Probab. Theory Relat. Fields 92, 1–20 (1992)

Probability Theory and Related Fields

Smoothed cross-validation

Peter Hall, J.S. Marron*, and Byeong U. Park*

Australian National University, GPO Box 4, Canberra ACT 2601, Australia
and Seoul National University, Korea

Received May 10, 1990; in revised form March 27, 1991

Summary. For bandwidth selection of a kernel density estimator, a generalization of the widely studied least squares cross-validation method is considered. The essential idea is to do a particular type of "presmoothing" of the data. This is seen to be essentially the same as using the smoothed bootstrap estimate of the mean integrated squared error. Analysis reveals that a rather large amount of presmoothing yields excellent asymptotic performance. The rate of convergence to the optimum is known to be best possible under a wide range of smoothness conditions. The method is more appealing than other selectors with this property, because its motivation is not heavily dependent on precise asymptotic analysis, and because its form is simple and intuitive. Theory is also given for choice of the amount of presmoothing, and this is used to derive a data-based method for this choice.

1 Introduction

The problem of bandwidth or smoothing parameter selection is most important for effective data analysis using any type of smoothing method. Good discussion of this point may be found in the monographs Silverman (1986), Eubank (1988), Müller (1988) and Härdle (1989). In the present paper, this problem is considered in the case of kernel density estimation, but the essential ideas are also applicable to a wide variety of other estimators and settings as well. Our ideas are developed and presented in this context, because its simplicity allows straightforward understanding of the main issues.

The centerpoint of research on bandwidth selection for the kernel density estimator has been least squares cross-validation, CV. The main idea behind this is a natural approximation to the integrated squared error, as proposed by Rudemo

* Research of the second author was done while on leave from the University of North Carolina. That of both the second and third was partially supported by National Science Foundation Grants DMS-8701201 and DMS-8902973

(1982) and Bowman (1984). The greatest appeal of the method is the compelling intuition used in its motivation. The first theoretical results, such as those of Hall (1983), Burman (1985) and the especially deep work of Stone (1984), were quite promising in showing that the performance of the resulting estimator is asymptotically as good as the best possible. However, the performance of this method in applications and simulations has been fairly disappointing. The problem is that the least squares cross-validated bandwidth is subject to too much sample variability, being too large for some data sets and too small for others, i.e. it is too noisy. This phenomenon has been quantified in Hall and Marron (1987) and Scott and Terrell (1987), where it is shown that the relative rate of convergence to the optimum is of the excruciatingly slow order of $n^{-1/10}$.

Because of the unacceptably large dependence of CV on sampling fluctuations, there has recently been a search for more stable methods. The first work in this important direction was the proposal of biased cross validation, which can be viewed as an attempt to cut down on the variability of CV, at the price of introducing some small bias, as proposed by Scott and Terrell (1987). Other work centers on various modifications of the "plug-in" idea, which has an early history, with versions dating back to Woodroofe (1970), Scott et al. (1977) and Scott and Factor (1981). This involves plugging estimates into an asymptotic representation of the optimal bandwidth, but the actual implementation has proved to be rather tricky. The essential idea behind the currently most-studied version involves solving a fixed point equation, and was developed in Hall (1981) and Sheather (1986). Very promising theoretical, simulation, and practical performance of this type of bandwidth selector was demonstrated by Park and Marron (1989) in the context of nonnegative kernels, and by Hall et al. (1989) for higher order kernels. Hall and Marron (1989) have shown that this type of bandwidth selector achieves the best possible relative rate of bandwidth convergence, under a very wide variety of smoothness conditions.

Despite these promising properties of various versions of the plug-in idea, they have two serious drawbacks. The first is that their motivation relies heavily on asymptotic arguments, and lacks the compelling simple intuition of CV. This point makes these methods difficult to sell to those data analysts who lack a mathematical background, and who thus tend to be uncomfortable (perhaps rightly so) with methods which place too much dependence on asymptotic ideas. The second drawback is that the methods are quite complicated, involving unappealing seventh, ninth and even thirteenth roots of various crucial quantities, so that the algorithms themselves are rather unintuitive, and again unconvincing in their own right.

The main point of this paper is the proposal of a data-based bandwidth selector which has the same desirable properties as those described for the plug-in method, but also has a simple, intuitive nonasymptotic motivation. We consider this method very appealing because it can be motivated in three separate and natural ways, described in detail in Sect. 2.

The first of these is a modification of CV that involves a presmoothing of the pairwise differences of the observations, before they are plugged into the usual cross-validation criterion. This motivates the name Smoothed Cross Validation. SCV includes ordinary CV as the special case of no presmoothing. Insight into why this type of presmoothing is so beneficial, in particular cutting down on the sample variability inherent to CV, is provided in Sect. 2.1 This comes through consideration of the family of SCV criteria from a viewpoint common in the statistical analysis of point patterns.

Section 2.2 gives a second motivation of SCV, through substituting a pilot kernel estimator into the integrated squared bias. This approach provides an estimate of the integrated squared bias which may be viewed as more direct than the usual asymptotic bias approximations, for example the type used in Biased Cross Validation and by the usual plug-in methods.

Another compelling feature of SCV, discussed in Sect. 2.3, is that it is essentially the same as the smoothed bootstrap estimate of the mean integrated squared error, which has been proposed by Faraway and Jhun (1990) and Taylor (1989). This connection yields benefits in two directions. First it provides yet another sense in which SCV is a very natural bandwidth selector, especially to those accustomed to thinking in bootstrap terms. Second it enables our analysis to automatically provide much deeper analysis than has been previously done concerning the choice of how much presmoothing should be initially done for the smoothed bootstrap.

Asymptotic analyses of SCV, including rates of convergence and theoretically optimal choice of the amount of presmoothing, are given in our main theorems, which are stated in Sect. 3. Data-based choice of the amount of presmoothing, together with a comparison with plug-in methods, is discussed in Sect. 4. Proofs of the main theorems are given in Sect. 5.

2. Definitions and ideas

A mathematical formulation of the density estimation problem involves using a sample $X_1, \ldots, X_n$ from a probability density f to estimate f. The kernel estimator is defined by

$$\hat{f}_h(x) = n^{-1} \sum_{i=1}^{n} K_h(x - X_i) ,$$

where $K_h(\cdot) = K(\cdot/h)/h$, K is called the kernel function, and h is called the bandwidth. Choice of h is crucial to the performance of $\hat{f}_h$: when h is too small the resulting curve is too wiggly, and when it is too large important features will be smoothed away. See for example Silverman (1986) or Devroye and Györfi (1985) for motivation, early references, and discussion of various aspects.

In this paper, the Smoothed Cross-Validation, SCV, method of using the data to choose the bandwidth is proposed and studied. There are three separate motivations for this method, which will be presented in turn.

2.1 Cross-Validation motivation of SCV

The method of least squares cross-validation, is to take the minimizer, $\hat{h}(0)$ say, of the function

$$\mathrm{CV}(h) = \int \hat{f}_h(x)^2 \, dx - 2n^{-1} \sum_{j=1}^{n} \hat{f}_{h,j}(X_j) ,$$

which is an estimate of a vertical shift of the integrated squared error, considered as a function of h. Here and below $\int$ with no explicit limits is understood to denote

integration over the real line, and $\hat{f}_{h,j}$ denotes the leave-one-out estimator defined by

$$\hat{f}_{h,j}(x) = (n-1)^{-1} \sum_{i \neq j}^{n} K_h(x - X_i) .$$

Scott and Terrell (1987) pointed out that $\mathrm{CV}(h)$ admits the representation (modulo the very accurate approximation $n \simeq n-1$)

$$\mathrm{CV}(h) = (nh)^{-1} R(K) + n^{-1}(n-1)^{-1} \sum\sum_{i \neq j} (K_h * K_h - 2K_h)(X_i - X_j) , \tag{2.1}$$

where $*$ denotes convolution, and where the functional $R(K) \equiv \int K^2$ (the same notation will be used for other functions as well as K). This representation is quite suggestive because the first term is a very good approximation to the integrated variance part of Mean Integrated Squared Error,

$$\mathrm{MISE} = E \int (\hat{f}_h - f)^2 ,$$

see for example (3.20) of Silverman (1986), while the second part can be viewed as an estimate of integrated squared bias,

$$B(h) = \int (K_h * f - f)^2 .$$

To see more clearly how the second term in (2.1) estimates bias, note that when there are no duplications among the data (which happens with probability one for truly continuous data), it can be written as

$$n^{-1}(n-1)^{-1} \sum\sum_{i \neq j} (K_h * K_h - 2K_h + K_0)(X_i - X_j) ,$$

where here and below K_0 denotes the Dirac delta function. This reveals that bias is essentially being estimated by a second differencing operation (this is made more precise later). It follows from the results of Hall and Marron (1987b), which provide pertinent limiting distributions for $\hat{h}(0)$ based on such quantities, that estimates of integrated squared bias tend to suffer from high variance when bandwidths of the order $n^{-1/5}$ are considered. Since this is the only random part of $\mathrm{CV}(h)$ it follows that this variability is the cause of the unacceptably large noise in $\hat{h}(0)$.

