV method. It would be little surprising if better results were obtained when mixing more sophisticated methods, see for example Mammen et al. (2011).

Our conclusion is, therefore, that among all existing (automatic) methods for kernel density estimation, concentrating on small or moderate samples and relatively smooth densities, the best strategies seem to be either a mixing or an indirect method. Among the two competing indirect methods (OSCV and SHS) the two OSCV seem to outperform SHS. However, if sample sizes increase a lot, and skewness becomes an important issue, then SHS is doubtless an interesting alternative for the reasons discussed. Depending on the boundary, one would apply left- or right-sided OSCV, respectively. For moderate sample sizes however, the mixture of CV and refPI seems to be an attractive alternative until n becomes that large that CV fails completely.

References

Ahmad, I.A., Ran, I.S.: Data based bandwidth selection in kernel density estimation with paramteric start via kernel contrasts. J. Nonparametr. Stat. **16**, 841–877 (2004)

Bean, S.J., Tsokos, C.P.: Developments in nonparametric density estimation. Int. Stat. Rev. **48**, 267–287 (1980)

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

Cao, R.: Bootstrapping the mean integrated squared error. J. Multivar. Anal. **45**, 137–160 (1993)

Cao, R., Cuevas, A., Gonzlez Manteiga, W.: A comparative study of several smoothing methods in density estimation. Comput. Stat. Data Anal. **17**, 153–176 (1994)

Chacon, J.E., Montanero, J., Nogales, A.G.: Bootstrap bandwidth selection using an h-dependent pilot bandwidth. Scand. J. Stat. **35**, 139–157 (2008)

Chaudhuri, P., Marron, J.S.: SiZer for exploration of structures in curves. J. Am. Stat. Assoc. **94**, 807–823 (1999)

Chiu, S.T.: Some stabilized bandwidth selectors for nonparametric regression. Ann. Stat. **19**, 1528–1546 (1991a)

Chiu, S.T.: Bandwidth selection for kernel density estimation. Ann. Stat. **19**, 1883–1905 (1991b)

Chiu, S.T.: An automatic bandwidth selector for kernel density estimation. Biometrika **79**, 771–782 (1992)

Chiu, S.T.: A comparative review of bandwidth selection for kernel density estimation. Stat. Sin. **6**, 129–145 (1996)

Devroye, L.: The double kernel method in density estimation. Annales de l'Institut Henri Poincaré **25**, 533–580 (1989)

Devroye, L.: Universal smoothing factor selection in density estimation: theory and practice. Test **6**, 223–320 (1997)

Devroye, L., Gyorfi, L.: Nonparametric Density Estimation: The L_1 View. Wiley, New York (1985)

Devroye, L., Lugosi, G.: A universal acceptable smoothing factor for kernel density estimation. Ann. Stat. **24**, 2499–2512 (1996)

Duin, R.P.W.: On the choice of smoothing parameters of Parzen estimators of probability density functions. IEEE Trans. Comput. **25**, 1175–1179 (1976)

Faraway, J.J., Jhun, M.: Bootstrap choice of bandwidth for density estimation. J. Am. Stat. Assoc. **85**, 1119–1122 (1990)

Feluch, W., Koronacki, J.: A note on modified cross-validation in density estimation. Comput. Stat. Data Anal. **13**, 143–151 (1992)

Fryer, M.J.: A review of some non-parametric methods of density estimation. J. Appl. Math. **20**(3), 335–354 (1977)

Godtliebsen, F., Marron, J.S., Chaudhuri, P.: Significance in scale space for bivariate density estimation. J. Comput. Graph. Stat. **11**, 1–21 (2002)

Grund, B., Polzehl, J.: Bias corrected bootstrap bandwidth selection. J. Nonparametr. Stat. **8**, 97–126 (1997)

Habbema, J.D.F., Hermans, J., van den Broek, K.: A stepwise discrimination analysis program using density estimation, In: Bruckman, G. (Ed.) COMPSTAT '74. Proceedings in Computational Statistics, pp. 101–110. Physica, Vienna (1974)

Hall, P.: Using the bootstrap to estimate mean square error and select smoothing parameters in nonparametric problems. J. Multivar. Anal. **32**, 177–203 (1990)

Hall, P., Johnstone, I.: Empirical functionals and efficient smoothing parameter selection. J. R. Stat. Soc. Ser. B **54**, 475–530 (1992)

Hall, P., Marron, J.S.: Extent to which least-squares cross-validation minimises integrated square error in nonparametric density estimation. Probab. Theory Relat. Fields **74**, 567–581 (1987a)

Hall, P., Marron, J.S.: Estimation of integrated squared density derivatives. Stat. Probab. Lett. **6**, 109–115 (1987b)

Hall, P., Marron, J.S.: Lower bounds for bandwidth selection in density estimation. Probab. Theory Relat. Fields **90**, 149–173 (1991)

Hall, P., Marron, J.S., Park, B.U.: Smoothed cross-validation. Probab. Theory Relat. Fields **92**, 1–20 (1992)

Hall, P., Sheater, S.J., Jones, M.C., Marron, J.S.: On optimal databased bandwidth selection in kernel density estimation. Biometrika **78**, 263–269 (1991)

Hanning, J., Marron, J.S.: Advanced distribution theory for SiZer. J. Am. Stat. Assoc **101**, 484–499 (2006)

Hardle, W., Muller, M., Sperlich, S., Werwatz, A.: Nonparametric and Semiparametric Models. Springer Series in Statistics, Berlin (2004)

Hardle, W., Vieu, P.: Kernel regression smoothing of time series. J. Time Ser. Anal. **13**, 209–232 (1992)

