A LOCAL CROSS-VALIDATION ALGORITHM

Peter Hall
Australian National University

William R. Schucany
Southern Methodist University

SMU/DS/TR-219

January 1988

A LOCAL CROSS-VALIDATION ALGORITHM

Peter Hall and William R. Schucany
Australian National University
Canberra, 2601, Australia

Southern Methodist University
Dallas, TX 75275, USA

Abstract: The usual form of cross-validation is global in character, and is designed to estimate a density in some "average" sense over its entire support. In this paper we present a local version of squared-error cross-validation, suitable for estimating a probability density at a given point. It is shown theoretically to be asymptotically optimal in the sense of minimizing mean squared error. Numerical examples illustrate finite sample characteristics, and show that local cross-validation is a practical algorithm.

Keywords: cross-validation, density estimation, kernel method, local estimation, window size.

1. Introduction.

The technique of squared-error cross-validation was suggested by Rudemo (1982) and Bowman (1984), and has had considerable influence on the practice and methodology of nonparametric density estimation. It permits simple, automatic selection of a smoothing parameter, for example of the window size in kernel density estimation. It is known to be free from difficulties which can arise with other sorts of cross-validation, such as likelihood cross-validation. In particular, it produces consistent, asymptotically optimal estimators under very mild conditions (Hall 1983, 1985; Stone 1984). However, cross-validation is presently available only for global problems — that is, problems where the density is estimated throughout its support. In the present paper we introduce a local version of cross-validation, suitable for estimating a density at a particular point.

Our argument runs like this. Suppose we wish to estimate a density f at a point x_0, which we take without loss of generality to be the origin. Let $\epsilon > 0$, and construct a small region $\mathcal{S}_\epsilon$ centred on $x_0 = 0$. If we were working in $p = 1$ dimension we might take $\mathcal{S}_\epsilon$ to be the interval $(-\epsilon, \epsilon)$; in p dimensions it might be a sphere of radius ϵ or the cube $(-\epsilon, \epsilon)^p$. When f is to be estimated globally, the cross-validatory criterion CV is constructed so that $CV + \int f^2$ is an unbiased estimator of global mean integrated square error. By analogy, in our local problem we construct $CV = CV_\epsilon$ so that $CV + \int_{\mathcal{S}_\epsilon} f^2$ is unbiased for mean squared error integrated over the region $\mathcal{S}_\epsilon$. Arguing as in Rudemo (1982), Hall (1983), Bowman (1984) or Stone (1984), this prescription produces the criterion

$$CV = \int_{\mathcal{S}_\epsilon} \hat{f}^2 - 2n^{-1} \sum_{j=1}^{n} I(X_j \in \mathcal{S}_\epsilon) \hat{f}_j(X_j) \,,$$

where n denotes sample size, $\hat{f}$ is the density estimator whose smoothing parameter we wish to select, $\hat{f}_j$ is the version of $\hat{f}$ computed on omitting the j'th observation X_j from the sample, and $I(X_j \in \mathcal{S}_\epsilon)$ denotes the random variable which equals 1 if $X_j \in \mathcal{S}_\epsilon$ and 0 otherwise. Our idea is that if the set $\mathcal{S}_\epsilon$ is allowed to shrink towards the origin at a suitable rate as sample size increases, then minimization of CV with respect to the smoothing parameter should be asymptotically equivalent to minimizing mean squared error of $\hat{f}(0)$.

In Section 2 we verify this claim theoretically, under very general conditions. There we take $\mathcal{S}_\epsilon$ to be the scaled-down set $\epsilon\mathcal{S} = \{\epsilon x : x \in \mathcal{S}\}$, where $\mathcal{S}$ is a set such as a sphere or a cube centred at the origin. We treat only the case of kernel estimators, although other types such as histogram estimators may be handled similarly.