The smoothed cross-validation criterion addresses this problem by modifying this term, by a type of presmoothing of the differences $X_i - X_j$ which still provides a bias estimate, and yet has better stability properties. In particular, define $\hat{h}(g)$ to be the minimizer of

$$\mathrm{SCV}_g(h) = (nh)^{-1} R(K) + \hat{B}_g(h) , \tag{2.2}$$

where

$$\hat{B}_g(h) = n^{-1}(n-1)^{-1} \sum\sum_{i \neq j} \{(K_h * K_h - 2K_h + K_0) * L_g * L_g\}(X_i - X_j) , \tag{2.3}$$

for the possibly different kernel function L and bandwidth g.

2.2 Pilot estimation motivation of SCV

Recall from the last section that the integrated squared bias, $B(h)$, has a simple representation in terms of the L^2 distance between f and $K_h * f$. In the closely

related context of nonparametric regression, Müller (1985) and Staniswallis (1989) have proposed the natural idea of replacing f in $B(h)$ with a pilot kernel estimator $\hat{f}_g$, which here has the possibly different bandwidth g and kernel L. Note that using some algebra,

$$(2.4)\quad \int (K_h * \hat{f}_g - \hat{f}_g)^2 = \int (K_h * \hat{f}_g - \hat{f}_g)^2 - h^{-1}(K_h * K_h - 2K_h + K_0) * L_g * L_g(0)$$
$$= \hat{B}_g(h)\,.$$

This means that the bias part of the function $\mathrm{SCV}_g(h)$ can be thought of as this very natural estimate $B(h)$. Since bias is the hardest part of the MISE to estimate, it follows that $\mathrm{SCV}_g(h)$ is providing a natural estimate of the curve $\mathrm{MISE}(h)$. This is why the form $L_g * L_g$ was used in (2.3), while we could just as well have used simply L_g at that point. See Chiu (1990) for bandwidth selection ideas related to this, but based on Fourier Transforms.

2.3 *Smoothed bootstrap motivation of* SCV

For a third motivation of SCV, let $X_1^*, \ldots, X_n^*$ denote the so called "smoothed bootstrap" random sample, i.e. a random sample chosen to have conditional (given $X_1, \ldots, X_n$) distribution with density $\hat{f}_g(x)$. Let E^* denote conditional expected value with respect to this distribution, and define

$$\hat{f}_h^*(x) = n^{-1} \sum_{i=1}^{n} K_h(x - X_i^*)\,.$$

Simple calculations show that

$$E^* \hat{f}_h^*(x) = K_h * \hat{f}_g(x)\,,$$

and so

$$\mathrm{MISE}^*(h) = E^* \int (\hat{f}_h^* - f\}^2 = (nh)^{-1} R(K) + n^{-1} \int (K_h * \hat{f}_g)^2 + \int (K_h * \hat{f}_g - \hat{f}_g)^2$$
$$= \mathrm{SCV}(h) + o((nh)^{-1})\,.$$

In this sense, SCV is just the smoothed bootstrap estimate of MISE. Note that unlike most nontrivial applications of the bootstrap, no simulation needs to be done to obtain the desired functional of the bootstrap distribution. This appears to be because MISE* depends on $\hat{f}_g$ in such a simple way.

2.4 *Further discussion*

At this point it is natural to consider choice of g. Our first guess was that the representation (2.4) seems to indicate that g should be fairly small, or perhaps that $g = h$ would be reasonable. However the theorems of the next section indicate that in fact g should be surprisingly large, and the choice $g = h$ gives a relative rate of convergence no better than for $g = 0$ (the case of ordinary CV), although the constant coefficient is better.

To understand why this should be the case, it is useful to consider the relationship between cross-validation and the function

$$\hat{K}(t) = n^{-1}(n-1)^{-1} \sum\sum_{i \neq j} K_t(X_i - X_j) .$$

This is essentially (exactly so if K is the uniform kernel) the cumulative distribution function of the population of pairwise differences, that has been used very successfully for analysis of spatial point patterns, see the monographs Ripley (1981) and Diggle (1983) on this topic. The connection between $\hat{K}$ and CV was pointed out by Diggle and Marron (1988), who use $\hat{K}$ for a different motivation of CV. To simplify the notation here, it is convenient to work with a version of $\hat{K}$ where the bandwidth argument is transformed, in particular define $\hat{\tilde{K}}(s) = \hat{K}(s^{1/2})$. The connection between this function and SCV is easiest to see when $K = L = \varphi$, the standard Gaussian density. In this case,

$$\hat{B}_g(h) = \hat{\tilde{K}}(2h^2 + 2g^2) - 2\hat{\tilde{K}}(h^2 + 2g^2) + \hat{\tilde{K}}(2g^2) . \tag{2.5}$$

The "second differencing" effect described above is now clearly seen. Observe that in the case of ordinary CV, i.e. $g = 0$, the second difference, in (2.5), occurs on a neighborhood of radius $2h^2$ about the origin. The noise of Cross-Validation may be attributed to the fact that for such arguments, not enough differences are used in construction of the cumulative $\hat{\tilde{K}}$ at these arguments, with the result that it is relatively noisy.

Now observe that the case $g = h$ should yield some improvement, but only in terms of constant coefficient, not in terms of exponent in the rate of convergence, because the number of differences being used is of the same order. This is precisely what happens when this case is analyzed; see Remark 3.6. However, if g is taken to be an order of magnitude larger than h, then there is potential for improvement in the exponent of convergence. This is what provides the large improvement in SCV over CV. Using this intuition in another direction, note that when g is an order of magnitude smaller than h (but still tends to 0 slower than n^{-1}), the third term in (2.5) will be even noisier than the others, so the resulting bandwidth will exhibit variability even larger than that of $\hat{h}(0)$. This can be verified by the techniques used in Sect. 5, but the interest in this case does not seem to justify the effort required. Although CV is SCV_g in the special case $g = 0$, the theory derived in this paper does not apply to CV, because there is a "discontinuity" in our asymptotic arguments at $g = 0$.

It may be expected that some price should be paid for the reductions in variability obtained in this way. This manifests itself in terms of a type of bias, which will be asymptotically quantified in Sect. 3. It is the trade-off between this bias and the reduced variability discussed above which leads to optimal choice of g as developed in Sect. 4.

Usual quantification of the effectiveness of a data driven bandwidth is based on some type of error criterion. We choose MISE for this purpose, for the reasons given in Hall and Marron (1989). That paper also provides bounds on the relative rate of convergence of any data driven bandwidth to h_0, the minimizer of $\text{MISE}(h)$, which set limits to how well the bandwidth may be selected under a variety of smoothness conditions. The bounds are seen to be sharp by showing that they are

attained by a plug-in type bandwidth selector. In the main theorem of the next section it is seen that $\hat{h}$ also achieves the optimal rate of relative convergence to h_0, for each amount of smoothness. However, unlike the plug-in bandwidth selectors, $\hat{h}$ is simple and straightforward.

Since the real goal of density estimation is estimation of f, it is natural to wonder why results, including rates of convergence and limiting distributions, are formulated in terms of h instead of MISE. This is done because, through standard Taylor expansion methods, see Theorem 2.2 in Hall and Marron (1987a), it is seen that the variability in the comparison of $\text{MISE}(\hat{h})$ to $\text{MISE}(h_0)$ is directly driven by the variability in the relative bandwidth error. It is simple to formulate corollaries to the limiting normal distributions, in terms of noncentral (since the asymptotic mean is not 0 in the present case) chi square distributions, but this idea is not new, and does not seem worth the space required for precise formulation.

