Partitioned Cross-Validation

by

J. S. Marron

University of North Carolina, Chapel Hill

April 19, 1987

ABSTRACT

Partitioned cross-validation is proposed as a method for overcoming the large amounts of across sample variability to which ordinary cross-validation is subject. The price for cutting down on the sample noise is that a type of bias is introduced. A theory is presented for optimal trade-off of this variance and bias. Comparison with other bandwidth selection methods is given.

AMS 1980 Subject Classification: primary 62G05

Key words and phrases: bandwidth, cross-validation, nonparametric curve estimation, smoothing parameter.

Research partially supported by NSF Grant DMS-8400602.

1. INTRODUCTION

In all smoothing methods for nonparametric curve estimation, the crucial problem is choice of the smoothing parameter. This problem is addressed here in the particular setting of kernel density estimation, however the ideas apply to all other curve estimation settings, such as nonparametric regression and hazard function estimation, and other estimators such as orthogonal series, smoothing splines and B-splines.

In all of these settings, the smoothing parameter essentially controls the amount of local averaging that the given estimator performs. If the local averaging is carried out through too narrow a window, the resulting estimate will be using too few observations in each window and will tend to oscillate wildly. On the other hand, if the window width is too large, points from too far away will be entered into the average and features of the underlying curve will be smoothed away. These ideas can be quantified statistically by observing that as the window width increases, the variance of the estimator decreases, but the bias increases. See for example Section 2.4 of Silverman (1986) for a detailed discussion of this.

Cross-validation provides an attractive data-based method for choosing the smoothing parameter. However it has recently been shown to be subject to considerable sample noise. Section 2 proposes a modification of cross-validation which eliminates some of the noise inherent to it.

The modification involves splitting the sample into subsamples, calculating the cross-validation score for each and minimizing the average of these score functions. This averaging has the effect of cutting way back on the sample noise, and the effect is strongest for many subsamples. However, the resulting bandwidth needs to be rescaled because it is appropriate only for the size of the subsample. The rescaling is only asymptotically correct and introduces some error in finite sample situations, which is worst for many subsamples.

The question of how to choose the number of subsamples is treated by asymptotic analysis in Section 3. It is seen that having more subsamples decreases "variance", but increases "bias", in a spirit which is quite analogous to the smoothing problem. This is quantified in a manner which allows theoretical determination of an optimal number of subsamples.

Comparison with other bandwidth selection methods, such as the plug-in method is given in Section 4. Suggestions for further improvements are also given in that section.

2. THE METHOD

Given a random sample $X_1,\ldots,X_n$ from a probability density $f(x)$, the kernel density estimator is defined by

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

where h is the smoothing parameter, often called the "bandwidth", and

$$K_h(\cdot) = K(\cdot/h)/h,$$

where K is called the kernel function and is here taken to be a probability density. See the monographs by Prakasa Rao (1983), Devroye and Györfi (1984) and Silverman (1986) for an access to the literature on this estimator.

Choice of h is crucial to the performance of the estimator and is focused on here. Rudemo (1982) and Bowman (1984) proposed the method of least squares cross-validation for choosing h. This is motivated by noting that the Integrated Squared Error,

$$ISE(h) = \int[\hat{f}_h(x) - f(x)]^2 dx,$$

may be estimated, up to a constant independent of h, by the cross-validation score function,

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

where $\hat{f}_{j,h}$ denotes the leave one out estimator,

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

Hence one may hope that the bandwidth $\hat{h}_{CV}$, the minimizer of CV(h), reasonably approximates $\hat{h}_{ISE}$, the minimizer of ISE(h), and hence approximates h_{MISE}, the minimizer of the Mean Integrated Squared Error,

$$MISE(h) = E[ISE(h)].$$

A sense in which this is true is called asymptotic optimality. See Hall (1983) and Stone (1984) for the best known results of this type. One way to formulate this is that, under some regularity conditions, as $n \to \infty$,

$$(2.1) \qquad \hat{h}_{CV}/h_{MISE} \to 1,$$

in some mode of convergence. There are several other equivalent formulations, such as h_{MISE} may be replaced by h_{ISE} or the ratio of bandwidths may be replaced by a ratio of error criteria evaluated at the two bandwidths. While such results are comforting in that they imply cross-validation is tracking the correct thing in the limit, there still should be considerable concern over what it implies about real data situations.

