Comparison of Two Bandwidth Selectors with Dependent Errors

C. K. Chu and J. S. Marron[1]

National Tsing Hua University and University of North Carolina

September 1, 1989

ABSTRACT

For nonparametric regression, in the case of dependent observations, cross-validation is known to be severely affected by dependence. This effect is precisely quantified through a limiting distribution for the cross-validated bandwidth. The performance of two methods, the "leave-$(2\ell+1)$-out" version of cross-validation and partitioned cross-validation, which adjust for the dependence effect on bandwidth selection is investigated. The bandwidths produced by these two methods are analyzed by further limiting distributions which reveal significantly different characteristics. Simulations demonstrate that the asymptotic effects hold for reasonable sample sizes.

AMS 1980 subject classifications: Primary 62G05; secondary 62G20.

Keywords: cross-validation, autoregressive-moving average process, bandwidth selector, nonparametric regression.

[1]This research is part of the Ph. D. dissertation of the first author, under the supervision of the second at the University of North Carolina, Chapel Hill. It was partially supported by NSF Grant DMS-8701201.

1. INTRODUCTION

Nonparametric regression is a smoothing method for recovering the mean function from noisy data. It has been well established as a powerful and useful data-analytic tool. See the monographs by Eubank (1988), Haerdle (1988), and Mueller (1988) for a large variety of interesting real data examples where applications of this method have yielded analysis essentially unobtainable by other techniques.

The simplest and most widely used regression smoothers are based on kernel methods. Kernel estimators are local weighted averages of the response variables. The width of the neighborhood in which averaging is performed is called the bandwidth or smoothing parameter. The magnitude of bandwidth controls the smoothness of the resulting estimate of the regression function. For the independent observations, cross-validation provides an attractive data-based method for choosing the bandwidth, although it suffers from considerable sample noise. See Haerdle, Hall, and Marron (1988) for a detailed discussion of this. For other bandwidth selectors, see also Rice (1984) and Marron (1988).

However, if the observations are dependent, then the bandwidth selectors designed for independent observations will not produce good bandwidths. For instance, if the observations are positively correlated, then cross-validation will produce small bandwidths which result in rough kernel estimates. On the other hand, if the observations are negatively correlated, then cross-validation will produce large bandwidths which result in oversmooth kernel estimates. See Hart and Wehrly (1986), Hart (1987), Chiu (1987), and Diggle and Hutchinson (1989) for a detailed discussion of the dependence effect on

bandwidth selection.

For dependent observations, a central limit theorem (CLT) for the cross-validated bandwidth is given in Section 3 which quantifies the dependence effect on cross-validation by showing what this bandwidth converges to and by giving the rate of convergence for the cross-validated bandwidth. The rate of convergence is of the same order as that given in Haerdle, Hall, and Marron (1988) for the case of the independent observations, although the convergence is now not to the optimal bandwidth. This quantification motivates a modification of cross-validation to eliminate the dependence effect.

This adjustment is called modified cross-validation (MCV) and is simply the "leave-$(2\ell+1)$-out" version of cross-validation. See Collomb (1985), Haerdle and Vieu (1987), and Vieu and Hart (1989) for earlier results on the application of this method to the settings of strong mixing data. Based on an autoregressive-moving average (ARMA) model for the dependent regression errors, a CLT is given in Section 3 for the modified cross-validated bandwidth, for each $\ell \geq 0$. This CLT shows clearly how the dependence effect on cross-validation is alleviated as the value of ℓ is increased. However, the value of ℓ does not appear in the rate of convergence.

There are other possibilities for overcoming the dependence effect. Marron (1987) proposed partitioned cross-validation (PCV) for kernel density estimation to eliminate the sample noise inherent to cross-validation. The idea of PCV is to split the observations into g subgroups by taking every g-th observation. For the correlated data, as long as g is large enough, the errors associated with each subgroup are essentially independent. Marron (1987) mentioned that this method of

cross-validation should effectively overcome the dependence effect. While this is true, the resulting bandwidth is poor for a surprising reason. In Section 3, a CLT for the partitioned cross-validated bandwidth is derived, for each $g \geq 1$. The rate of convergence is faster than that for the modified cross-validated bandwidth. This rate of convergence is of the same order as that given in Marron (1987) for kernel density estimation. However, the asymptotic expectation reveals that there is a significant distance between the partitioned cross-validated bandwidth and the optimal bandwidth which minimizes the mean average square error. In fact the limiting distribution of this bandwidth is centered at the bandwidth which is optimal for no dependence, which is different from the true optimal. Essentially partitioned cross-validation does not work well because it is too effective at removing the dependence.