3 Main theorems

The asymptotic distribution of $(\hat{h} - h_0)/h_0$ depends on four constants $c_1, \ldots, c_4$, which are defined at the end of this section. Let K and L be functions satisfying $\int K = \int L = 1$, and put

$$\kappa_j = (-1)^j (j!)^{-1} \int x^j K(x)\,dx, \quad \lambda_j = (-1)^j (j!)^{-1} \int x^j L(x)\,dx .$$

Assume that K is a kernel of order r, and that L is a kernel of order at least s, in the sense that

$$(3.1) \quad \kappa_j = 0, \text{ for } j = 1, \ldots, r-1, \quad \kappa_r \neq 0, \quad \lambda_j = 0, \text{ for } j = 1, \ldots, s-1 ,$$

where $r \geqq 2$ is an integer and $s \geqq 2$ is an *even* integer.
Additional regularity conditions include:

(3.2) K, L have compact support and are Hölder continuous; L' exists and is Hölder continuous;

(3.3) $L^{(r)}$ exists and is continuous;

(3.4) f is compactly supported and has $\max(2r, r + (s/2))$ bounded derivatives, and for some $0 \leqq q \leqq 1$,

$$\sup_{-\infty < x, y < \infty} |f^{(r+(s/2))}(x) - f^{(r+(s/2))}(y)| / |x - y|^q < \infty .$$

Put $v = (s/2) + q$. Then, roughly speaking, f has "$\max(2r, r + v)$ derivatives" under condition (3.4).

Theorem 3.1 *Assume either that $h_0/g \to c$ where $0 < c < \infty$, and* (3.1), (3.2) *and* (3.4) *hold; or else that $h_0/g \to 0$ and* (3.1)–(3.4) *hold. Then, if $\hat{h}$ denotes the minimizer of* $\text{SCV}_g(h)$,

$$(3.5) \quad (\hat{h} - h_0)/h_0 = c_1^{-1} n^{2(r-1)/(2r+1)} h_0^{2r-2}$$
$$\{(c_2 n^{-2} g^{-(4r+1)} + c_3 n^{-1})^{1/2} Z_n + c_4 g^s + O(g^{2v})\} ,$$

where Z_n is asymptotically normal $N(0, 1)$.

Remark 3.1. At the expense of additional algebraic effort, and moment conditions, each of K, L and f may be taken to have support $(-\infty, \infty)$.

Remark 3.2. If we assume instead of (3.4) that

(3.4′) f is compactly supported and has $\max(2r, r + (s/2))$ continuous derivatives, then Theorem 3.1 remains true provided we replace the $O(g^{2\nu})$ term in (3.5) by $o(g^{2\nu})$.

Remark 3.3. (This shows that $\hat{h}$ may be made $n^{1/2}$ consistent for h_0.) Assume for the sake of simplicity that $\lambda_s = 0$, which means that L is a kernel of order at least $s + 1$. In this case $c_4 = 0$, and so

$$(\hat{h} - h_0)/h_0 = c_5\{(c_2 n^{-2} g^{-(4r+1)} + c_3 n^{-1})^{1/2} Z_n + O(g^{2\nu})\}\,, \tag{3.6}$$

where the constant $c_5 > 0$ is such that

$$(n^{1/(2r+1)} h_0)^{2(r-1)} \to c_1 c_5\,. \tag{3.7}$$

Choose g in (3.6) so that the terms involving g are balanced:

$$(n^{-2} g^{-(4r+1)})^{1/2} = g^{2\nu}, \quad \text{i.e. } g = n^{-1/(2r+2\nu+(1/2))}\,.$$

If $\nu \geqq r + (1/4)$ (observe that this assumption relates to both s and the smoothness of f) then it follows from (3.6) that

$$(\hat{h} - h_0)/h_0 = O_p(n^{-1/2} + n^{-2\nu/(2r+2\nu+(1/2))}) = O_p(n^{-1/2})\,.$$

In other words, $\hat{h}$ is $n^{1/2}$ consistent for h_0, in relative terms.

Remark 3.4. The following result may be concluded from the discussion in Remark 3.3: If K and L are kernels of orders r and $2r + 2$ respectively, if $g = n^{-1/(4r+1)}$, and if f has $r + (1/4)$ derivatives, then $\hat{h}$ is $n^{1/2}$ consistent for h_0. In the important case $r = 2$, this requires L of order 6, f to have 2.25 "derivatives", and $g \sim n^{-1/9}$.

Remark 3.5. An interesting special case is that where $K = L$ and $g = h_0$. Then by (3.5),

$$(\hat{h} - h_0)/h_0 = n^{-1/\{2(2r+1)\}} c_6 Z_n''\,,$$

where Z_n'' is asymptotically normal $N(0, 1)$ and $c_6 > 0$ is a constant. Note particularly that the bandwidth difference $\hat{h} - h_0$ now has an asymptotic distribution with *zero mean*; the asymptotic mean is usually nonzero in other cases discussed above. Furthermore, the convergence rate of relative error is $n^{-1/\{2(2r+1)\}}$, exactly that which arises in ordinary (unsmoothed) cross-validation. To establish this result from Theorem 3.1 requires us to assume $2r$ derivatives of f, but a direct proof demands only $r + (1/4) + \varepsilon$ derivatives, for any $\varepsilon > 0$. This is because the $2r$ derivative assumption is needed only for the term $(c_3 n^{-1})^{1/2}$ in (3.6), and that quantity is negligible in the present case.

Remark 3.6. Observe that the setting of Remark 3.5 is different from the closely related setting where $g = h$ (because here the parameter g appears in the minimization process), as considered by Taylor (1989). However it is straightforward to extend our calculations to cover this case, and the result is the same except for the

value of the constants. This quantifies the intuition expressed in Sect. 2 that the rate of convergence for $h = g$ should be the same as for $\hat{h}(0)$. One feature of interest about the constants is that the analog of c_6 is much smaller than for $\hat{h}(0)$, and in fact is even substantially smaller than the corresponding quantity for BCV as discussed in Scott and Terrell (1987).

A case not covered in Theorem 3.1 is the behaviour of $\hat{h}$ under weaker conditions than existence of $2r$ derivatives. In that direction one may prove the following result, in which assumption (3.4) is replaced by:

(3.4″) f is compactly supported and has $r + r'$ derivatives, where $0 \leqq r' \leqq r - 1$ is an integer and for some $0 \leqq q \leqq 1$,

$$\sup_{-\infty < x, y < \infty} |f^{(r+r')}(x) - f^{(r+r')}(y)|/|x - y|^q < \infty .$$

(If (3.4″) holds then, defining $v = r' + q \leqq r$, we see that f has "$r + v$ derivatives".) Let $[v]$ denote the integer part of v.

Theorem 3.2 *Take $s > 2[v] + 2$. Assume either $h_0/g \to c$ where $0 < c < \infty$, and* (3.1), (3.2) *and* (3.4″) *hold; or $h_0/g \to 0$ and* (3.1)–(3.3) *and* (3.4″) *hold. Then*

$$(\hat{h} - h_0)/h_0 = O_p\{(n^{-2}g^{-(4+1)} + n^{-(2v+1)/(2r+1)} + g^{4v})^{1/2}\} .$$

Theorem 3.2 shows that the SCV bandwidth $\hat{h}$ can achieve the optimal rate of convergence established in Hall and Marron (1989) for $v > 0$.

Next we give explicit forms for the constants $c_1, \ldots, c_4$. Define $L_0 = L$, $L_1 = L'$, $K_0 = K$, $K_1(x) = -xK(x)$,

$$M_i(c, u) = \int \{L_i(u - cx) - L_i(u)\} K_i(x)\, dx ,$$

$$M_{ij}(c, u) = \int M_i(c, v) M_j(c, u + v)\, dv,$$

$$C(c, i, j, k, l) = \int M_{ij}(c, u) M_{kl}(c, u)\, du,$$

for i, j, k, l equal to 0's and 1's. Note that $M_1 = (\partial/\partial c) M_0$. We now define the constants c_i.