Our theoretical results provide concise sufficient conditions on the rate at which ϵ is permitted to decrease to zero, for asymptotic optimality to obtain. Numerical work in Section 3 explains the effect which selection of ϵ has on the cross-validatory window and on the final density estimate, for "finite samples". It turns out that the influence is relatively minor, much less than the effect which varying window size in the same manner has on the final density estimate.

2. Methodology

We begin with notation. Given a random sample $X_1, \ldots, X_n$ from a distribution with p-variate density f, construct kernel estimators

$$\hat{f}(x \mid h) \equiv (nv_h)^{-1} \sum_{i=1}^{n} K\{(x - X_i)/h\} , \quad \hat{f}_j(x \mid h) \equiv \{(n-1)v_h\}^{-1} \sum_{i \neq j} K\{(x - X_i)/h\} ,$$

where K is a kernel function (a p-variate function integrating to unity), $h = (h_1, \ldots, h_p)$ is the "window" (a vector of positive numbers), $v_h \equiv \prod_k h_k$, and for any p-vector $x = (x_1, \ldots, x_p)$, $x/h \equiv (x_1/h_1, \ldots, x_p/h_p)$ and $xh \equiv (x_1 h_1, \ldots, x_p h_p)$. Of course, $\hat{f}_j$ is the version of $\hat{f}$ when X_j is omitted from the sample, and is used to define the cross-validation criterion. Our aim is to select the window so as to estimate f at, or in the vicinity of, $x = 0$.

Let $\mathcal{J}$ denote the unit radius p-dimensional open sphere centred at $x = 0$, and let $\mathcal{S}$ be any other subset of $\mathbb{R}^p$ satisfying $r_1 \mathcal{J} \subseteq \mathcal{S} \subseteq r_2 \mathcal{J}$ for some $0 < r_1 < r_2 < \infty$, where $r\mathcal{J} \equiv \{rx : x \in \mathcal{J}\}$. In this sense, $\mathcal{S}$ is "sphere-like". We wish to choose h so as to minimize integrated square error over $\epsilon\mathcal{S}$, where $\epsilon = \epsilon(n) > 0$, is bounded, and may decrease to zero as $n \to \infty$. Define

$$ISE \equiv \int_{\epsilon\mathcal{S}} (\hat{f} - f)^2 , \quad MISE \equiv \int_{\epsilon\mathcal{S}} E(\hat{f} - f)^2 \qquad \text{and}$$

$$CV \equiv \int_{\epsilon\mathcal{S}} \hat{f}^2 - 2n^{-1} \sum_{j=1}^{n} I(X_j \in \epsilon\mathcal{S}) \hat{f}_j(X_j)$$

to be integrated square error, mean integrated square error and the cross-validation criterion, respectively. Since $MISE = E(ISE) = E(CV) + \int f^2$ then it is to be hoped that the operations of minimizing ISE, MISE and CV will produce very similar windows. We shall prove that this is indeed the case.

Assume that h is selected from a set $\mathcal{H}_n$ of values, containing no more than $O(n^c)$ elements for some arbitrary but fixed $c > 0$, and satisfying

$$C_1 n^{-1+\eta} \leq v_h \quad \text{and} \quad \max h_i \leq C_2 n^{-\eta} , \quad \text{all } h \in \mathcal{H}_n \text{ and all } n \geq 1 , \tag{2.1}$$

for some $\eta, C_1, C_2 > 0$. To ensure that ϵ does not converge to zero too rapidly we ask that

$$\epsilon^{2p} \geq C v_h^{1-\eta} \,, \quad \text{all } h \in \mathcal{H}_n \text{ and all } n \geq 1 \,, \tag{2.2}$$

for some $\eta, C > 0$. We suppose that the kernel K is bounded and compactly supported, that f is bounded away from zero in a neighbourhood of the origin, and that f and its one-dimensional marginals are bounded. Finally, defining $\Delta \equiv \int_{\epsilon S}(E\hat{f} - f)^2$, we ask that for ϵ sufficiently small

$$C_1 \Delta \leq \epsilon^p \inf_{x \in \epsilon S} |E\hat{f}(x) - f(x)|^2 \leq \epsilon^p \sup_{x \in \epsilon S} |E\hat{f}(x) - f(x)|^2 \leq C_2 \Delta \tag{2.3}$$

for all $h \in \mathcal{H}_n$, all $n \geq 1$, and some $C_1, C_2 > 0$. Then we have the following theorem.