One possible approach to this problem has been taken by Hall and Marron (1987a). In that paper it is seen that under standard assumptions (essentially that K is a probability density and f has two continuous derivatives), the rate of convergence in (2.1) is excruciatingly slow. In particular,

$$(2.2) \qquad (\hat{h}_{CV} - h_{MISE})/h_{MISE} \sim n^{-1/10}.$$

This result quantifies the fact that cross-validation is subject to a good deal of sample noise, and that noise goes down very slowly with the sample size.

At this point a natural question arises: is their some inefficiency in cross-validation, or this rate of $n^{-1/10}$ inherent to the problem? A partial answer was provided by Hall and Marron (1987b), who show that in a rather compelling minimax sense, the rate is the best possible. However these results implicitly assume that $f(x)$ has no more than two derivatives and leave open the possibility that the rate of convergence in (2.2) can be improved if more derivatives are assumed.

A method for using extra smoothness on f is the following. Suppose first that the integer m divides n (this is strictly for notational convenience, the generalization to arbitrary m is straightforward).

Randomly (independent of $X_1,\dots,X_n$) partition the integers $1,\dots,n$ into subsets $S_1,\dots,S_m$ each with cardinality n/m. For $j=1,\dots,m$, let $CV_j(h)$ and $ISE_j(h)$ denote the cross-validation score and the integrated squared error respectively for the sample, $\{X_i : i \in S_j\}$. Then define

$$CV^*(h) = m^{-1} \sum_{j=1}^{m} CV_j(h),$$

$$ISE^*(h) = m^{-1} \sum_{j=1}^{m} ISE_j(h),$$

and let $\hat{h}^*_{CV}$ denote the minimizer of $CV^*(h)$. A result of the type (2.2) can be used to show that the bandwidth $\hat{h}^*_{CV}$ should be reasonably close to h^*_{MISE}, the minimizer of the subsample Mean Integrated Square Error,

$$MISE^*(h) = E[ISE^*(h)] = E[ISE_1(h)].$$

Hence $\hat{h}^*_{CV}$ will be appropriate for a sample of size m/n.

A reasonable full sample bandwidth can then be obtained by using the fact that h_{MISE} satisfies the asymptotic relationship

$$h_{MISE} \sim C_0 n^{-1/5}, \tag{2.3}$$

where

$$C_0 = [\frac{\int K^2}{(\int x^2 K)^2 \int (f'')^2}]^{1/5},$$

when f has two uniformly continuous derivatives (see for example (3.21) of Silverman 1986). From this it follows that the appropriate rescaling of the subsample bandwidth is

$$\hat{h}_{PCV} = m^{-1/5} \hat{h}^*_{CV},$$

called the partitioned cross-validated bandwidth.

3. ANALYSIS

The issue of choice of the number of subsamples is addressed by the following asymptotic results.

Assume that:

(1) K is a twice continuously differentiable, compactly supported, symmetric probability density,

(2) f has an integrable, uniformly continuous second derivative,

Theorem 1: Under the above assumptions, as $n \to \infty$,

$$(n/m)^{3/10} m^{1/2} (\hat{h}^*_{CV} - h^*_{MISE}) \xrightarrow{d} N(0,\sigma^2), \tag{3.1}$$

where

$$\sigma^2 = 8(\int f^2)(\int[K*(K-L)-(K-L)]^2)/25(\int K^2)^{7/5}(\int x^2K)^{6/5}(\int(f'')^2)^{3/5}.$$

Note that theorem 1 quantifies the noise inherent to the partitioned cross-validated bandwidth because it implies that

$$(\hat{h}_{PCV} - m^{-1/5}h^*_{MISE})/m^{-1/5}h^*_{MISE} \sim n^{-1/10}m^{-1/5}.$$

Thus the variability of $\hat{h}_{PCV}$ across different samples may be made much smaller than that of $\hat{h}_{CV}$ (compare with (2.2)) by taking m large. Unfortunately, there is a limitation on how large m should be that is imposed by the fact that (2.3) is only an approximation, and hence $m^{-1/5}h^*_{MISE}$ is not exactly the same as h_{MISE}. Most of the work in understanding this issue is contained in

Theorem 2: If f has a uniformly continuous fourth derivative, then

$$h_{MISE} - C_0 n^{-1/5} = C_1 n^{-3/5} + o(n^{-3/5}),$$

where,

$$C_1 = -(\int K^2)^{3/5}(\int x^4 K)(\int (f''')^2)/20(\int x^2 K)^{11/5}(\int (f'')^2)^{8/5}.$$

It follows from theorem 2 that

$$(3.2) \qquad h_{MISE} - m^{-1/5} h^*_{MISE} = -C_1 n^{-3/5} m^{2/5} + o(n^{-3/5} m^{2/5}),$$

as $m, n \to \infty$ with $n/m \to \infty$. Note that while (3.1) quantifies the sample noise or "variance" of $\hat{h}_{PCV}$, (3.2) can be thought of as measuring a type of "bias". The problem of choice of m is analogous to the smoothing problem of choosing h as discussed in section 1. If m is large there is too much bias, while if m is small there is too much variance.