(i) c_1: The asymptotic formula for MISE(h) is

$$\text{AMISE}(h) = (nh)^{-1} R(K) + h^{2r} \kappa_r^2 R(f^{(r)}) .$$

Let h_1 (equal to a constant multiple of $n^{-1/(2r+1)}$) be the minimizer of AMISE, and let $c_1 > 0$ be the constant determined by

$$\text{AMISE}''(h_1) = c_1 n^{-2(r-1)/(2r+1)} .$$

(ii) c_2: Here it is necessary to treat the cases $h/g \to 0$ and $h/g \to c$ $(0 < c < \infty)$ a little differently. In the former circumstance, assume L is r times differentiable and put

$$c_2 = 8r^2 \kappa_r^4 R(f) R(L^{(r)} * L^{(r)}) .$$

In the latter case, put

$$c_2 = c_2(c) = 4c^{-(4r-2)} R(f)\{C(c, 1, 1, 0, 0) + C(c, 1, 0, 0, 1)\} .$$

(iii) c_3: Assume f has $2r$ derivatives, and put

$$c_3 = 16r^2 \kappa_r^4 \{ \int (f^{(2r)})^2 f - R(f^{(r)})^2 \} .$$

(iv) c_4: Assume f has $r + (s/2)$ derivatives, and put

$$c_4 = (-1)^{(s/2)+1} 4r \kappa_r^2 \lambda_s R(f^{(r+(s/2))}) .$$

4 Data based presmoothing

In this section optimal choice of g is considered. This choice is then used to motivate a practical data-based value for g. The essential idea is related to the plug-in idea discussed in Park and Marron (1989), where the unknown parts are replaced by their counterparts for some reference distribution, whose scale is adjusted to be the same as the sample standard deviation.

Assume that in (3.1), $\lambda_s \neq 0$. For optimal choice of g, note that the variance and bias parts of Theorem 3.1 (in the case $h/g \to 0$) can be combined into a "Mean Square Error", which is proportional to the criterion

$$C^* = c_2 n^{-2} g^{-(4r+1)} + c_3 n^{-1} + c_4^2 g^{2s} .$$

But C^* is minimized by

$$g_1 = \{(4r + 1)c_2/(2sc_4^2)\}^{1/(4r+2s+1)} n^{-2/(4r+2s+1)} .$$

The choice g_1 is unavailable because c_2 and c_4 depend on the unknown density f. However the asymptotics of the last section reveal that this dependence is less crucial than the dependence of h on f. Hence we replace f by some reference distribution $\tilde{f}$, for example a Gaussian distribution. One aspect of $\tilde{f}$ that clearly has an important influence on the effectiveness of SCV is its scale. Hence we adjust by assuming that $\tilde{f}$ has variance one, and we actually replace f in the above formulae by the rescaling $\tilde{f}_{\hat{\sigma}}$, where $\tilde{f}_\sigma(\cdot) = \tilde{f}(\cdot/\sigma)/\sigma$ and $\hat{\sigma}$ denotes the sample standard deviation (any other measure of scale, for example something more robust, will work here as well).

Hence, using the invariance property $R(\tilde{f}_\sigma^{(k)}) = R(\tilde{f}^{(k)})/\sigma^{2k+1}$, we see that

$$\hat{g} = \hat{\sigma} c_7 n^{-2/(4r+2s+1)} , \tag{4.1}$$

where

$$c_7 = [\{(4r + 1)R(L^{(r)} * L^{(r)}) R(\tilde{f})\}/\{4s\lambda_s^2 R(\tilde{f}^{(r+(s/2))})^2\}]^{1/(4r+2s+1)} .$$

In the important special case that $r = s = 2$, that $K = L = \varphi$ (standard Gaussian), and that $\tilde{f}$ is also standard Gaussian, straightforward calculations show that,

$$g = \hat{\sigma}\{21/(40.2^{1/2})\}^{1/13} n^{-2/13} \cong \hat{\sigma} 0.9266\, n^{-2/13} . \tag{4.2}$$

It is interesting to compare how this Smoothed Cross-Validated bandwidth $\hat{h}(\hat{g})$ compares with the plug-in bandwidth considered in Park and Marron (1989). Theorem 3.1 can be used together with the fact that $\hat{\sigma} = \sigma + O_p(n^{-1/2})$, where σ is

the standard deviation of the f distribution, to give the following analog of Park and Marron's Theorem 3.3:

$$n^{4/13}(\hat{h}(\hat{g}) - h_0)/h_0 \to N(\mu_{\mathrm{SCV}}, \sigma^2_{\mathrm{SCV}}) , \tag{4.3}$$

where

$$\mu_{\mathrm{SCV}} = [\{9R(\varphi'' * \varphi'')R(\tilde{f}_\sigma)\}/\{2R(\tilde{f}_\sigma^{(3)})^2\}]^{2/13}[R(f^{(3)})/\{5R(f^{(2)})\}] ,$$

$$\sigma^2_{\mathrm{SCV}} = [\{2^9 R(\varphi'' * \varphi'')^4 R(\tilde{f}_\sigma^{(3)})^{18}\}/\{9^9 R(\tilde{f}_\sigma)^9\}]^{1/13}[2R(f)/\{25R(f^{(2)})^2\}] .$$

When comparison is done with the corresponding quantities μ_{PI} and σ^2_{PI} in Park and Marron,

$$\mu_{PI}/\mu_{\mathrm{SCV}} = \{R(\varphi^{(4)})/R(\varphi'' * \varphi'')\}^{2/13}\{R(\tilde{f}_\sigma^{(2)})/R(f^{(2)})\}^{4/13} ,$$

$$\sigma^2_{PI}/\sigma^2_{\mathrm{SCV}} = \{R(\varphi'' * \varphi'')/R(\varphi^{(4)})\}^{9/13}\{R(f^{(2)})/R(\tilde{f}_\sigma^{(2)})\}^{18/13}$$

$$= (\mu_{\mathrm{SCV}}/\mu_{PI})^{9/2} .$$

It follows from this that no clear comparison can be made between the two bandwidth selectors. When one of them has larger variance, the other has larger bias.

5 Proofs

5.1 Preparatory lemmas

In this section, we state and prove a sequence of seven lemmas which are needed for the proof of Theorem 3.1. Notation is drawn from Sect. 3. Our first lemma is a relatively straightforward result in functional analysis; a version of it is proved in detail in Lemma 5.4 of Hall and Marron (1989).

Lemma 5.1 *If the function H vanishes outside a compact set and satisfies*

$$|H(x + y) - H(x)| \leqq C_1|y|^q, \quad \text{for } -\infty < x < \infty,$$

where $0 \leqq q \leqq 1$; and if $\varepsilon \to 0$ and $\sup_\varepsilon |C_2(\varepsilon)|$ is bounded; then

$$\int H(z)[H\{z + \varepsilon + C_2(\varepsilon)\} - H\{z + C_2(\varepsilon)\}]\,dz = O(|\varepsilon|^{2q}) .$$

Define $D = K_h * L_g - L_g$, $a(x) = E\{D(x - X)\}$, $b(x) = E\{\hat{f}(x)\} - f(x)$. Put $a_h = (\partial/\partial h)a$, and define b_h, D_h similarly.

Lemma 5.2 *Assume conditions* (3.1), (3.2) *and* (3.4). *Then as $h, g \to 0$,*

$$b(x) = \kappa_r h^r f^{(r)}(x) + o(h^r),\ a(x) - b(x) = o(h^r) , \tag{5.1}$$

$$b_h(x) = r\kappa_r h^{r-1} f^{(r)}(x) + o(h^{r-1}) ,$$

uniformly in x. If in addition h/g is bounded then

$$\int (a_h - b_h)(a + b) = -(c_4/2)h^{2r-1}g^s + O(h^{2r-1}g^{2\nu}) , \tag{5.2}$$

$$\int (a - b)(a_h + b_h) = -(c_4/2)h^{2r-1}g^s + O(h^{2r-1}g^{2\nu}) . \tag{5.3}$$

Proof of Lemma 5.2. Results (5.1) follow by the usual Taylor expansion arguments used to approximate bias in density estimation. Of results (5.2) and (5.3), we shall only prove (5.2) here.
Observe that

$$a(x) = \iint K(u)L(v)\{f(x - hu - gv) - f(x - gv)\}\,du\,dv = b(x) + \delta(x)\,,$$

where

$$\delta(x) = \iint K(u)L(v)[f(x - hu - gv) - f(x - gv) - \{f(x - hu) - f(x)\}]\,du\,dv\,. \tag{5.4}$$

To prove (5.2) we must show that

$$\int \delta_h(2b + \delta) = -(c_4/2)\,h^{2r-1}g^s + O(h^{2r-1}g^{2v})\,,$$

for which it is sufficient to establish that

$$\int \delta_h b = (-1)^{s/2} r\kappa_r^2 \lambda_s h^{2r-1} g^s \int (f^{(r+(s/2))})^2 + O(h^{2r-1}g^{2v})\,, \tag{5.5}$$

$$\int \delta_h \delta = O(h^{2r-1}g^{2v})\,. \tag{5.6}$$

We shall prove only (5.5), since the argument leading to (5.6) is similar.