Theorem. *Under the conditions above,*

$$CV = ISE - \int_{\epsilon S} f^2 - 2n^{-1} \sum_{i=1}^{n} f(X_i) I(X_i \in \epsilon S) + R_1 \quad \text{and} \quad ISE = MISE + R_2 \,,$$

where

$$\max_{h \in \mathcal{H}_n} (|R_1| + |R_2|)/MISE \to 0 \tag{2.4}$$

almost surely as $n \to \infty$.

Remarks. (We assume η has been chosen so small that the window h which minimizies MISE satisfies (2.1) for all sufficiently large n.)

(i) In the case of a Hölder-continuous kernel, result (2.4) is readily extended to the case where $\mathcal{H}_n$ denotes the set of all values h satisfying (2.1). This extension is readily obtained in other cases, such as that where K is the uniform kernel, by using special properties of the kernel. It is convenient to assume, as we do below, that the extension is possible.

(ii) Let $\hat{h}$ denote that window which minimizes CV over $\mathcal{H}_n$. It follows from our theorem that

$$ISE(\hat{h}) / \inf_{h \in \mathcal{H}_n} ISE \to 1 \quad \text{and} \quad ISE(\hat{h}) / \inf_{h \in \mathcal{H}_n} MISE \to 1$$

almost surely, so that $\hat{h}$ provides asymptotic minimization of both ISE and MISE.

(iii) When estimating f at the origin we would want ϵ to shrink to zero as $n \to \infty$, sufficiently slowly for (2.2) to hold. We claim that in this circumstance, the window which minimizes CV is asymptotic to the window which minimizes $\delta(h) \equiv E\{\hat{f}(0 \mid h) - f(0)\}^2$. To appreciate why, observe that in regular cases, $MISE(h) \sim \epsilon^p \|\mathcal{S}\| \delta(h)$ where $\|\mathcal{S}\|$ denotes the p-dimensional content of $\mathcal{S}$. We may now deduce from our theorem that if $\hat{h}$ is the window which minimizes CV then

$$MISE(\hat{h}) \sim \inf_{h \in \mathcal{H}_n} MISE(h) \quad \text{almost surely} ,$$

or equivalently

$$\delta(\hat{h}) \sim \inf_{h \in \mathcal{H}_n} \delta(h) \quad \text{almost surely} , \tag{2.5}$$

which (again in regular cases) means that if h_n minimizes $\delta(h)$ then $\hat{h}/h_n \to 1$ almost surely.

In the case of $p = 1$ dimension, an example of the "regular cases" which we have in mind is that where f has r bounded and continuous derivatives, $f(0) \neq 0 \neq f^{(r)}(0)$, and K is an r'th order kernel:

$$\int K(z)\, dz = 1 \quad \text{and} \quad \int z^j K(z)\, dz = 0 \quad \text{for } 1 \le j \le r-1 .$$

The assumption $f^{(r)}(0) \neq 0$ ensures condition (2.3), and continuity of $f^{(r)}$ implies first that (2.5) holds and thence that $\hat{h}/h_n \to 1$ almost surely. In this circumstance, the window which minimizes $\delta(h)$ is of size $n^{-1/(2r+1)}$, and condition (2.2) governing the size of $\epsilon(n)$ asks that ϵ shrink to zero more slowly than $n^{-1/\{2(2r+1)\}}$.

