Local Minima in Cross-Validation Functions

Peter Hall, J. S. Marron

Local Minima in Cross-validation Functions

By PETER HALL† and J. S. MARRON

Brown University, Providence, USA, and Universität Bonn, FRG

[Received November 1988. Final revision December 1989]

SUMMARY
The method of least squares cross-validation for choosing the bandwidth of a kernel density estimator has been the object of considerable research, through both theoretical analysis and simulation studies. The method involves the minimization of a certain function of the bandwidth. One of the less attractive features of this method, which has been observed in simulation studies but has not previously been understood theoretically, is that rather often the cross-validation function has multiple local minima. The theoretical results of this paper provide an explanation and quantification of this empirical observation, through modelling the cross-validation function as a Gaussian stochastic process. Asymptotic analysis reveals that the degree of wigginess of the cross-validation function depends on the underlying density through a fairly simple functional, but dependence on the kernel function is much more complicated. A simulation study explores the extent to which the asymptotic analysis describes the actual situation. Our techniques may also be used to obtain other related results—e.g. to show that spurious local minima of the cross-validation function are more likely to occur at too small values of the bandwidth, rather than at too large values.

Keywords: BANDWIDTH SELECTION; CROSS-VALIDATION; KERNEL DENSITY ESTIMATION; LOCAL MINIMA; SMOOTHING PARAMETER SELECTION

1. INTRODUCTION

For the smooth estimation of a density \( f(x) \), using a random sample, \( X_1, \ldots, X_n \), from \( f \), the kernel estimator is given 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 often taken to be a symmetric probability density and \( h \) is called the bandwidth or smoothing parameter. See Silverman (1986) for motivation of this type of estimator, and also for discussion of how to use it for effective data analysis. The choice of \( h \) is crucial to the performance of the estimator.

Several means of using the data to yield an objective choice of \( h \) have been proposed; see Marron (1988) for a survey and Park and Marron (1989) for a deeper comparison of available methods. The most widely studied of these is called least squares cross-validation. This was developed by Rudemo (1982) and Bowman (1984), who proposed to choose the bandwidth which is the minimizer of the least squares cross-validation function \( CV(h) \), where

†Address for correspondence: Department of Statistics, Australian National University, GPO Box 4, Canberra, ACT 2601, Australia.
\[
CV(h) = \int \hat{f}_h^2 - 2n^{-1} \sum_{j=1}^{n} \hat{f}_{h,j}(X_j),
\]
\[
f_{h,j}(x) = (n-1)^{-1} \sum_{i \neq j} K_h(x - X_i).
\]

Minimization of the function \(CV(h)\) can be motivated by the fact that \(CV(h)\) provides an unbiased estimate of a vertical shift of the mean integrated squared error,
\[
MISE(h) = E \left[ (\hat{f}_h - f)^2 \right].
\]

See Diggle and Marron (1988) for a very different motivation.

Hall (1983, 1985), Stone (1984), Burman (1985), Marron (1987) and Hall and Marron (1987a, b) have investigated theoretical properties of least squares cross-validation. An unpleasant aspect of least squares cross-validation, which has been noticed in simulation studies and in applications to real data sets, but which does not show up in any of the existing theoretical analyses, is that the function \(CV(h)\) often has multiple local minima. See Fig. 1 for an example of several such curves. These curves are chosen from 100 simulated data sets of size 100 from the standard normal distribution, when the kernel is also standard normal. The logarithm of the bandwidth is used as abscissa because \(h\) affects the performance of the kernel estimator in terms of the scale of the data. A vertical shift of the curve \(MISE(h)\) is also overlayed for comparison.

In this paper we provide insight into the properties of \(f\) which cause the wiggliness of the cross-validation function. We show that, asymptotically, the expected number of local maxima is an increasing function of the scale invariant quantity
\[
\rho = \rho_f = \int f^2 / (\int f'^2)^{1/2},
\]
which is the ratio of two measures of the roughness of \(f\). This result is supported both by theoretical analysis, in Section 2, and a simulation study, in Section 3. The techniques in Section 2 also reveal other features of cross-validation—for example, they

---

Fig. 1. Shifted MISE(h) (----) and several simulated realizations of CV(h) (-----) for different samples of 100 \(N(0, 1)\) observations
show that spurious turning points of $\text{CV}(h)$ tend to occur on the low side of the optimal $h$, rather than on the high side.