Another approach to bandwidth selection for correlated data is that of Hart (1987), Chiu (1989), and Diggle and Hutchinson (1989) who proposed methods to estimate the covariance function of the regression errors, plugging the estimated covariance function into bandwidth selectors. They showed, in simulation studies, that these plug-in bandwidth selectors would produce good bandwidths.

When the dependent observations are considered in nonparametric regression, a convenient dependence structure for analysis is the class of ARMA processes in time series analysis. Section 2 describes the regression setting and the precise formulation of these two bandwidth selectors. The asymptotic behaviors of bandwidth estimates produced by these two methods are given as theorems in Section 3. Section 4 contains simulation results which give additional insight into what the

theoretical results mean. Finally, sketches of proofs are given in Section 5.

2. REGRESSION MODEL AND BANDWIDTH SELECTORS

In this paper, the equally spaced fixed design and the short range dependence nonparametric regression model is considered. The model is given by, for $j = 1, 2, \ldots, n$,

$$Y_j = m(x_j) + \epsilon_j. \tag{2.1}$$

Here m is a smooth unknown regression function defined on the interval [0,1] (without loss of generality), x_j are equally spaced fixed design points, i.e. $x_j = j/n$, ϵ_j are an unknown ARMA process, and Y_j are noisy observations of the regression function m at the design points x_j. Using Definition 3.1.2 of Brockwell and Davis (1987), the process ϵ_j is an ARMA(p,q) process if ϵ_j is a stationary process and if there are positive integers p and q such that

$$\phi(B)\epsilon_j = \theta(B)e_j,$$

for all j. Here e_j are uncorrelated random variables with mean zero and finite variance σ^2, $\phi(z)$ and $\theta(z)$ are the p-th and q-th degree polynomials and B is the backward shift operator.

To estimate the regression function m(x), we consider a kernel estimator as introduced by Nadaraya (1964) and Watson (1964). Given a kernel function K and a bandwidth h, for $0 < x < 1$, the Nadaraya-Watson estimator is defined by

$$\hat{m}(x) = [n^{-1} \sum_{i=1}^{n} K_h(x-x_i)Y_i] \,/\, [n^{-1} \sum_{i=1}^{n} K_h(x-x_i)], \tag{2.2}$$

where $K_h(\cdot) = h^{-1}K(\cdot/h)$ (if the denominator is zero, take $\hat{m}(x) = 0$). See Chu and Marron (1989) for the comparison of this estimator to other types of kernel estimator.

The optimal bandwidth, h_M, is taken as the minimizer of the mean average square error (MASE). The MASE function is defined by

$$d_M(h) = E\{ n^{-1} \sum_{j=1}^{n} [\hat{m}(x_j) - m(x_j)]^2 W(x_j) \}, \tag{2.3}$$

where $\hat{m}(x_j)$ are kernel estimators of $m(x_j)$. The weight function W is introduced to allow elimination (or at least significant reduction) of boundary effects by taking W to be supported on a subinterval of the unit interval (see Gasser and Mueller (1979)).