3. Numerical results

The cross-validatory criterion CV is particularly easy to compute in the univariate case when K is the uniform kernel, $K(z) = \frac{1}{2}$ for $|z| \le 1$ and $K(z) = 0$ otherwise. There, if $\mathcal{S}_\epsilon \equiv (-\epsilon, \epsilon)$ and f is to be estimated at the origin,

$$CV = (2nh)^{-2} \sum_i \sum_j \max\{0, \min(\epsilon, X_i + h, X_j + h) + \min(\epsilon, -X_i + h, -X_j + h)\}$$
$$- \{n(n-1)h\}^{-1} \sum\sum_{i \neq j} I(|X_i - X_j| \leq h\,,\ -\epsilon \leq X_j \leq \epsilon)\,. \tag{3.1}$$

Thus, only counting methods are needed for calculation. For a general nonnegative symmetric kernel, and provided the origin is not a point of inflexion of f, the window which minimizes $E\{\hat{f}(0) - f(0)\}^2$ is asymptotic to

$$h_0 \equiv n^{-1/5}(\kappa_2/\kappa_1^2)^{1/5}\{f(0)/f''(0)^2\}^{1/5}\ , \tag{3.2}$$

where $\kappa_1 \equiv \int z^2 K(z)\,dz$ and $\kappa_2 \equiv \int K(z)^2\,dz$. In the special case of the uniform kernel, the ratio κ_2/κ_1^2 equals 9/2. Thus, for particularly simple calculations in the case of a general nonnegative symmetric kernel K, $\hat{h}$ may be chosen to minimize CV defined at (3.1), and $\tilde{h}$, an asymptotically optimal window for the given kernel K, be taken equal to $(2\kappa_2)^{1/5}(9\kappa_1^2)^{-1/5}\hat{h}$. It should be stressed that our use of the uniform is a device for reducing the heavy computational burden that would be required for calculation of local bandwidths at a sequence of X values

We applied formula (3.1), with various values of ϵ, to numerous samples of size $n = 50$ from the Standard Normal distribution. There, when K is the uniform kernel, the "optimal" window h_0 defined by (3.2) is $h_0 = .74$. Figure 1 illustrates the behavior of CV as a function of $\log_{10}(h)$ for our first six data sets with $\epsilon = 1.0$.

Figure 1 near here

These are fairly typical of the difficulties posed by the better-known global squared-error cross validation. As usual, the possibility exists that a spurious, extremely small value of h may be indicated.

The cross-validatory score function would be less noisy if one were to employ the triangular kernel. This or even Epanechnikov's quadratic kernel would increase the computation somewhat and yet still permit a reasonably manageable formula along the lines of (3.1). To illustrate the potential effect of such an approach we overlay the results of a five-point moving average smoother on the first realization. There is no change in location of the minimum. The values of $\hat{h}$ that minimize CV in (3.1) are presented in Table 1 along with the corresponding values of $\hat{f}(0;\hat{h})$.

Table 1 near here

The selection of a value for ϵ is not extremely critical. A reasonable value may be based upon a "pilot" estimate for h. For a Standard Normal kernel Silverman (1986) recommends $h = .9An^{-1/5}$, where $A = \min\{s, \hat{\sigma}\}$, s is the sample standard deviation and $\hat{\sigma}$ = interquartile range/1.34. Converting this coefficient to pertain to the uniform kernel and using $n^{1/10}$, we obtain a candidate value for ϵ that adequately reflects scale and sample size. Furthermore, to illustrate the relative insensitivity to the choice of ϵ the second of our samples (for which $\hat{h} = .77$) was reanalyzed for numerous alternative values of ϵ ranging from 0.8 to 3.0. The value of $\hat{h}$ was the same for all of these. At $\epsilon = 0.75$ the minimum of (3.1) switched locations to $h = .12$.

With an increased sample size of $n = 100$, which is still not large for density estimation, we estimated f at x_0 = 0.0, 0.5 and 1.0. (The latter choice violates (2.3); see following paragraph. However, it is just this feature that makes $x_0 = 1.0$ interesting.) By translating the data so that x_0 becomes the origin, formula (3.1) may be used to obtain the corresponding adaptive windows. A typical set of results for the Standard Normal again with $\epsilon = 1.0$ is summarized in Table 2. These examples are not offered in support of a recommendation of our implementation. We only wish to show that the approach is feasible and yields different bandwidths at different x values. Refinements are in order and the analyst is well advised to examine plots of the cross-validatory score function.