2. THEORETICAL ANALYSIS

2.1. Introduction and Summary

Recall the definition of $\rho = \rho(f)$ at equation (1.1). Let $h_0 = h_0(f)$ be the minimizer of $\text{MISE}(h)$, let $0 < s_1 < s_2 < \infty$ be arbitrary constants and allow the sample size to increase. We argue that if $f_1$ and $f_2$ are two densities with $\rho(f_1) > \rho(f_2)$ then the expected number of local maxima of $\text{CV}(h)$ within interval $I(f) = (s_i h_0, s_2 h_0)$ (an index of the wiggliness of $\text{CV}(h)$) decreases to zero more slowly when $f = f_i$. Section 2.2 introduces an approximation to the expected number of local maxima, derived by considering downcrossings of a certain Gaussian process. Section 2.3 calculates key components of this approximation and draws the conclusions already described. Section 2.4 discusses generalizations.

2.2. Describing Local Maxima by Downcrossings.

Define

$$a(h) \equiv E\{\text{CV}(h)\} + \{f^2, \rho(h)\} \equiv E\{\text{CV}(h)\}.$$

The cross-validated bandwidth is a minimizer of the function $a(h) + \epsilon(h)$. Each local maximum of $a(h) + \epsilon(h)$ occurs at a downcrossing $h^*$ of the stochastic process $a'(h) + \epsilon'(h)$, where the prime denotes differentiation. A downcrossing is a point $h^*$ at which $a'(h) + \epsilon'(h)$ changes, as $h$ increases, from positive to negative. It is mathematically convenient to focus on the number of downcrossings of $a' + \epsilon'$; if $a(h) + \epsilon(h)$ is non-decreasing as $h \downarrow 0$ and as $h \uparrow \infty$, then the number of downcrossings is one less than the number of local minima of $a + \epsilon$.

Put

$$m(t) = n^{3/5} a'(n^{-1/5}), \quad X(t) \equiv n^{7/10} \epsilon'(n^{-1/5}t), \quad \lambda \equiv n^{1/10}.$$

The process $n^{3/5}(a' + \epsilon')$ has exactly the same upcrossing structure as $a' + \epsilon'$, and so we may concentrate on the former process. Define

$$X_\lambda(t) \equiv n^{3/5} a'(n^{-1/5}t) + n^{3/5} \epsilon'(n^{-1/5}t) = m(t) + \lambda^{-1} X(t),$$

where $m$ is a non-random function and $X$ is a stochastic process with zero mean. It may be shown that, as $n \to \infty$, $X$ converges weakly to a certain Gaussian process on $(0, \infty)$; see Appendix A.

Assume that $X$ is exactly Gaussian with zero mean and covariance function $r(t, u) \equiv E\{X(t)X(u)\}$. Then the expected number of downcrossings of $X_\lambda$ within interval $(t_1, t_2)$ equals

$$d_\lambda \equiv \int_{t_1}^{t_2} \gamma \sigma^{-1}(1 - \mu^2)^{1/2} \phi(\lambda m/\sigma)[\phi(\lambda \eta) - \lambda \eta \{1 - \Phi(\lambda \eta)\}] \, dt,$$  

(2.1)

where $\sigma, \gamma, \mu, \eta$ and $m$ are functions of $t$, $\sigma(t)^2 \equiv r(t, t)$,
\[\alpha(t) = \frac{\partial r(t, u)}{\partial u} \bigg|_{u=t}, \quad \gamma(t)^2 = \frac{\partial^2 r(t, u)}{\partial t \partial u} \bigg|_{u=t}, \quad \mu(t) = \frac{\alpha(t)}{\sigma(t)} \gamma(t), \] (2.2)

\[\eta(t) = \left\{ m'(t) - \gamma(t) \mu(t) \sigma(t)^{-1}\right\} / \left\{ 1 - \mu(t)^2 \right\}^{1/2} \gamma(t) \]

and \(\phi\) and \(\Phi\) denote the standard normal density and distribution functions respectively. See Cramér and Leadbetter (1967).

The functions \(\sigma, \gamma, \mu, \eta\) and \(m\) all have proper limits as \(n \to \infty\). In this abbreviated account of asymptotic theory we shall not distinguish between these functions and their limits. Given \(0 < t_1 < t_2 < \infty\), put

\[T = \min_{t_i < t < t_2} \left\{ m(t)^2 \sigma(t)^{-2} + \eta(t)^2 \right\} I\{\eta(t) > 0\}. \] (2.3)

We claim that the expected number of downcrossings of \(X_\lambda\) within \((t_1, t_2)\) decreases to zero more rapidly for processes \(X\) with larger values of \(T\). To appreciate why, observe first that \(\psi(x) \equiv \phi(x) - x \left\{ 1 - \Phi(x) \right\} > 0\) for all \(x\), \(\psi(x) \sim x^{-2} \phi(x)\) as \(x \to \infty\), \(\psi(x) \sim |x|\) as \(x \to -\infty\). We may deduce from formula (2.1) that, for any \(\delta > 0\), \(d_\lambda\) decreases to zero at a rate governed by the maximum size of

\[\phi(\lambda m/\sigma)[\phi(\lambda \eta) - \lambda \eta \left\{ 1 - \Phi(\lambda \eta) \right\}],\]

which equals \(\exp[-\left\{ \frac{1}{2} + o(1) \right\} \lambda^2 T]\). From this follows the statement in italics.