For any $\ell \geq 0$, the "leave-$(2\ell+1)$-out" version of MCV is to choose the bandwidth by minimizing the modified cross-validation score

$$CV_\ell(h) = n^{-1} \sum_{j=1}^{n} [\hat{m}_j(x_j) - Y_j]^2 W(x_j).$$

Here $\hat{m}_j(x_j)$ are a "leave-$(2\ell+1)$-out" version of $\hat{m}(x_j)$, i.e. the observations (x_{j+i}, Y_{j+i}), $-\ell \leq i \leq \ell$, are left out in constructing $\hat{m}(x_j)$. For the Nadaraya-Watson estimator, $\hat{m}_j(x_j)$ are defined by

$$\hat{m}_j(x_j) = [(n-2\ell-1)^{-1} \sum_{i:|i-j|>\ell} K_h(x_j-x_i)Y_i] \,/\, [(n-2\ell-1)^{-1} \sum_{i:|i-j|>\ell} K_h(x_j-x_i)].$$

The amount of dependence between $\hat{m}_j(x_j)$ and Y_j is reduced as ℓ is increased. When $\ell = 0$, MCV is ordinary cross-validation. The minimizer of $CV_\ell(h)$ is denoted by $\hat{h}_{CV(\ell)}$.

For any $g \geq 1$, PCV is to calculate the ordinary cross-validation score $CV_{0,k}(h)$ of the k-th subgroup of observations, $k = 1, 2, \ldots, g$, and minimize the average of these score functions

$$CV^*(h) = g^{-1} \sum_{k=1}^{g} CV_{0,k}(h).$$

The minimizer of $CV^*(h)$ is denoted by $\hat{h}^*_{CV}$. Since $\hat{h}^*_{CV}$ is appropriate for sampled size only n/g, the partitioned cross-validated bandwidth $\hat{h}_{PCV(g)}$ is defined to be the rescaled $\hat{h}^*_{CV}$, $\hat{h}_{PCV(g)} = g^{-1/5}\hat{h}^*_{CV}$. This scale

factor is given in Section 5. When $g = 1$, PCV is ordinary cross-validation. For g large enough, the dependence effect inherent to $CV_{0,k}(h)$, for all k, becomes negligible.

3. RESULTS

In this section, we shall study the asymptotic behaviors of $\hat{h}_{CV(\ell)}$ and $\hat{h}_{PCV(g)}$ for any $\ell \geq 0$ and $g \geq 1$. In order to derive this, using the regression model (2.1) and the Nadaraya-Watson estimator (2.2), we must impose the following assumptions:

(A.1) The regression function $m(x)$ supported on the interval $[0,1]$ has a uniformly continuous fourth derivative $m^{(4)}(x)$ on the interval $(0,1)$.

(A.2) The kernel function K is a symmetric probability density function with support contained in the interval $[-1,1]$. The second derivative K'' of K is Hoelder continuous of order 1.

(A.3) The weight function W compactly supported on the interval $(0,1)$ with a nonempty interior has a uniformly continuous first derivative W'.

(A.4) The regression errors ϵ_j are an unknown ARMA(p,q) process for which the polynomials $\phi(z)$ and $\theta(z)$ have no common zeros, $\phi(z)$ has no zeros on $|z| \leq 1$, and e_j are independent and identically distributed (IID) random variables with mean zero and all finite moments.

(A.5) The autocovariance function $\gamma(\cdot)$ of the regression errors ϵ_j has a positive sum, i.e. $0 < \sum_{k=-\infty}^{\infty} \gamma(k) < \infty$.

(A.6) The total number of observations in this regression setting is n, with $n \longrightarrow \infty$. The "leave-$(2\ell+1)$-out" version of MCV is applied, with $\ell \ll n^{1/2}$. The number of subgroups of PCV is g, with $g \ll n^{1/2}$. The number of observations of each subgroup of PCV is $\eta = n/g$. For simplicity of notation, n is assumed to be a multiple of g.

(A.7) For any $\ell \geq 0$, the minimizer of $CV_\ell(h)$ is searched on the interval $H_n = [an^{-1/5}, bn^{-1/5}]$ for $n = 1, 2, \ldots$. For any $g \geq 1$, the minimizer of $CV^*(h)$ is searched on the interval $H_{n,g} = [a\eta^{-1/5}, b\eta^{-1/5}]$ for $\eta = 1, 2, \ldots$. Here the constant a is arbitrarily small and b is arbitrarily large.