Table 2 near here

It may be shown as in Hall and Marron (1987a, p.572) that if an r'th order kernel is used in a global setting, then the relative error of the cross-validatory window $\hat{h}$ is of size $n^{-1/\{2(2r+1)\}}$. (Our Remark (iii) at the end of Section 2 defines an r'th order kernel.) Therefore numerical results can be quite erratic if r is large. This is a feature of the problem, not of the method of cross-validation, as has been shown by Hall and Marron (1987b). Exactly the same behaviour is observed when estimating the density at a point of inflexion, such as $x_0 = \pm 1$ in the case of our Standard Normal example. There, the estimation problem has all the features of using a fourth order kernel, since terms in h, h^2 and h^3 vanish from the bias. In consequence, relative error increases from order $n^{-1/10}$ at points $x_0 \neq \pm 1$, to $n^{-1/18}$ at $x_0 = \pm 1$. Numerical studies do reveal a considerable increase in the fluctuation of $\hat{h}$ when estimating at ± 1.

Acknowledgements

We are grateful to a referee for a thoughtful and helpful review. This research was partially supported by ONR Contract N00014-85-K-0340 and completed while the second author was on leave at The Institute for Advanced Studies, Australian National University.

Appendix: Proof of Theorem

Put $U(x,y) \equiv v_h^{-1} K\{(y-x)/h\} I(y \in \epsilon\mathcal{S}) - \int_{\epsilon\mathcal{S}} f E(\hat{f})$,

$$V(x,y) \equiv \int_{\epsilon\mathcal{S}} \left[v_h^{-2} K\{(u-x)/h\} K\{(u-y)/h\} - \{E\hat{f}(u)\}^2\right] du ,$$

$U_1(x) \equiv E\{U(x,X)\}$, $U_2(x) \equiv E\{U(X,x)\}$, $V_1(x) \equiv E\{V(x,X)\} = E\{V(X,x)\}$, $A(x,y) \equiv U(x,y) - U_1(x) - U_2(y)$ and $B(x,y) \equiv V(x,y) - V_1(x) - V_1(y)$. Then

$$\tfrac{1}{2}\left(ISE - CV - \textstyle\int_{\epsilon\mathcal{S}} f^2\right) = \{n(n-1)\}^{-1} \sum\sum_{i \neq j} A(X_i, X_j) + n^{-1} \textstyle\sum_i U_2(X_i) , \qquad (A.1)$$

$$\begin{aligned} ISE - MISE = n^{-2} \sum\sum_{i \neq j} B(X_i, X_j) + 2n^{-1} \sum_i \{V_1(X_i) - U_1(X_i)\} - 2n^{-2} \sum_i V_1(X_i) \\ + n^{-2} \sum_i \{V(X_i, X_i) - EV(X_1, X_1)\} . \qquad (A.2) \end{aligned}$$

Put $u(x) \equiv U_2(x) - \left\{f(x) I(x \in \epsilon\mathcal{S}) - \int_{\epsilon\mathcal{S}} f^2\right\}$. By (A.1) and (A.2) we have for each integer $\nu \geq 1$ and each $t > 0$, using Markov's inequality,

$$\pi_1(h) \equiv P\left[\left|CV - ISE + 2n^{-1} \sum_i f(X_i) I(X_i \in \epsilon\mathcal{S}) - \int_{\epsilon\mathcal{S}} f^2\right| > t\{\epsilon^p (nv_h)^{-1} + \Delta\}\right]$$

$$\le C(\nu)t^{-2\nu}\{\epsilon^p(nv_h)^{-1}+\Delta\}^{-2\nu}\Big\{E\Big|n^{-2}\sum\sum_{i\neq j}A(X_i,X_j)\Big|^{2\nu} + E\Big|n^{-1}\sum_i u(X_i)\Big|^{2\nu}\Big\}\,, \tag{A.3}$$