Assume temporarily that f has $r + s$ derivatives. Differentiating under the integral sign in (5.4), and then Taylor expanding the integrand using an integral formula for the remainder, we see that

$$\begin{aligned}
h\,\delta_h(x) &= \iint K(u)L(v)(-hu)\{f'(x - hu - gv) - f'(x - hu)\}\,du\,dv \\
&= \{(r-2)!\}^{-1} \iint K(u)L(v)(-hu)^r \int_0^1 (1-t)^{r-2}\{f^{(r)}(x - hut - gv) \\
&\quad - f^{(r)}(x - hut)\}\,dt\,du\,dv \\
&= \{(r-2)!\,(s-1)!\}^{-1} \iint K(u)\,L(v)\,(-hu)^r(-gv)^s \\
&\quad \times \int_0^1 \int_0^1 (1-t_1)^{r-2}(1-t_2)^{s-1} f^{(r+s)}(x - hut_1 - gvt_2)\,dt_1\,dt_2\,du\,dv\,.
\end{aligned}$$

A simpler argument shows that

$$\begin{aligned}
b(x) &= \int K(u)\{f(x - hu) - f(x)\}\,du \\
&= \{(r-1)!\}^{-1} \int K(u)(-hu)^r \int_0^1 (1-t)^{r-1} f^{(r)}(x - hut)\,dt\,du\,.
\end{aligned}$$

Therefore

$$\begin{aligned}
(r-1)!\,(r-2)!\,(s-1)!\,h \int \delta_h b &= \iiint K(u)L(v)K(w)(-hu)^r(-gv)^s(-hw)^r \\
&\int_0^1\int_0^1\int_0^1 (1-t_1)^{r-2}(1-t_2)^{s-1}(1-t_3)^{r-1}\,U\,dt_1\,dt_2\,dt_3\,du\,dv\,dw\,,
\end{aligned} \tag{5.7}$$

where

$$\begin{aligned}
U &= U(u, v, w, t_1, t_2, t_3) \\
&= \int f^{(r+s)}(x - hut_1 - gvt_2) f^{(r)}(x - hwt_3)\,dx \\
&= (-1)^{s/2} \int f^{(r+(s/2))}(x - hut_1 - gvt_2) f^{(r+(s/2))}(x - hwt_3)\,dx\,,
\end{aligned}$$

on integrating by parts $s/2$ times. If U has this form then identity (5.7) for $\int \delta_h b$ requires only $r + (s/2)$ derivatives of f, not $r + s$ derivatives.
From (5.7) and Lemma 5.1 we conclude that

$$\int \delta_h b = (-1)^{s/2} r\kappa_r^2 \lambda_s h^{2r-1} g^s \int (f^{(r+(s/2))})^2 + O(h^{2r-1} g^{s'}) ,$$

where $s' = s + 2\{v - (s/2)\} = 2v$. This establishes (5.5).
Define

$$I_1 = \iint E\{D(x - X) D(y - X)\} a_h(x) a_h(y) \, dx \, dy ,$$

$$I_2 = \iint E\{D_h(x - X) D_h(y - X)\} a(x) a(y) \, dx \, dy ,$$

$$I_3 = \iint E\{D(x - X) D_h(y - X)\} a_h(x) a(y) \, dx \, dy ,$$

$$I_4 = \iint E\{D_h(x - X) D(y - X)\} a(x) a_h(y) \, dx \, dy .$$

Lemma 5.3 *Let I denote any one of $I_1, \ldots, I_4$, and assume conditions* (3.1), (3.2) *and* (3.4). *Then as $h, g \to 0$,*

$$I \sim h^{4r-2} \kappa_r^4 r^2 \int (f^{(2r)})^2 f .$$

Proof of Lemma 5.3. Recall the definitions $L_0 = L, L_1 = L', K_0 = K, K_1(x) = -xK(x)$.
Define

$$Q_i(x) = \int [L_i\{(x - h\xi)/g\} - L_i(x/g)] K_i(\xi) \, d\xi .$$

Then

$$D = g^{-1} Q_0, \quad D_h = g^{-2} Q_1 . \tag{5.8}$$

We begin by developing a formula for

$$R_1 = g^{-2} \iint E\{Q_i(x - X) Q_j(y - X)\} A_i(x) A_j(y) \, dx \, dy ,$$

where A_i, A_j are given functions. For that task it is helpful to note that

$$(-1)^k (k!)^{-1} \int x^k K_i(x) \, dx = \begin{cases} 0 & \text{for } 1 \leqq k \leqq r_i - 1 \\ \kappa(i) & \text{for } k = r_i , \end{cases} \tag{5.9}$$

where $r_0 = r, r_1 = r - 1, \kappa(0) = \kappa_r, \kappa(1) = r\kappa_r$.
Observe that

$$E\{Q_i(x - X) Q_j(y - X)\} = \iiint [L_i\{(x - u - h\xi)/g\} - L_i\{(x - u)/g\}] \tag{5.10}$$
$$[L_j\{(y - u - h\eta)/g\} - L_j\{(y - u)/g\}] f(u) K_i(\xi) K_j(\eta) \, du \, d\xi \, d\eta .$$

Therefore

$$R_1 = \iiint U(u, \xi, \eta) f(u) K_i(\xi) K_j(\eta) \, du \, d\xi \, d\eta ,$$

where

$$\begin{aligned} U(u, \xi, \eta) &= g^{-2} \iint [L_i\{(x - u - h\xi)/g\} - L_i\{(x - u)/g\}] \\ &\quad [L_j\{(y - u - h\eta)/g\} - L_j\{(y - u)/g\}] A_i(x) A_j(y) \, dx \, dy \\ &= \iint \{L_i(x - hg^{-1}\xi) - L_i(x)\} \{L_j(y - hg^{-1}\eta) - L_j(y)\} \\ &\quad A_i(u + gx) A_j(u + gy) \, dx \, dy \\ &= \iint L_i(x) L_j(y) \{A_i(u + h\xi + gx) - A_i(u + gx)\} \\ &\quad \{A_j(u + h\eta + gy) - A_j(u + gy)\} \, dx \, dy . \end{aligned}$$

Hence

$$
\begin{aligned}
(5.11)\quad R_1 &= \iiint f(u) L_i(x) L_j(y) \Big[\int \{ A_i(u + h\xi + gx) - A_i(u + gx) \} K_i(\xi)\, d\xi \Big] \\
&\quad \times \Big[\int \{ A_j(u + h\eta + gy) - A_j(u + gy) \} K_j(\eta)\, d\eta \Big]\, du\, dx\, dy \\
&= \iiint f(u) L_i(x) L_j(y) \Big[\{(r_i - 1)!\}^{-1} \int_0^1 (1 - t)^{r_i - 1} \int (h\xi)^{r_i} \\
&\quad \times A_i^{(r_i)}(u + ht\,\xi + gx) K_i(\xi)\, d\xi\, dt \Big] \\
&\quad \times \Big[\{(r_j - 1)!\}^{-1} \int_0^1 (1 - t)^{r_j - 1} \int (h\eta)^{r_j} A_j^{(r_j)}(u + ht\eta + gy) \\
&\quad \times K_j(\eta)\, d\eta\, dt \Big]\, du\, dx\, dy \\
&= h^{r_i + r_j} \{(r_i - 1)!\,(r_j - 1)!\}^{-1} \int_0^1 (1 - t_1)^{r_i - 1} \int_0^1 (1 - t_2)^{r_j - 1} \\
&\quad \times \int \xi^{r_i} K_i(\xi) \int \eta^{r_j} K_j(\eta) \int f(u) \\
&\quad \times \int A_i^{(r_i)}(u + ht_1\xi + gx) L_i(x)\, dx \int A_j^{(r_j)}(u + ht_2\xi + gy) L_j(y) \\
&\quad \times dy\, du\, d\xi\, d\eta\, dt_1\, dt_2 \\
&= (-1)^{i+j} h^{r_i + r_j} g^{i+j} \{(r_i - 1)!\,(r_j - 1)!\}^{-1} \int_0^1 (1 - t_1)^{r_i - 1} \int_0^1 (1 - t_2)^{r_j - 1} \\
&\quad \times \int \xi^{r_i} K_i(\xi) \int \eta^{r_j} K_j(\eta) \int f(u) \int A_i^{(r_i + i)}(u + ht_1\xi + gx) L(x) \\
&\quad \times \int A_j^{(r_j + j)}(u + ht_2\eta + gy) L(y)\, dy\, dx\, du\, d\eta\, d\xi\, dt_2\, dt_1,
\end{aligned}
$$