Let the notation $X_n = o_u(v_n)$ mean that, as $n \longrightarrow \infty$, $|X_n/v_n| \longrightarrow 0$ almost surely, and uniformly on H_n if v_n involves h. Under the above assumptions, it is shown briefly in Section 5 that $d_M(h)$ can be asymptotically expressed as

$$d_M(h) = a_1 n^{-1}h^{-1} + b_1 h^4 + b_2 h^6 + o(h^6), \tag{3.1}$$

where

$$a_1 = \left(\sum_{k=-\infty}^{\infty} \gamma(k)\right)\int K^2 \int W,$$

$$b_1 = (1/4)\left(\int u^2 K\right)^2 \int (m'')^2 W,$$

$$b_2 = (-1/24)\left(\int u^2 K\right)\left(\int u^4 K\right)\left(\int (m^{(3)})^2 W + \int m'' m^{(3)} W'\right).$$

Here and throughout this paper, the notation $\int$ denotes $\int du$. For the components of MASE, the terms $a_1 n^{-1}h^{-1}$ and $b_1 h^4 + b_2 h^6$ represent the variance and the bias square respectively. A consequence of (3.1) is that the optimal bandwidth h_M can be asymptotically expressed as

$$h_M = C_0 n^{-1/5} + B_0 n^{-3/5} + o(n^{-3/5}), \tag{3.2}$$

where

$$C_0 = [a_1 / (4b_1)]^{1/5} = \left[\left(\sum_{k=-\infty}^{\infty} \gamma(k)\right)\int K^2 \int W \left(\int u^2 K\right)^{-2}\left(\int (m'')^2 W\right)^{-1}\right]^{1/5},$$

$$B_0 = (1/20)\left[\left(\sum_{k=-\infty}^{\infty} \gamma(k)\right)\int K^2 \int W\right)^{3/5}\left(\int u^4 K\right)\left(\int (m^{(3)})^2 W + \int m'' m^{(3)} W'\right)\right] / \left[\left(\int u^2 K\right)^{11}\left(\int (m'')^2 W\right)^8\right]^{1/5}.$$

We now quantify the dependence effect on the methods of cross-validation, the MCV for each $\ell \geq 0$ and the PCV for each $g \geq 1$, through the following limiting distributions. Let the coefficients $a^S_{1\ell}$

and $C^S_{0\ell}$ be the coefficients a_1 and C_0 with $[\sum_{k=-\infty}^{\infty} \gamma(k)\int K^2]$ replaced by $[\sum_{k=-\infty}^{\infty} \gamma(k)\int K^2 - 4K(0)\sum_{k>\ell} \gamma(k)]$ in each case. Let the coefficients a^S_{1g} and C^S_{0g} be the coefficients a_1 and C_0 with $[\sum_{k=-\infty}^{\infty} \gamma(k)\int K^2]$ replaced by $[\sum_{k=-\infty}^{\infty} \gamma(gk)\int K^2 - 4K(0)\sum_{k>0} \gamma(gk)]$ in each case.

<u>Theorem 1</u>: Under the above assumptions, as $n \longrightarrow \infty$, $b_1 > 0$, and $a^S_{1\ell} > 0$ for $\hat{h}_{CV(\ell)}$ and $a^S_{1g} >$ for $\hat{h}_{PCV(g)}$, then

$$(3.3)\quad n^{1/10}[\hat{h}_{CV(\ell)}/h_M - C^S_{0\ell}/C_0] \Rightarrow N(0,\ (C^S_{0\ell}/C_0)^{-7}(\sum_{k=-\infty}^{\infty} \gamma(k))^{1/5}\mathrm{Var}_M),$$

$$(3.4)\quad g^{2/5}n^{1/10}[\hat{h}_{PCV(g)}/h_M - C^S_{0g}/C_0] \Rightarrow N(0,\ v_g(\sum_{k=-\infty}^{\infty} \gamma(k))^{-2/5}\mathrm{Var}_M),$$

where

$$\mathrm{Var}_M = (8/25)\int(K*(K-L)-(K-L))^2\int W^2 \ /\ [(\int K^2)^9(\int W)^9(\int u^2K)^2(\int (m'')^2W)]^{1/5},$$

$$v_g = [\sum_{i=-\infty}^{\infty}\sum_{j=-\infty}^{\infty} \gamma(j)\gamma(j-ig)][\sum_{k=-\infty}^{\infty} \gamma(gk)\int K^2 - 4\sum_{k>0} \gamma(gk)K(0)]^{-7/5},$$

and where $L(u) = -uK'(u)$ and $*$ means convolution.