### 2.3. Calculations Based on Covariance of \(CV(h)\)

We begin by stating asymptotic formulae for functions appearing in expression (2.3). For \(m\),

\[m(t) = t^3 k_1 \int f^{"} - t^{-2} k_2,\]

where \(k_j\) denotes a positive constant depending only on \(K\), not on \(f\). (This expression is easily derived via the classical formula for the mean integrated squared error of a non-parametric density estimator.) The asymptotic covariance function of \(X, r(t_1, t_2)\), is given in Appendix A. From that we may deduce after some tedious algebra that

\[\sigma(t)^2 = t^{-3} k_3 \int f^2, \quad \gamma(t)^2 = t^{-5} k_4 \int f^2, \quad \mu(t) = \sigma'(t)/\gamma(t) = k_5.\]

It now follows from expressions (2.2) that

\[\eta(t) = k_6 \left\{ 9 t^2 k_1 \int f^{"2} + k_2,\right\}

and so by expression (2.3)

\[T = \min_{t_i < t < t_2} \left\{ t^{-1}(\int f^2)^{-1} \left\{ k_7(t^5 k_1 \int f^{"2} - k_2)^2 + k_8(9 t^2 k_1 \int f^{"2} + k_2)^2 \right\}\right\}.

Take \(t_0 = (k_2/k_1)^{1/5}\), which happens to be the solution of \(m(t) = 0\). Changing variable from \(t\) to \(s \equiv t/t_0\) and writing \(s_t \equiv t_1/t_0\), we deduce that

\[T = \rho^{-1} \min_{s_t < s < s_2} \left\{ s^{-1} \left\{ k_9(s^5 - 1)^2 + k_10(9 s^5 + 1)^2 \right\}\right\},\]

where \(\rho \equiv (\int f^2/\left\{ \int f^{"2}\right\})^{1/5}\). In particular, \(T\) is inversely proportional to \(\rho\), the constant of proportionality depending only on \(K, s_1\) and \(s_2\). It follows that densities \(f\) with larger values of \(\rho\) exhibit cross-validation functions with more propensity to form local maxima, because this propensity decreases with increasing \(T\).
This representation for $T$ also allows us to make some statements concerning the location of the local minima of $CV(h)$ with respect to $h_0$. In particular, for $k_9, k_{10} > 0$, the minimizer over $s > 0$ of $s^{-1} \{ k_9(s^5 - 1)^2 + k_{10}(9s^5 + 1)^2 \}$ can be shown by straightforward calculus to be less than unity. (The important part of the derivative is a quadratic in $s^5$, whose roots are thus easy to bound.) Applying this result for various subintervals $(s_1, s_2)$ indicates that the most likely place to find a local maximum in $CV(h)$ is for $h$ values smaller than $h_0$. Hence when two local minima appear in $CV(h)$ the minimum located nearer to the origin will usually correspond to an unreasonably small amount of smoothing. This provides a theoretical basis for some observations of Scott and Terrell (1987), and also for the empirically derived suggestion of Park and Marron (1989) that the largest local minimum be selected when $CV(h)$ has multiple local minima.

2.4. Similar Results in Other Circumstances

2.4.1. Higher order kernels

The previous argument may be pursued almost without change for $r$th-order kernels, which satisfy

$$\int z^j K(z) \, dz = \begin{cases} 1 & \text{if } j = 0, \\ 0 & \text{if } 1 \leq j \leq r - 1, \\ k \neq 0 & \text{if } j = r. \end{cases}$$

The only alteration of intrinsic importance is to replace $\rho$ by

$$\rho_r \equiv \{ \int f^{2r} \}^{1/(2r + 1)} / \{ \int f^{r+1} \},$$

which is a scale invariant ratio of zeroth order and $r$th-order descriptions of the roughness of $f$. If $CV(h)$ is computed for an $r$th-order kernel then the propensity of $CV(h)$ to have multiple local minima is an increasing function of $\rho_r$.