$$\pi_2(h)\equiv P\big[|ISE-MISE|>t\{\epsilon^p(nv_h)^{-1}+\Delta\}\big]\le C(\nu)t^{-\nu}\{\epsilon^p(nv_h)^{-1}+\Delta\}^{-2\nu}$$
$$\times\Big[E\Big|n^{-2}\sum\sum_{i\neq j}B(X_i,X_j)\Big|^{2\nu}+E\Big|n^{-1}\sum_i\{V_1(X_i)-U_1(X_i)\}\Big|^{2\nu} + E\Big|n^{-2}\sum_i V_1(X_i)\Big|^{2\nu}\Big]\,. \tag{A.4}$$

Let D denote either A or B, and observe that $E\{D(X_i,X_j)\mid X_i\}=E\{D(X_i,X_j)\mid X_j\}=0$. Therefore, with $Y_i\equiv\sum_{j\le i-1}\{D(X_i,X_j)+D(X_j,X_i)\}$ we have $E(Y_i\mid X_1,\ldots,X_{i-1})=0$, implying that the Y_i's are martingale differences. It now follows by Burkholder's and Rosenthal's inequalities (Hall and Heyde 1980, p.87) that for any $\nu\ge 1$,

$$E\Big|\sum\sum_{i\neq j}D(X_i,X_j)\Big|^{2\nu}=E\Big|\sum_{i=2}^{n}Y_i\Big|^{2\nu}\le C(\nu)n^{\nu-1}\sum_{i=2}^{n}E|Y_i|^{2\nu}\,.$$

Conditional on X_i, Y_i is a sum of $i-1$ independent and identically distributed random variables, whence it follows by Rosenthal's inequality (Hall and Heyde 1980, p.23) that

$$E(|Y_i|^{2\nu}\mid X_i)\le C\big(n^{\nu}\big[E\{D^2(X_i,X_1)\mid X_i\}\big]^{\nu}+n^{\nu}\big[E\{D^2(X_1,X_i)\mid X_i\}\big]^{\nu} + nE\{|D(X_i,X_1)|^{2\nu}+|D(X_1,X_i)|^{2\nu}\mid X_i\}\big)$$

for $2\le i\le n$. A little algebra shows that $E\{D^2(x,X)+D^2(X,x)\}\le Cv_h^{-1}$ uniformly in x, and $E\{|D(X_1,X_2)|^{2\nu}\}\le Cv_h^{-2\nu+1}$. Therefore

$$E\Big|n^{-2}\sum\sum_{i\neq j}D(X_i,X_j)\Big|^{2\nu}\le C(n^{-2\nu}v_h^{-\nu}+n^{-3\nu+1}v_h^{-2\nu+1})$$
$$=C\{\epsilon^p(nv_h)^{-1}\}^{2\nu}\{(v_h/\epsilon^{2p})^{\nu}+(n\epsilon^{2p})^{-\nu}nv_h\}\,.$$

Since $C_1n^{-1+\eta}\le v_h\le C_2n^{-\eta}$, and $\epsilon^{2p}\ge Cv_h^{1-\eta}$, then for some $\zeta>0$ and all $\nu\ge 1$,

$$E\Big|n^{-2}\sum\sum_{i\neq j}D(X_i,X_j)\Big|^{2\nu}\le C(\nu)\{n^{-\zeta}\epsilon^p(nv_h)^{-1}\}^{2\nu}\,. \tag{A.5}$$