Remark 3.1: If $a^S_{1\ell} \le 0$, then $CV_\ell(h)$ is minimized at the left end of H_n asymptotically. If $a^S_{1g} \le 0$, then $CV^*(h)$ is also minimized at the left end of $H_{n,g}$ asymptotically. If $b_1 = 0$, then $CV_\ell(h)$ and $CV^*(h)$ are minimized at the right or the left ends of H_n and $H_{n,g}$ respectively, depending on the values of $a^S_{1\ell}$ and a^S_{1g}.

Remark 3.2: The rates of convergence for h_M, $\hat{h}_{CV(\ell)}$, and $\hat{h}_{PCV(g)}$ are of the same order as those given in Haerdle, Hall, and Marron (1988) and Marron (1987) for the respective cases with independent observations.

Remark 3.3: In the case of independent observations, the ratios, $C^S_{0\ell}/C_0$ and C^S_{0g}/C_0, are equal to 1 for any values of ℓ and g. However, for the

dependent observations, these two ratios have different values. For MCV, if $\ell \longrightarrow \infty$ with $\ell \ll n^{1/2}$, then $C^S_{0\ell}/C_0 \longrightarrow 1$ at a polynomial rate for the geometric boundedness of $\gamma(k)$ as given in Exercise 3.11 of Brockwell and Davis (1987). This means that MCV would produce asymptotically unbiased optimal bandwidth with respect to h_M whenever ℓ is moderately large. Thus MCV is adequate to the case that the expectation of $\hat{m}$ is important. For PCV, if $g \longrightarrow \infty$ with $g \ll n^{1/2}$, then $C^S_{0g}/C_0 \longrightarrow [\gamma(0) / \sum_{k=-\infty}^{\infty} \gamma(k)]^{1/5}$ at a polynomial rate. This means that PCV would produce asymptotically biased bandwidth with respect to h_M, no matter how large the value of g is. This asymptotic bias is caused by the distance g/n among the observations of each subgroup. An immediate remedy for this bias is to split the observations into g subgroups by taking every g-th cluster. Each cluster is composed of ζ consecutive observations. Thus PCV would be able to reflect the dependence structure of the whole data set. Since $\gamma(k)$ is geometrically bounded, it is enough to take the value of ζ as $O(\log n)$. A drawback of this approach is that it requires too many observations. Since $\hat{h}_{PCV(g)}$ has a faster rate of convergence than $\hat{h}_{CV(\ell)}$, then PCV is adequate to the case that the variance of $\hat{m}$ is important.

Remark 3.4: See sections 3.5, 4.3, and 5.4 of Chu (1989) for more detailed comparison of these two methods in the important special cases where the regression errors are a MA(1) process and an AR(1) process.

Remark 3.5: See sections 5.4 and 6.3 of Chu (1989) for the choice of the optimal value of g for PCV.