2.4.2. Local minima of integrated squared error

A similar argument may be developed for the integrated squared error

$$ISE(h) = \int (f_h - f)^2.$$

In this case, the main change necessary in Section 2.2 is to replace $\epsilon(h)$ by $\delta(h)$ $ISE(h) - MISE(h)$. The new (asymptotic) covariance function is $r(t, u) = r_1(t, u) + r_2(t, u)$, where $r_1$ and $r_2$ are defined at the end of Appendix A. If the theory in Section 2.3 is pursued in this context then it will be found that $T$ is a decreasing function of $\rho$, and also a function of

$$\rho^* \equiv \{ \int f''^2 f - (\int f'f)^2 \} / \{ \int f^2 \} / f''.$$

Although $T$ does not depend on $f$ except through $\rho$ and $\rho^*$, the additional quantity $\rho^*$ does make it difficult to give a simple description of the properties of $f$ which influence wiggleness of $ISE(h)$.

There is also a high order kernel analogue of these results.
3. SIMULATIONS

We took eight different normal mixture distributions and ranked them in order of decreasing $\rho(f)$, defined at equation (1.1):

(1) $\frac{1}{2} N(-\frac{1}{4}, \frac{4}{9}) + \frac{1}{2} N(\frac{1}{2}, \frac{4}{9})$,

(2) $N(0, 1)$,

(3) $0.81 N(0, 1) + 0.19 N(\frac{3}{2}, \frac{1}{4})$,

(4) $\frac{1}{2} N(0, 1) + \frac{1}{2} N(0, 0.224)$,

(5) $\frac{1}{2} N(-1, \frac{4}{9}) + \frac{1}{2} N(1, \frac{4}{9})$,

(6) $\frac{1}{2} N(0, 1) + \frac{1}{2} N(0, 0.0696)$,

(7) $\frac{1}{4} N(0, 1) + \frac{1}{4} N(5, 1) + \frac{1}{4} N(10, 1)$

(8) $\frac{1}{4} N(0, 1) + \frac{1}{4} N(15, 1)$

(9) $0.7 N(0, 1) + 0.3 N(0, 0.01)$

For each we simulated 500 data sets of sizes 25, 100, 400 and 1600. To effect the calculations in a reasonable time, a binned implementation of the calculation of $CV(h)$ was used, as described in Scott and Terrell (1987). However, the discrete binning of Scott and Terrell was replaced by the linear binning ideas of Jones and Lotwick (1983), as described in Section 3.5 of Silverman (1986). In all cases 800 bins were used on the interval $(-3, 3)$ (and density (7) was rescaled appropriately). The grid of $h$ values that was considered consisted, in each case, of 20 values logarithmically spaced from $h_0/4$ to $4h_0$, where $h_0$ denotes the minimizer of the actual $MISE(h)$ (not an asymptotic representation).

Table 1 shows the average number of local maxima in each case. To give a good idea of variability, we computed 95% confidence intervals based on treating the observed number of local maxima ($N_k$, for $1 \leq k \leq 500$) as independent Poisson($\lambda$) variables and using the approximation

$$\frac{\bar{N} - \lambda}{\lambda - 500}^{1/2} \approx N(0, 1),$$

where $\bar{N} = (1/500)\sum N_k$ (Bickel and Doksum, 1977). This produces a 95% interval with end points $\bar{N} + t^2/2 \pm t(\bar{N} + t^2/4)^{1/2}$, where $t = 1.96/500^{1/2}$. The data in Table 1 are presented in the form $x \pm y$, where $x = \bar{N} + (t^2/2)$ and $y = t(\bar{N} + t^2/4)^{1/2}$.

The main point to be noted from Table 1 is that, for given $n$, the expected number of local maxima is an increasing function of $\rho$. Section 2 provides asymptotic theory explaining this empirical observation. The major exception to the trend occurs for $n = 25$, which is hardly surprising since $n = \infty$ cannot be expected to provide a good guide to this situation.

**Table 1**

95% confidence intervals for the expected number of local maxima in the cross-validation function, for the eight normal mixture distributions