Note that $E\{u(X)\}=0$ and $|u(x)| \le C(\Delta/\epsilon^p)^{\frac{1}{2}}$ uniformly in x. Therefore by Rosenthal's inequality, and for any $\delta>0$,

$$\begin{aligned} E\left|n^{-1}\sum_i u(X_i)\right|^{2\nu} &\le C_1\{(\Delta/n\epsilon^p)^\nu + n^{1-2\nu}(\Delta/\epsilon^p)^\nu\} \le 2C_1(\Delta/n\epsilon^p)^\nu \\ &\le 4C_1\left[\{n^{-\delta}\epsilon^p(nv_h)^{-1}\}^{2\nu} + (n^\delta v_h\Delta/\epsilon^{2p})^{2\nu}\right] \\ &\le C_2\left[n^{-\zeta}\{\epsilon^p(nv_h)^{-1}+\Delta\}\right]^{2\nu}\,, \end{aligned} \tag{A.6}$$

on choosing $\delta>0$ and $\zeta>0$ small. An identical argument, noting that $E\{U_1(X)-V_1(X)\}=0$ and $|U_1(x)-V_1(x)| \le C(\Delta/\epsilon^p)^{\frac{1}{2}}$ gives

$$E\left|n^{-1}\sum_i\{U_1(X_i)-V_1(X_i)\}\right|^{2\nu} \le C\left[n^{-\zeta}\{\epsilon^p(nv_h)^{-1}+\Delta\}\right]^{2\nu}. \tag{A.7}$$

And since $E\{V_1(X)\}=0,\ |V_1(x)|\le C$ and $|V(x,x)| \le Cv_h^{-1}$ then

$$E\left|n^{-2}\sum_i V_1(X_i)\right|^{2\nu} \le C_1 n^{-3\nu} \le C_2\{n^{-\frac{1}{2}}\epsilon^p(nv_h)^{-1}\}^{2\nu}\,, \tag{A.8}$$

$$\begin{aligned} E\left|n^{-2}\sum_i\{V(X_i,X_i)-EV(X_1,X_1)\}\right|^{2\nu} &\le Cn^{-3\nu}v_h^{-2\nu} \\ &= C\{(n^{\frac{1}{2}}\epsilon^p)^{-1}\epsilon^p(nv_h)^{-1}\}^{2\nu}\,. \end{aligned} \tag{A.9}$$

Formulae (A.5)–(A.9) provide bounds for the right-hand sides of (A.3) and (A.4), from which we may now deduce that for each $c>0$,

$$\max_{h\in\mathcal{H}_n}\{\pi_1(h)+\pi_2(h)\} = O(n^{-c})\,.$$

Since $\#\mathcal{H}_n = O(n^d)$ for some $d>0$ then by the Borel-Cantelli lemma,

$$\begin{aligned} \max_{h\in\mathcal{H}_n}\left\{\left|CV-ISE+2n^{-1}\sum_i f(X_i)I(X_i\in\epsilon S)-\int_{\epsilon S} f^2\right| + |ISE-MISE|\right\} \\ \times\{\epsilon^p(nv_h)^{-1}+\Delta\}^{-1}\to 0 \end{aligned} \tag{A.10}$$

almost surely as $n\to\infty$. Finally, observe that

$$MISE = \int_{\epsilon S}\mathrm{var}\,(\hat f)+\Delta \ge C\epsilon^p(nv_h)^{-1}+\Delta$$

for some $C > 0$, and so the theorem follows from (A.10).

References

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

Hall, P. (1983). Large sample optimality of least squares cross-validation in density estimation. *Ann. Statist.* **11**, 1156–1174.

Hall, P. (1985). Asymptotic theory of minimum integrated square error for multivariate density estimation. In: *Multivariate Analysis VI*, Ed. P.R. Krishnaiah, pp.289–309.

Hall, P. and C.C. Heyde (1980). *Martingale Limit Theory and its Application.* (Wiley, New York).

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

Hall, P. and J.S. Marron, (1987b). On the amount of noise inherent in bandwidth selection for a kernel density estimator. *Ann. Statist.* **15**, 163–181.