4. SIMULATIONS

To investigate the practical implications of the asymptotic results for $\hat{h}_{CV(\ell)}$ and $\hat{h}_{PCV(g)}$ presented in Section 3, an empirical study was carried out. We shall first introduce the simulated regression settings. The sample size was n = 200. The regression model (2.1) and the kernel estimator (2.2) were considered. The regression function was $m(x) = x^3(1-x)^3$ for $0 \leq x \leq 1$. The regression errors ϵ_j were an AR(1) process, i.e. $\epsilon_j = \phi\epsilon_{j-1} + e_j$, where e_j were IID $N(0,\sigma^2)$. The AR(1) parameters ϕ and σ were given as the following five combinations 0, 0.0177; 0.6, 0.0071; 0.6, 0.0018; -0.6, 0.0283; -0.6, 0.0029. Only the second combination is treated the numerical results in Table 1. The values of σ make h_M correspond roughly 1/5, 1/4, or 1/2, and give different amounts of sample variability of the regression settings with the same value of ϕ. The kernel function was $K(x) = (15/8)(1-4x^2)^2$ for $-1/2 \leq x \leq 1/2$. The weight function was $W(x) = 5/3$ for $1/5 \leq x \leq 4/5$. The same functions K and m were also used in Rice (1984) and Haerdle, Hall, and Marron (1988). For each combination of ϕ and σ, 1000 independent sets of data were generated. For MCV, the values of ℓ were 0, 1, 2, ..., 14. For PCV, the values of g were 1, 2, ..., 15. The values of the score functions, the average square function in $d_M(h)$, $CV_\ell(h)$, and $CV^*(h)$, were calculated on an equally spaced logarithmic grid of 11 values. The endpoints of the grid were different for the different setting, and chosen to contain essentially all the bandwidths of interest. The expectation in $d_M(h)$ was empirically approximated by averaging the average square error over the 1000 pseudo data sets. The minimizers, h_M, $\hat{h}_{CV(\ell)}$, and $\hat{h}^*_{CV}$ of the score functions $d_M(h)$, $CV_\ell(h)$, and $CV^*(h)$ respectively, were calculated. After evaluation on the grid,

a one step interpolation improvement was done, with the results taken as the selected bandwidths. If the score functions had multiple minimizers on the grid, the algorithm chose the smaller of them (this choice was made arbitrarily).

The sample variances, the sample bias-squares, and the MSE of the bandwidth estimates $\hat{h}/h_M$ were summarized, where $\hat{h}$ denotes $\hat{h}_{CV(\ell)}$ and $\hat{h}_{PCV(g)}$. The sample bias-square of the bandwidth estimates was taken as the square of the average of the 1000 values of $\hat{h}/h_M - 1$. The MSE was the sum of the sample variance and the sample bias-square. For the first combination, where $\phi = 0$ (independent observations) and $\sigma = 0.0177$ (h_M roughly equals 1/2), the bias-squares for $\hat{h}_{CV(\ell)}$ and $\hat{h}_{PCV(g)}$ were roughly constant over ℓ and g as predicted by our theorem. As ℓ and g increased, the variances for $\hat{h}_{CV(\ell)}$ stayed the same, but the variances for $\hat{h}_{PCV(g)}$ decreased also as predicted. In this case, PCV is preferred to MCV. The numerical results as given in Table 1 represent the second combination, where $\phi = 0.6$, and $\sigma = 0.0071$ (h_M roughly equals 1/2). In this case, the bias-squares for $\hat{h}_{CV(\ell)}$ decreased to 0 as ℓ increased. However, the bias-squares for $\hat{h}_{PCV(g)}$ converged to a nonzero constant as g increased. In contrast to the bias-squares, the variances for $\hat{h}_{CV(\ell)}$ stayed the same for all ℓ and the variances for $\hat{h}_{PCV(g)}$ decreased monotonely as g increased. Here, variance is the dominant term in MSE. Thus, using PCV to reduce the variance of bandwidth estimate would result in a smaller value of MSE than using MCV to reduce the bias-square of the bandwidth estimate.

[Put Table 1 about here.]

In the case of the third combination, variance and bias-square had the same tendency as the second combination. In this case, bias-square is

the dominant term in MSE. Thus using MCV to reduce bias-square would give better MSE than using PCV to reduce variance. In the final two cases where $\phi = -0.6$, and $\sigma = 0.0283$ (h_M roughly equals 1/2) and $\sigma = 0.0029$ (h_M roughly equals 1/5), the variances and the bias-squares for $\hat{h}_{PCV(g)}$ decreased along even g's and odd g's separately. This is because $\phi^g = (\phi^2)^k$ where $g = 2k$ for the even number of g and some k. The conclusions for these two cases are the same as those for the second and the third combinations. Finally, the choice between MCV and PCV should be made on the basis of which component, variance or bias-square, is the dominant term in MSE.