<table>
<thead>
<tr>
<th>Distribution</th>
<th>$\rho$</th>
<th>$n = 25$</th>
<th>$n = 100$</th>
<th>$n = 400$</th>
<th>$n = 1600$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.394</td>
<td>0.115 (± 0.029)</td>
<td>0.070 (± 0.023)</td>
<td>0.128 (± 0.086)</td>
<td>0.111 (± 0.146)</td>
</tr>
<tr>
<td>2</td>
<td>0.385</td>
<td>0.128 (± 0.031)</td>
<td>0.114 (± 0.029)</td>
<td>0.072 (± 0.023)</td>
<td>0.072 (± 0.023)</td>
</tr>
<tr>
<td>3</td>
<td>0.329</td>
<td>0.178 (± 0.080)</td>
<td>0.050 (± 0.019)</td>
<td>0.020 (± 0.012)</td>
<td>0.008 (± 0.007)</td>
</tr>
<tr>
<td>4</td>
<td>0.329</td>
<td>0.061 (± 0.021)</td>
<td>0.050 (± 0.019)</td>
<td>0.018 (± 0.011)</td>
<td>0.008 (± 0.007)</td>
</tr>
<tr>
<td>5</td>
<td>0.251</td>
<td>0.055 (± 0.020)</td>
<td>0.045 (± 0.018)</td>
<td>0.024 (± 0.013)</td>
<td>0.020 (± 0.012)</td>
</tr>
<tr>
<td>6</td>
<td>0.251</td>
<td>0.013 (± 0.009)</td>
<td>0.005 (± 0.005)</td>
<td>0.004 (± 0.004)</td>
<td>0.004 (± 0.004)</td>
</tr>
<tr>
<td>7</td>
<td>0.125</td>
<td>0.312 (± 0.049)</td>
<td>0.004 (± 0.004)</td>
<td>0.004 (± 0.004)</td>
<td>0.004 (± 0.004)</td>
</tr>
<tr>
<td>8</td>
<td>0.125</td>
<td>0.205 (± 0.039)</td>
<td>0.004 (± 0.004)</td>
<td>0.004 (± 0.004)</td>
<td>0.004 (± 0.004)</td>
</tr>
</tbody>
</table>
Table I also shows that, for a given density, the expected number of local maxima is a decreasing function of \( n \). The methods used in Section 2 are readily applied to support this observation.

ACKNOWLEDGEMENTS

The comments of two referees and the Editor have proved most helpful.

The research of the first author was done while on leave from the Australian National University. The research of the second author was done while on leave from the University of North Carolina and was partially supported by National Science Foundation grant DMS-8701201 and Deutsche Forschungsgemeinschaft SFB 303.

APPENDIX A: LIMIT THEOREM FOR DERIVATIVE OF CV(h)

We first introduce the notation. Let \( K \) be a compactly supported, symmetric function on \( \mathbb{R} \) with Hölder-continuous derivative \( K' \) and satisfying \( \int K = 1, \kappa = \int z^2 K(z) \, dz \neq 0 \). (A function \( g \) is Hölder continuous if \( |g(x) - g(y)| \leq C|x - y|^\delta \) for some \( C, \delta > 0 \) and all \( x, y \).) Assume that \( f \) is bounded and twice differentiable, \( f' \) and \( f'' \) are bounded and integrable, and \( f'' \) is uniformly continuous.

Define \( L(z) = -z K'(z), \epsilon(h) = CV(h) - E\{CV(h)\}, X(t) = n^{7/10} \epsilon(n^{-1/5} t), \beta_1(w; t_1, t_2) = \int K(t_1 z + t_2^{-1} w)K(t_2 z) \, dz, \beta_2(w; t_1, t_2) = \int K(t_1 z + t_2^{-1} w) \, L(t_2 z) \, dz, \)

\[
r_1(t_1, t_2) = 8(\int f^2)(t_1, t_2)^{-1} \{ \beta_1(w; t_1, t_2) - \beta_2(w; t_1, t_2) \} \{ \beta_1(w; t_1, t_2) - \beta_2(w; t_2, t_1) \} \, dw, \]

\[
a(t_1, t_2) = \{ \{ K(t_1 w) - L(t_1 w) \} \{ K(t_2 w) - L(t_2 w) \} \} \, dw, \]

\[
b(t_1, t_2) = \{ \{ K(y_1) \{ K(y_2) - L(y_2) \} \{ K(t_1, t_2)^{-1}(y_1 - y_2) \} \} - L(t_1, t_2)^{-1}(y_1 - y_2) \} \, dy_1 \, dy_2, \]

\[
u(t_1, t_2) = 8(\int f^2)(t_1, t_2)^{-1} a(t_1, t_2) \quad \text{and} \quad \nu(t_1, t_2) = 4(\int f^2)(t_1, t_2)^{-1} b(t_1, t_2). \]

Theorem. For each \( 0 < a < b < \infty \), \( X(\cdot) \) converges weakly in the space of continuous functions on \([a, b]\) to a Gaussian process with zero mean and covariance

\[
r(t_1, t_2) = r_1(t_1, t_2) + u(t_1, t_2) - v(t_1, t_2) - v(t_2, t_1). \]

The proof is rather standard, although intricate, and so is not given.

REFERENCES