the last identity following on integrating by parts.
Note that $r_i + i = r_j + j = r$, and assume that for $k = i, j$ we have

$$
A_k^{(r)}(x) = c_k h^{p_k} f^{(2r)}(x) + o(h^{p_k}) \tag{5.12}
$$

uniformly in x, where p_k is a positive integer and c_k is a constant. Then results (5.9) and (5.11) imply that

$$
\begin{aligned}
(5.13)\quad R_1 &\sim (-1)^{i+j+r_i+r_j} h^{r_i + r_j + p_i + p_j} g^{i+j} c_i c_j \kappa(i) \kappa(j) \\
&\quad \iiint f(u) f^{(2r)}(u)^2 L(x) L(y)\, du\, dx\, dy \\
&= h^{2r + p_i + p_j - i - j} g^{i+j} c_i c_j \kappa_r^2 r^{i+j} \int (f^{(2r)})^2 f .
\end{aligned}
$$

Recall $a(x) = E\{D(x - X)\}$ and note that by Lemma 5.2, $a(x) = \kappa_r h^r f^{(r)}(x) + o(h^r)$. Similarly it may be proved that $a^{(r)}(x) = \kappa_r h^r f^{(2r)}(x) + o(h^r)$ and $a_h^{(r)}(x) = r\kappa_r h^{r-1} f^{(2r)}(x) + o(h^{r-1})$, uniformly in x. Hence if we define

$$
A_k(x) = (\partial/\partial h)^{q_k} a(x)
$$

where $q_k = 0$ or 1, then (5.12) holds with $c_k = r^{q_k} \kappa_r$ and $p_k = r - q_k$. Therefore by (5.13),

$$
R_1 \sim h^{4r - (i + j + q_i + q_j)} g^{i+j} \kappa_r^4 r^{i + j + q_i + q_j} \int (f^{(2r)})^2 f . \tag{5.14}
$$

In the cases $I = I_1, I_2, I_3, I_4$ we have $(i, j, q_i, q_j) = (0, 0, 1, 1), (1, 1, 0, 0), (1, 0, 1, 0), (1, 0, 0, 1)$, respectively. On each occasion $i + j + q_i + q_j = 2$, and by (5.8), $I = g^{-(i+j)} R_1$. Therefore the lemma follows from (5.14).

Define

$$J_1 = \iint E\{D_h(x - X) D_h(y - X)\}\, E\{D(x - X) D(y - X)\}\, dx\, dy\,,$$

$$J_2 = \iint E\{D_h(x - X) D(y - X)\}\, E\{D(x - X) D(y - X)\}\, dx\, dy\,.$$

Recall the definition of $C(c, i, j, k, l)$ given in Sect. 3.

Lemma 5.4 (i) *Assume conditions* (3.1)–(3.4), *and that* $h, g, h/g \to 0$. *Let* J *denote either* J_1 *or* J_2. *Then*

$$J = h^{4r-2} g^{-(4r+1)} \kappa_r^4 r^2 R(f)\, R(L^{(r)} * L^{(r)}) + o(h^{4r-2} g^{-(4r+1)}).$$

(ii) *Assume conditions* (3.1), (3.2) *and* (3.4), *and that* $h, g \to 0$ *and* $h/g \to c$ *where* $0 < c < \infty$. *Then*

$$J_1 = g^{-3} R(f)\, C(c, 1, 1, 0, 0) + o(g^{-3})$$

$$J_2 = g^{-3} R(f)\, C(c, 1, 0, 0, 1) + o(g^{-3})$$

Proof of Lemma 5.4. Adopt notation from the proof of Lemma 5.3. We begin by developing a formula for

$$R_2 = g^{-3} \iint E\{Q_i(x - X) Q_j(y - X)\}\, E\{Q_k(x - X) Q_l(y - X)\}\, dx\, dy\,,$$

where i, j, k, l are 0's and 1's. By (5.10),

$$\begin{aligned}
(5.15)\quad R_2 &= g^{-3} \iint [\iiint [L_i\{(x - u - h\xi)/g\} - L_i\{(x - u)/g\}] \\
&\quad \times [L_j\{(y - u - h\eta)/g\} - L_j\{(y - u)/g\}] f(u) K_i(\xi) K_j(\eta)\, du\, d\xi\, d\eta] \\
&\qquad [\text{same with } (k, l) \text{ replacing } (i, j)]\, dx\, dy \\
&= g^{-1} \iint [\iiint \{L_i(u - hg^{-1}\xi) - L_i(u)\} [L_j\{u + (y - x)g^{-1} - hg^{-1}\eta\} \\
&\quad - L_j\{u + (y - x)g^{-1}\}] f(x - gu) K_i(\xi) K_j(\eta)\, du\, d\xi\, d\eta] \\
&\qquad [\text{same with } (k, l) \text{ replacing } (i, j)]\, dx\, dy \\
&= \iint [\int f(x - gu) [\int \{L_i(u - hg^{-1}\xi) - L_i(u)\} K_i(\xi)\, d\xi] \\
&\quad \times [\int \{L_j(u + v - hg^{-1}\eta) - L_j(u + v)\} K_j(\eta)\, d\eta]\, du] \\
&\qquad [\text{same with } (k, l) \text{ replacing } (i, j)]\, dx\, dv\,.
\end{aligned}$$

If $h/g \to 0$ then by (5.15), with $r_* = r_i + r_j + r_k + r_l$ and $\kappa_* = \kappa(i)\kappa(j)\kappa(k)\kappa(l)$, we have

$$\begin{aligned}
(5.16)\quad R_2 &= \iint [\int f(x - gu) \{\int L_i^{(r_i)}(u) (-hg^{-1}\xi)^{r_i} (r_i!)^{-1} K_i(\xi)\, d\xi\} \\
&\quad \times \{\int \{L_j^{(r_j)}(u + v) (-hg^{-1}\eta)^{r_j} (r_j!)^{-1} K_j(\eta)\, d\eta\}\, du] \\
&\qquad [\text{same with } (k, l) \text{ replacing } (i, j)]\, dx\, dv + o\{(h/g)^{r_*}\} \\
&= (-h/g)^{r_*} \kappa_* (\int f^2) \int \{\int L_i^{(r_i)}(u)\, L_j^{(r_j)}(u + v)\, du\} \\
&\quad \times \{\int L_k^{(r_k)}(u)\, L_l^{(r_l)}(u + v)\, du\}\, dv + o\{(h/g)^{r_*}\}\,.
\end{aligned}$$

If $m = i, j, k$ or l then $r_m + m = r$, and so $L_m^{(r_m)} = L^{(r)}$. Since $i + j + k + l = 2$,

$$r_* = 4r - (i + j + k + l) = 4r - 2, \quad \kappa_* = \kappa_r^4 r^{i+j+k+l} = \kappa_r^4 r^2 .$$

Therefore by (5.16),

$$R_2 = (h/g)^{4r-2} \kappa_r^4 r^2 R(f) R(L^{(r)} * L^{(r)}) + o\{(h/g)^{4r-2}\} . \tag{5.17}$$

Note finally that by (5.8), $J_1 = g^{-3} R_2$ with $(i, j, k, l) = (1, 1, 0, 0)$, and $J_2 = g^{-3} R_2$ with $(i, j, k, l) = (1, 0, 0, 1)$. Hence the lemma in the case $h/g \to 0$ follows from (5.17).