5. SKETCHES OF PROOFS

The following notation and results will be used in this section. For all integers i and j, let X_i be IID random variables with mean zero and all finite moments, and a_i and b_{ij} be real numbers such that $\sum_{i=-\infty}^{\infty} |a_i| < \infty$ and $\sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} |b_{ij}| < \infty$. Using Theorem 2 of Whittle (1960) and Theorem A of section 1.4 of Serfling (1980), then, for all positive integers k, we have

$$E((\sum_{i=-\infty}^{\infty} a_i X_i)^{2k}) \le c_1(\sum_{i=-\infty}^{\infty} a_i^2)^k,$$

$$E((\sum_{i=-\infty, i\neq j}^{\infty} \sum_{j=-\infty}^{\infty} b_{ij} X_i X_j)^{2k}) \le c_2(\sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} b_{ij}^2)^k,$$

where c_1 and c_2 are constants involving k and moments of X. Let ϵ_j be a linear process defined by $\epsilon_j = \sum_{i=0}^{\infty} \psi_i e_{j-i}$, for $j = 1, 2, \ldots, n$, where ψ_i are real numbers with $\sum_{i=0}^{\infty} |\psi_i| < \infty$ and e_j are IID random variables with mean zero and all finite moments. Using Fubini's Theorem, Theorem 2 of Whittle (1960), Minkowski's inequality, and Theorem A of Serfling

(1980), then, for all positive integers k, we have

$$E(\sum_{j=1}^{n} \epsilon_j)^{2k} = O(n^k).$$

For any $\ell \ll n^{1/2}$ and each x_j with $W(x_j) \neq 0$, or $h < x_j < 1-h$, under the assumptions given in Section 3, we have the following asymptotic results:

$$n^{-1} \sum_{i=1}^{n} K_h(x_j-x_i) = 1 + O(n^{-1}h^{-1}),$$

$$(n-2\ell-1)^{-1} \sum_{i:|i-j|>\ell} K_h(x_j-x_i) = 1 + O(\ell n^{-1}h^{-1}),$$

$$\begin{aligned} b_j &= [n^{-1} \sum_{i=1}^{n} K_h(x_j-x_i)(m(x_i)-m(x_j))] \,/\, [n^{-1} \sum_{i=1}^{n} K_h(x_j-x_i)] \\ &= (1/2)h^2 m''(x_j)\int u^2 K + (1/24)h^4 m^{(4)}(x_j)\int u^4 K + o(h^4), \end{aligned}$$

$$\begin{aligned} v_j &= [n^{-1} \sum_{i=1}^{n} K_h(x_j-x_i)\epsilon_i] \,/\, [n^{-1} \sum_{i=1}^{n} K_h(x_j-x_i)] \\ &= n^{-1} \sum_{i=1}^{n} K_h(x_j-x_i)\epsilon_i + o_u((nh)^{-7/5}) = o_u((nh)^{-2/5}). \end{aligned}$$

Proofs of (3.1):

Since $d_M(h) = n^{-1} \sum_{j=1}^{n} b_j^2 W(x_j) + n^{-1} \sum_{j=1}^{n} E(v_j^2)W(x_j)$, using the asymptotic results of b_j and v_j as given above, through a straightforward calculation, then the proof of (3.1) is complete.

Proof of Theorem 1:

We first give asymptotic expressions of $\hat{h}_{CV(\ell)}$ and $\hat{h}_{PCV(g)}$ for each $\ell \geq 0$ and $g \geq 1$. Through adding and subtracting the terms $\hat{m}(x_j)$ and $m(x_j)$, then $CV_\ell(h)$ can be expressed as

$$(5.4)\quad CV_\ell(h) = n^{-1} \sum_{j=1}^{n} \epsilon_j^2 W(x_j) + d_A(h) - 2\mathrm{Cross}_\ell(h) + \mathrm{Remainder}_\ell(h),$$

where

$$\text{Cross}_\ell(h) = n^{-1} \sum_{j=1}^{n} \epsilon_j[\hat{m}_j(x_j) - m(x_j)]W(x_j),$$

$$\text{Remainder}_\ell(h) = n^{-1} \sum_{j=1}^{n} [\hat{m}_j(x_j) - \hat{m}(x_j)][\hat{m}_j(x_j) + \hat{m}(x_j) - 2m(x_j)]W(x_j).$$