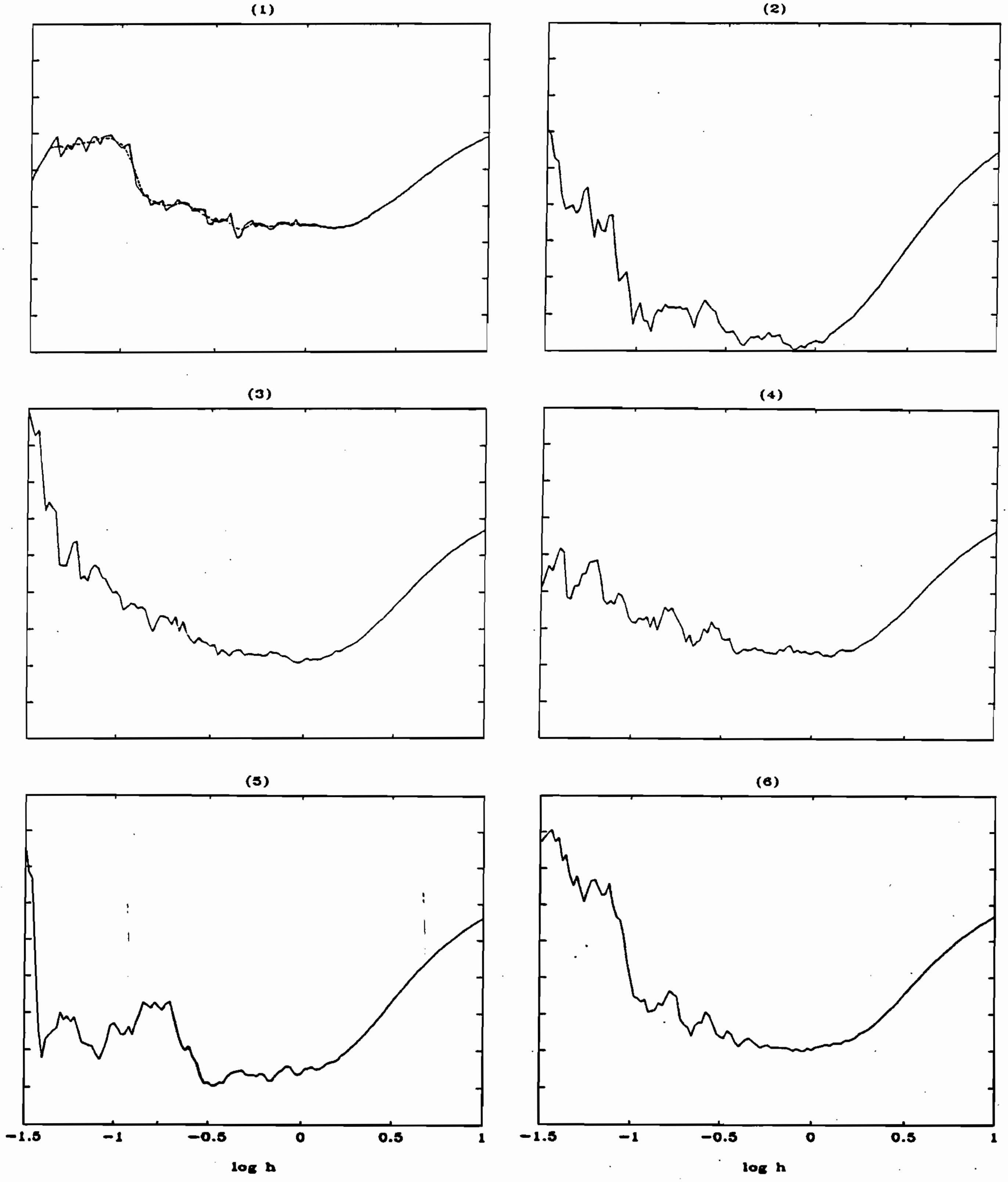
Rudemo, M. (1982). Empirical choice of histogram and kernel density estimators. *Scand. J. Statist.* **9**, 65–78.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis.* (Chapman and Hall, London).

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

Figure 1

Squared-error Cross-validation Criterion as a function of Window Width.


(1)
(2)
(3)
(4)
(5)
(6)
log h
log h