Hart, J.D., Yi, S.: One-sided cross validation. J. Am. Stat. Assoc. **93**, 620–631 (1998)

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

Jones, M.C., Marron, J.S., Park, B.U.: A simple root *n* bandwidth selector. Ann. Stat. **19**, 1919–1932 (1991)

Jones, M.C., Marron, J.S., Sheather, S.J.: A brief survey of bandwidth selection for density estimation. J. Am. Stat. Assoc. **91**, 401–407 (1996a)

Jones, M.C., Marron, J.S., Sheather, S.J.: Progress in data-based bandwidth selection for kernel density estimation. Comput. Stat. **11**, 337–381 (1996b)

Jones, M.C., Sheather, S.J.: Using non-stochastic terms to advantage in kernel-based estimation of integrated squared density derivatives. Stat. Probab. Lett. **11**, 511–514 (1991)

Kim, W.C., Park, B.U., Marron, J.S.: Asymptotically best bandwidth selectors in kernel density estimation. Stat. Probab. Lett. **19**, 119–127 (1994)

Loader, C.R.: Bandwidth selection: classical or plug-in? Ann. Stat. **27**(2), 415–438 (1999)

Mammen, E., Martínez-Miranda, M.D., Nielsen, J.P., Sperlich, S.: Do-validation for kernel density estimation. J. Am. Stat. Assoc. **106**, 651–660 (2011)

Marron, J.S.: Convergence properties of an empirical error criterion for multivariate density estimation. J. Multivar. Anal. **19**, 1–13 (1986)

Marron, J.S.: Automatic smoothing parameter selection: a survey. Empir. Econ. **13**, 187–208 (1988a)

Marron, J.S.: Partitioned cross-validation. Econ. Rev. **6**, 271–283 (1988b)

Marron, J.S.: Bootstrap bandwidth selection. In: LePage, R., Billard, L. (eds.) Exploring the Limits of Bootstrap, pp. 249–262. Wiley, New York (1992)

Marron, J.S.: Visual understanding of higher order kernels. J. Comput. Graph. Stat. **3**, 447–458 (1994)

Marron, J.S., Nolan, D.: Canonical kernels for density estimation. Stat. Probab. Lett. **7**, 195–199 (1988)

Marron, J.S., Wand, M.P.: Exact mean integrated squared errors. Ann. Stat. **20**, 712–736 (1992)

Martinez-Miranda, M.D., Nielsen, J., Sperlich, S.: One sided cross validation in density estimation. In: Gregoriou, G.N. (ed.) Operational Risk Towards Basel III: Best Practices and Issues in Modeling, Management and Regulation. Wiley, Hoboken (2009)

Park, B.U., Marron, J.S.: Comparison of data-driven bandwidth selectors. J. Am. Stat. Assoc. **85**, 66–72 (1990)

Park, B.U., Turlach, B.A.: Practical performance of several data driven bandwidth selectors, CORE Discussion Paper 9205 (1992)

Rigollet, P., Tsybakov, A.: Linear and convex aggregation of density estimators. Math. Methods Stat. **16**, 260–280 (2007)

Rudemo, M.: Empirical choice of histograms and kernel density estimators. Scand. J. Stat. **9**, 65–78 (1982)

Ruppert, D., Cline, B.H.: Bias Reduction in kernel density estimation by smoothed empirical transformations. Ann. Stat. **22**, 185–210 (1994)

Samarov, A., Tsybakov, A.: Aggregation of density estimators and dimension reduction. In: Nair, V. (ed.) Advances in Statistical Modeling and Inference: essays in honor of Kjell A. Doksum, pp. 233–251 (2007)

Savchuk, O.J., Hart, J.D., Sheather, S.J.: Indirect cross-validation for density estimation. J. Am. Stat. Assoc. **105**, 415–423 (2010)

Silverman, B.W.: Density estimation for statistics and data analysis. Monographs on Statistics and Applied Probability, vol. 26. Chapman and Hall, London (1986)

Scott, D.W., Terrell, G.R.: Biased and unbiased cross-validation in density estimation. J. Am. Stat. Assoc. **82**, 1131–1146 (1987)

Sheather, S.J.: Density estimation. Stat. Sci. **19**, 588–597 (2004)

Sheather, S.J., Jones, M.C.: A reliable data-based bandwidth selection method for kernel density estimation. J. R. Stat. Soc. Ser. B **53**, 683–690 (1991)

Stone, C.J.: An asymptotically optimal window selection rule for kernel density estimates. Ann. Stat. **12**, 1285–1297 (1984)

Stute, W.: Modified cross validation in density estimation. J. Stat. Plan. Inference **30**, 293–305 (1992)

Tartar, M.E., Kronmal, R.A.: An introduction to the implementation and theory of nonparametric density estimation. Am. Stat. **30**, 105–112 (1976)

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

Turlach, B.A.: Bandwidth selection in kernel density estimation: a review. Working Paper (1994)

Wand, M.P., Jones, M.C.: Kernel smoothing. Monographs on Statistics and Applied Probability, vol. 60. Chapman and Hall, London (1995)

Wand, M.P., Marron, J.S., Ruppert, D.: Transformations in density estimation. J. Am. Stat. Assoc. **86**, 343–353 (1991)

Wegkamp, M.H.: Quasi universal bandwidth selection for kernel density estimators. Can. J. Stat. **27**, 409–420 (1999)

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

Wertz, W., Schneider, B.: Statistical density estimation: a bibliography. Int. Stat. Rev. **47**, 155–175 (1979)

Yang, Y.: Mixing strategies for density estimation. Ann. Stat. **28**, 75–87 (2000)

Yang, L., Marron, S.: Iterated transformation-kernel density estimation. J. Am. Stat. Assoc. **94**, 580–589 (1999)