If $h/g \to c$ where $0 < c < \infty$ then the lemma follows from (5.15) and the fact that $J = g^{-3} R_2$.

Define $G_i(x) = D(x - X_i) - E\{D(x - X_i)\}$, $V_i(h) = \int G_i a$, $V_{ij}(h) = \int G_i G_j$.

Lemma 5.5 *Assume conditions* (3.1), (3.2) *and* (3.4). *Then as* $h, g \to 0$,

$$E\{V_1'(h)^2\} \sim 4h^{4r-2} \kappa_r^4 r^2 [\int (f^{(2r)})^2 f - R(f^{(r)})^2] .$$

Proof of Lemma 5.5. Recall that $a(x) = E\{D(x - X)\}$ and $a_h = (\partial/\partial h) a$, and define $G(x) = D(x - X) - a(x)$, $G_h = (\partial/\partial h) G$. Then

$$V_1(h) \overset{\text{dist.}}{=} V(h) \equiv \int Ga$$

$$V_1'(h) \overset{\text{dist.}}{=} V'(h) \equiv \int (G_h a + G a_h) ,$$

$$\begin{aligned} E\{V'(h)^2\} &= \iint [E\{G_h(x) G_h(y)\} a(x) a(y) \\ &\quad + 2E\{G_h(x) G(y)\} a(x) a_h(y) + E\{G(x) G(y)\} a_h(x) a_h(y)] \, dx \, dy \\ &= \iint [E\{D_h(x - X) D_h(y - X)\} a(x) a(y) \\ &\quad + 2E\{D_h(x - X) D(y - X)\} a(x) a_h(y) \\ &\quad + E\{D(x - X) D(y - X)\} a_h(x) a_h(y) \\ &\quad - 4a(x) a(y) a_h(x) a_h(y)] \, dx \, dy . \end{aligned}$$

Now apply Lemmas 5.2 and 5.3 to evaluate each of these double integrals. Note that there exists a compact set $[u, v]$, not depending on h, outside which both a and a_h vanish for all sufficiently small h.

Lemma 5.6 (i) *Assume conditions* (3.1)–(3.4), *and that* $h, g, h/g \to 0$. *Then*

$$E\{V_{12}'(h)^2\} \sim 4h^{4r-2} g^{-(4r+1)} \kappa_r^4 r^2 R(f) R(L^{(r)} * L^{(r)}) .$$

(ii) *Assume conditions* (3.1), (3.2) *and* (3.4), *and* $h, g \to 0$, *where* $h/g \to c$ *for* $0 < c < \infty$. *Then*

$$E\{V_{12}'(h)^2\} \sim 2g^{-3} R(f) \{C(c, 1, 1, 0, 0) + C(c, 1, 0, 1, 0)\} .$$

Proof of Lemma 5.6. Note that

$$V_{12}(h) = \int G_1 G_2,\ V'_{12}(h) = \int (G_1 G_{2h} + G_{1h} G_2)\,,$$

$$\begin{aligned}(1/2)E\{V'_{12}(h)^2\} &= \iint [E\{G_1(x)G_1(y)\}\, E\{G_{2h}(x)G_{2h}(y)\} \\ &\quad + E\{G_1(x)G_{1h}(y)\}\, E\{G_{2h}(x)G_2(y)\}]\,dx\,dy \\ &= \iint [E\{D(x-X)D(y-X)\}\, E\{D_h(x-X)D_h(y-X)\} \\ &\quad + E\{D_h(x-X)D(y-X)\}\, E\{D(x-X)D_h(y-X)\} \\ &\quad - E\{D(x-X)D(y-X)\}\, a_h(x)a_h(y) \\ &\quad - E\{D_h(x-X)D_h(y-X)\}\, a(x)a(y) \\ &\quad - 2E\{D_h(x-X)D(y-X)\}\, a_h(x)a(y) \\ &\quad + 2a(x)a(y)a_h(x)a_h(y)]\,dx\,dy\,.\end{aligned}$$

Now apply Lemmas 5.2–5.4.

Lemma 5.7 *Assume conditions* (3.2) *and* (3.4), *and that* $g \to 0$ *and* h_0/g *is bounded. Then for some* $\varepsilon > 0$ *and any* $0 < a < b < \infty$,

$$\sup_{a \le t \le b} \left| n^{-1} \sum_i \{V'_i(th_0) - V'_i(h_0)\} \right| = O_p(h_0^{2r-1} n^{-(1/2)-\varepsilon})\,,$$

$$\sup_{a \le t \le b} \left| n^{-2} \sum\sum_{i \ne j} \{V'_{ij}(th_0) - V'_{ij}(h_0)\} \right| = O_p(h_0^{2r-1} g^{-(2r+(1/2))} n^{-1-\varepsilon})\,.$$

The proof of Lemma 5.7 is very close to that of Lemma 3.2 in Hall and Marron (1987a), and so will not be given here.

5.2 Proof of Theorem 3.1. Put $\Delta = \mathrm{SCV}_g - \mathrm{MISE}$. Then

$$0 = \mathrm{SCV}'_g(\hat{h}) = \Delta'(\hat{h}) + \mathrm{MISE}'(\hat{h})\,,$$

$$\mathrm{MISE}'(\hat{h}) = \mathrm{MISE}'(h_0) + (\hat{h} - h_0)\,\mathrm{MISE}''(h^*) = (\hat{h} - h_0)\,\mathrm{MISE}''(h^*)\,,$$

where h^* lies between h_0 and $\hat{h}$. Therefore

$$\Delta'(\hat{h}) + (\hat{h} - h_0)\,\mathrm{MISE}''(h^*) = 0\,,$$

whence

$$\hat{h} - h_0 = -\Delta'(\hat{h})/\mathrm{MISE}''(h^*)\,. \tag{5.18}$$

Since $h^*/h_0 \to 1$ in probability, it is simple to prove that

$$\mathrm{MISE}''(h^*)/\mathrm{MISE}''(h_0) \to 1$$

in probability. Straightforward calculations give

$$\mathrm{MISE}''(h_0) \sim c_1 n^{-2(r-1)/(2r+1)}\,.$$

Combining the estimates from (5.18) down we see that

$$\hat{h} - h_0 = -c_1^{-1} n^{2(r-1)/(2r+1)} \{1 + o_p(1)\} \Delta'(\hat{h}) . \tag{5.19}$$

Recall from (2.2) and (2.3) that

$$\mathrm{SCV}_g(h) = (nh)^{-1} R(K) + \hat{B}_g(h)$$

where, with $D = K_h * L_g - L_g$,

$$\hat{B}_g(h) = \{n(n-1)\}^{-1} \sum\sum_{i \neq j} \int D(x - X_i) D(x - X_j) \, dx .$$

Recalling the notation $b(x) = E\hat{f}(x) - f(x)$ from lemma 5.2, we have

$$\mathrm{MISE} = \int \mathrm{var}(\hat{f}) + R(b) = (nh)^{-1} R(K) - n^{-1} R(E\hat{f}) + R(b) .$$

Therefore

$$\Delta = \mathrm{SCV} - \mathrm{MISE} = \Delta_1 + n^{-1} R(E\hat{f}) ,$$

where

$$\Delta_1 = \hat{B}_g - R(b) . \tag{5.20}$$

Straightforward calculations give

$$(\partial/\partial h) \, R(E\hat{f}) = O(h^{r-1}) ,$$

and so

$$\Delta'(\hat{h}) = \Delta_1'(\hat{h}) + O_p(n^{-3r/(2r+1)}) . \tag{5.21}$$

Put $B = E(\hat{B}_g)$, and note that

$$\hat{B}_g(h) - B(h) = \{n\{n-1\}\}^{-1} \sum\sum_{i \neq j} V_{ij}(h) + 2n^{-1} \sum_i V_i(h) ,$$

$$B(h) - R(b) = \int (a^2 - b^2) = \int (a - b)(a + b) .$$

Hence, using as above the subscript h to denote partial differentiation with respect to h, we see from (5.20) that

$$\Delta_1'(h) = \{n(n-1)\}^{-1} \sum\sum_{i \neq j} V_{ij}'(h) + 2n^{-1} \sum_i V_i'(h) \tag{5.22}$$

$$+ \int \{(a_h - b_h)(a + b) + (a - b)(a_h + b_h)\} .$$

Lemma 5.2 implies that

$$\int \{(a_h - b_h)(a + b) + (a - b)(a_h + b_h)\} = -c_4 h^{2r-1} g^s + O(h^{2r-1} g^{2v}) .$$