This trade off can be better understood by looking at an analogue of the Mean Square Error. In particular put the "variance" of (3.1) and the squared "bias" of (3.2) into a type of Asymptotic Mean Square Error,

$$AMSE(\hat{h}_{PCV}, h_{MISE}) = \sigma^2 n^{-3/5} m^{-4/5} + C_1^2 n^{-6/5} m^{4/5}.$$

This asymptotic expression may now be minimized over m giving,

$$m_0 = (\sigma/C_1)^{5/4} n^{3/8}.$$

Note that while this result is of theoretical interest, the practical applicability of it is somewhat hampered by the fact that it depends on the unknowns, $\int (f'')^2$ and $\int (f''')^2$, although this objection is no more serious than the analagous one which may be raised about the plug-in bandwidth selectors considered in the next section. For this choice of m note that

$$(3.3) \qquad (\hat{h}_{PCV} - h_{MISE})/h_{MISE} \sim n^{-1/4},$$

which provides substantial improvement over (2.2).

4. COMPARISON WITH PLUG IN METHODS

While the method of partitioned cross-validation represents, at least in principle, a substantial improvement over ordinary cross-validation, an important question is: how does it compare with other bandwidth selection methods? An important competitor to cross-validation is the so-called plug-in methods, most recently studied by Sheather (1986). The motivation for this type of bandwidth selector comes from (2.3). In particular note that the only unknown part of C_0 is $\int(f'')^2$, which can be rather well estimated.

Hall and Marron (1987c) have shown that, by using a kernel type estimator to construct a $\hat{C}_0$, with careful choice of its smoothing parameter (observe that estimation of the integrated squared second derivative requires a much different amount of smoothing than curve estimation), under the assumption of four continuous derivatives,

$$\hat{C}_0 - C_0 = O(n^{-4/13})$$

Hence, again using (2.3) the bandwidth

$$\hat{h}_{PI} = \hat{C}_0 n^{-1/5},$$

satisfies

$$(\hat{h}_{PI} - h_{MISE})/h_{MISE} = O(n^{-4/13}),$$

which is better than both (2.2) and (3.3).

So while the partitioned cross-validated bandwidth represents an improvement over ordinary cross-validation, there is still some loss of efficiency compared to the theoretically best plug-in bandwidth. It seems that one possible reason for this is that when the sample is partitioned, there is some loss of "information" because the observations in each subsample are allowed to "interact" only with each other, and there is no possibility of "interaction" across subsamples.

N. I. Fisher has suggested overcoming this problem by replacing the average over only the partitioned subsets by an average over all subsamples of size n/m. Note that there are far too many subsamples to actually do this, but the effect could be well achieved by averaging over a randomly selected collection of subsamples. Analysis of this bandwidth selector goes beyond the scope of this paper, but it is speculated that the rate of convergence for this selector will be the same as for the plug-in selector. A further piece of speculation is that this rate will turn out to be the best possible in a minimax sense very similar to that described in Hall and Marron (1987b).

A very attractive potential application of the idea of partitioned cross-validation comes in the setting of nonparametric regression estimation. In that situation, it is well known that serial correlation of the error variables can be devastating to the performance of cross-validation (see Diggle 1985). Results very similar to those of this paper may be obtained in that context. A major difference is that the partitioning need not be done randomly, indeed one may consider taking every m-th point to form the subsamples. Note that in addition to having the enhanced sample stability properties described above, this method of cross-validation will also effectively overcome the serial correlation problem, as long as m is large enough that the errors associated with each sample are essentially independent. A reasonable choice of m in this context might be that value for which the corresponding lagged correlation is not significantly different from zero.

ACKNOWLEGEMENT

The ideas presented in this paper came from a number of interesting discussions with P. Hall and N. I. Fisher.

APPENDIX

Sketch of Proof of Theorem 1: This proof is essentially the same as the proofs of Theorems 2.1 and 2.2 of Hall and Marron (1987a). One big difference is that this result is stated for the difference between $\hat{h}^*_{CV}$ and h^*_{MISE} which is essentially a combination of the the two results. This is easily done by basing the proof on the following analogue of (2.5) of Hall and Marron:

$$0 = \mathrm{MISE}^{*\prime\prime}(h^{**})(\hat{h}^*_{CV} - h_{MISE}) + D^{*\prime}(\hat{h}^*_{CV}) + \delta^{*\prime}(\hat{h}^*_{CV}),$$

where

$$D^*(h) = \mathrm{ISE}^*(h) - \mathrm{MISE}^*(h),$$

$$\delta^*(h) = \mathrm{CV}^*(h) - \mathrm{ISE}^*(h) + \int f^2$$

and where h^{**} is between $\hat{h}^*_{CV}$ and h_{MISE}. Now proceed as in Hall and Marron (1987a), with the only other real difference being that Lemmas 3.4 and 3.5 are replaced by:

$$(n/m)^{7/10} m^{1/2} \begin{bmatrix} D^{*\prime}(h_{MISE}) \\ \delta^{*\prime}(h_{MISE}) \end{bmatrix} \xrightarrow{d} N\left[\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{0C}^2 \\ \sigma_{0C}^2 & \sigma_C^2 \end{bmatrix}\right]$$

(where σ_0^2, σ_{0C}^2 and σ_C^2 are the same as in Hall and Marron 1987a).

Sketch of Proof of Theorem 2: Using a Taylor expansion argument such as that of Rosenblatt (1971), as $n \to \infty$, $h \to 0$ with $nh \to \infty$,

$$\text{MISE} = a_1 n^{-1} h^{-1} - a_2 n^{-1} - a_3 n^{-1} h^2 + a_4 h^4 + a_5 h^6 + o(n^{-1} h^2 + h^6),$$

where

$$a_1 = \int K^2,$$
$$a_2 = \int f^2,$$
$$a_3 = (\int x^2 K) \int (f')^2,$$
$$a_4 = (\int x^2 K)^2 \int (f'')^2 / 4,$$
$$a_5 = (\int x^2 K)(\int x^4 K) \int (f''')^2 / 24.$$

Now let

$$D(h) = \text{MISE}(h) - [a_1 n^{-1} h^{-1} + a_4 h^4].$$

It can be shown that

$$\text{(A.A)} \qquad D'(h_{MISE}) = 6 C_0^5 a_5 n^{-1} + o(n^{-1}).$$

But,

$$\text{(A.B)} \quad 0 = M'(h_{MISE}) = -a_1 n^{-1} h_{MISE}^{-2} + 4 a_4 h_{MISE}^{3} + D'(h_{MISE})$$
$$= (h_{MISE} - C_0 n^{-1/5})[2 a_1 n^{-1} \tilde{h}^{-3} + 12 a_4 \tilde{h}^2] + D'(h_{MISE}),$$

where $\tilde{h}$ is between h_{MISE} and $C_0 n^{-1/5}$.

Theorem 2 follows from plugging (A.A) into (A.B), solving for $h_{MISE} - C_0 n^{-1/5}$, and then applying (2.3).

REFERENCES

Bowman, A. (1984), "An alternative method of cross-validation for the smoothing of density estimates," *Biometrika*, 65, 521-528.

Devroye, L. and Györfi, L. (1984). *Nonparametric Density Estimation: The L_1 View*. Wiley, New York.

Diggle, P. J. (1985), discussion of "Some aspects of the spline smoothing approach to nonparametric regression curve fitting" by B. W. Silverman, *Journal of the Royal Statistical Society, Series B*, 47, 28-29.

Hall, P. (1983), "Large sample optimality of least square cross-validation in density estimation," *Annals of Statistics* 11, 1156-1174.

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

Hall, P. and Marron, J. S. (1987b), "The amount of noise inherent in bandwidth selection for a kernel density estimator, " *Annals of Statistics*, to appear.

Hall, P. and Marron, J. S. (1987c), "Estimation of integrated squared density derivatives", North Carolina Institute of Statistics, Mimeo Series #1720.

Prakasa Rao, B. L. S. (1983), *Nonparametric Functional Estimation*, Academic Press, New York.

Rosenblatt, M. (1971), "Curve estimates," *Annals of Mathematical Statistics*, 42, 1815-1842.

Rudemo, M. (1982), "Empirical choice of histograms and kernel density estimators," *Scandanavian Journal of Statistics*, 9, 65-78.

Sheather, S. J. (1986). "An improved data-based algorithm for choosing the window width when estimating the density at a point," *Computational Statistics and Data Analysis*, 4, 61-65.

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

Stone, C. J. (1984), "An asymptotically optimal window selection rule for kernel density estimates," *Annals of Statistics*, 12, 1285-1297.