Since $\hat{h}/h_0 \to 1$ in probability then by Lemma 5.7,

$$n^{-2} \left| \sum\sum_{i \neq j} \{V_{ij}'(\hat{h}) - V_{ij}'(h_0)\} \right| + n^{-1} \left| \sum_i \{V_i'(\hat{h}) - V_i'(h_0)\} \right|$$

$$= o_p\{h_0^{2r-1} (n^{-1} g^{-(2r+(1/2))} + n^{-1/2})\} .$$

Combining estimates from (5.22) down we conclude that

$$\Delta_1'(\hat{h}) = \{n(n-1)\}^{-1} \sum\sum_{i \neq j} V_{ij}'(h_0) + 2n^{-1} \sum_i V_i'(h_0)$$
$$- c_4 h_0^{2r-1} g^s + O(h_0^{2r-1} g^{2v}) + o_p\{h_0^{2r-1}(n^{-1} g^{-(2r+(1/2))} + n^{-1/2})\} .$$

This result, (5.19) and (5.21) give

$$\hat{h} - h_0 = -c_1^{-1} n^{2(r-1)/(2r+1)} [\{n(n-1)\}^{-1} \sum\sum_{i \neq j} V_{ij}'(h_0) + 2n^{-1} \sum_i V_i'(h_0)$$
$$- c_4 h_0^{2r-1} g^s + O(h_0^{2r-1} g^{2v})$$
$$+ o_p\{h_0^{2r-1}(n^{-1} g^{-(2r+(1/2))} + n^{-1/2})\}] . \tag{5.23}$$

The Cramér-Wold device and a martingale argument, as in Hall (1984) and on p. 578 of Hall and Marron (1987a), may be used to prove that

$$S = \{n(n-1)\}^{-1} \sum\sum_{i \neq j} V_{ij}'(h_0) + 2n^{-1} \sum_i V_i'(h_0)$$

is asymptotically normally distributed with zero mean and variance

$$E(S^2) = 2\{n(n-1)\}^{-1} E\{V_{12}'(h_0)^2\} + 4n^{-1} E\{V_i'(h_0)^2\}$$
$$\sim c_2 n^{-2} h_0^{4r-2} g^{-(4r+1)} + c_3 n^{-1} h_0^{4r-2} ,$$

where we have used Lemmas 5.5 and 5.6. Hence by (5.23),

$$\hat{h} - h_0 = c_1^{-1} n^{2(r-1)/(2r+1)} h_0^{2r-1} \{(c_2 n^{-2} g^{-(4r+1)} + c_3 n^{-1})^{1/2} Z_n$$
$$+ c_4 g^s + O(g^{2v})\} ,$$

where Z_n is asymptotically normal $N(0, 1)$. This completes the proof of Theorem 3.1.

Acknowledgement. A key idea in the start of this research was suggested by D. Nychka. The $\hat{K}$ interpretation of smoothed cross-validation is a consequence of a remark made by P.J. Diggle.

References

Bickel, P., Ritov, Y.: Estimating integrated squared density derivatives. Sankhyā Ser. A. **50**, 381–393 (1988)

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

Burkholder, D.L.: Distribution function inequalities for martingales. Ann. Probab. **1**, 19–42 (1973)

Burman, P.: A data dependent approach to density estimation, Z. Wahrscheinlichkeits theor. Verw. Geb. **69**, 609–628 (1985)

Chiu, S.-T.: Bandwidth selection for kernel density estimation. Ann. Stat. to appear (1991)

Devroye, L., Györfi, L.: Nonparametric density estimation: The L_1 View. New York: Wiley 1984

Diggle, P.J.: Statistical analysis of point patterns London: Academic Press 1983

Diggle, P.J.: A kernel method for smoothing point process data. Appl. Stat. **34**, 138–147 (1985)

Diggle, P.J., Marron, J.S.: Equivalence of smoothing parameter selectors in density and intensity estimation. J. Am. Stat. Assoc. **83**, 793–800 (1988)

Eubank, R.L.: Spline smoothing and nonparametric regression. New York: Dekker 1988

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

Gasser, T., Müller, H-G., Mammitzsch, V.: Kernels for nonparametric curve estimation. J. R. Stat. Soc., Ser. **B47**, 238–252 (1985)
Härdle, W.: Applied nonparametric regression. Econometrics Society Monograph Series, No. 19, Cambridge: Cambridge University Press 1989
Härdle, W., Hall, P., Marron, J.S.: How far are automatically chosen regression smoothers from their optimum? (with discussion). J. Am. Stat. Assoc. **83**, 86–95 (1988)
Hall, P.: Objective methods for the estimation of window size in the nonparametric estimation of a density (unpublished manuscript, 1980)
Hall, P.: Large sample optimality of least squares cross-validation in density estimation. Ann. Stat. **11**, 1156–1174 (1983)
Hall, P.: Central limit theorem for integrated squared error of multivariate density estimators. J. Multivariate Anal. **14**, 1–16 (1984)
Hall, P., Marron, J.S.: Extent to which least-squares cross-validation minimises integrated square error in nonparametric density estimation. Probab. Th. Rel. Fields **74**, 567–581 (1987a)
Hall, P. Marron, J.S.: On the amount of noise inherent in bandwidth selection for a kernel density estimator. Ann. Stat. 15, 163–181 (1987b)
Hall, P., Marron, J.S.: Estimation of integrated squared density derivatives. Stat. Probab. Lett. **6**, 109–115 (1987c)
Hall, P., Marron, J.S.: Lower bounds for bandwidth selection in density estimation (unpublished manuscript, 1989)
Hall, P., Sheather, S., Jones, M.C., Marron, J.S.: On optimal data-based bandwidth selection in kernel density estimation. (unpublished manuscript, 1989)
Marron, J.S.: Automatic smoothing parameter selection: A survey. Emp. Econ. **13**, 187–208 (1988)
Müller, H.G.: Empirical bandwidth choice for nonparametric kernel regression by means of pilot estimators. Stat. Decis. **2**, [Suppl] 193–206 (1985)
Müller, H.G.: Nonparametric analysis of longitudinal data. Berlin Heidelberg New York: Springer 1988
Park, B.U., Marron, J.S.: Comparison of data-driven bandwidth selectors. J. Am. Stat. Assoc. **85**, 66–72 (1990)
Ripley, B.D.: Spatial statistics New York: Wiley 1981
Rudemo, M.: Empirical choice of histograms and kernel density estimators. Scand. J. Stat. **9**, 65–78 (1982)
Scott, D.W.: Averaged shifted histograms: effective nonparametric density estimation in several dimensions. Ann. Stat. **4**, 1024–1040 (1985)
Scott, D.W., Factor, L.E.: Monte Carlo study of three data-based nonparametric density estimators. J. Am. Stat. Assoc. **76**, 9–15 (1981)
Scott, D.W., Terrell, G.R.: Biased and unbiased cross-validation in density estimation. J. Am. Stat. Assoc. **82**, 1131–1146 (1987)
Scott, D.W., Tapia, R.A., Thompson, J.W.: Kernel density estimation revisited. J. Nonlinear Anal., Theor. Methods Appl. **1**, 339–372 (1977)
Sheather, S.J.: An improved data-based algorithm for choosing the window width when estimating the density at a point. Comput. Stat. Data Anal. **4**, 61–65 (1986)
Silverman, B.W.: Density estimation for statistics and data analysis. New York: Chapman and Hall 1986
Staniswallis, J.G.: Local bandwidth selection for kernel estimates J. Am. Stat. Assoc. **84**, 284–288 (1987)
Stone, C.J.: An asymptotically optimal window selection rule for kernel density estimates. Ann. Stat. **12**, 1285–1297 (1984).
Taylor, C.C.: Bootstrap choice of the smoothing parameter in kernel density estimation Biometrika **76**, 705–712 (1989)
Woodroofe, M.: On choosing a delta sequence. Ann. Math. Stat. **41**, 1665–1671 (1970)