Using the asymptotic results of b_j and v_j, through a straighforward calculation, then, as $n \longrightarrow \infty$,

$$\text{(5.5)} \qquad \text{Cross}_\ell(h) = 2n^{-1}h^{-1}\Big(\sum_{k>\ell} \gamma(k)\Big)K(0)\int W + o_u(d_M(h)),$$

$$\text{(5.6)} \qquad \text{Remainder}_\ell(h) = o_u(d_M(h)).$$

As $n \longrightarrow \infty$, $a^S_{1\ell} > 0$, and $b_1 > 0$, through a straightforward calculation, then

$$\hat{h}_{CV(\ell)} = C^S_{0\ell}n^{-1/5}(1 + o_u(1)).$$

Using the results of (5.4) through (5.6), through a straightforward calculation, then $CV^*(h)$ can be asymptotically expressed as

$$CV^*(h) = n^{-1} \sum_{j=1}^{n} \epsilon_j^2 W(x_j) + a^S_{1g}\eta^{-1}h^{-1} + b_1h^4 + o_u(\eta^{-1}h^{-1}+h^4).$$

This implies that, as $n \longrightarrow \infty$, $a^S_{1g} > 0$, and $b_1 > 0$, then

$$\hat{h}^*_{CV} = C^S_{0g}\eta^{-1/5}(1 + o_u(1)).$$

Since the optimal bandwidth h_M is of the order $n^{-1/5}$ and $\hat{h}^*_{CV}$ is of the order $\eta^{-1/5} = g^{1/5}n^{-1/5}$, then $\hat{h}_{PCV(g)}$ is defined as $g^{-1/5}\hat{h}^*_{CV}$.

Using the unique linear solution of the ARMA process ϵ_j, the asymptotic properties given above, and Fubini's Theorem, the proof of Theorem 1 is essentially the same as the proofs of Theorem 1 of Marron (1987), and Theorems 1 and 2 of Haerdle, Hall, and Marron (1988). The only difference is that $\hat{h}_{CV(\ell)}$ should be close to $C^S_{0\ell}n^{-1/5}$ and $\hat{h}_{PCV(g)}$ close to $C^S_{0g}n^{-1/5}$, not h_M.

REFERENCES

Brockwell, P. J. and Davis, R. A. (1987). Time Series: Theory and Methods. Springer Series in Statistics, Springer-Verlag, New York.

Chiu, S. T. (1987). Estimating the parameters of the noise spectrum for a time series with trend: with application to bandwidth selection for nonparametric regression. To appear.

Chiu, S. T. (1989). Bandwidth selection for kernel estimation with correlated noise. To appear in Statistics and Probability Letters.

Chu, C. K. (1989). Some results in nonparametric regression. Ph. D. Dissertation, Department of Statistics, University of North Carolina.

Chu, C. K. and Marron, J. S. (1989). Comparison of kernel regression estimators, unpublished paper.

Clark, R. M. (1975). A calibration curve for radiocarbon data. Antiquity, 49, 251-266.

Collomb, G. (1985). Non parametric time series analysis and prediction: uniform almost sure convergence of the window and K-NN autoregressive estimates. Statistics, 16, 297-307.

Diggle, P. J. and Hutchinson, M. F. (1989). On spline smoothing with autocorrelated errors. Australia Journal of Statistics, 31, 166-182.

Eubank, R. L. (1988). Spline Smoothing and Nonparametric Regression. Marcel Dekker Inc., New York.

Gasser, T. and Mueller, H. G. (1979). Kernel estimation of regression functions. In Smoothing techniques for curve estimation, Lecture Notes in Math., No. 757, 23-68, Spring-Verlag, New York.

Haerdle, W. (1988). Applied nonparametric regression, unpublished monograph.

Haerdle, W., Hall, P., and Marron, J. S. (1988). How far are autometically chosen regression smoothing parameters from their optimum? Journal of the American Statistical Association, 83, 86-101.

Haerdle, W. and Marron, J. S. (1985). Optimal bandwidth selection in nonparametric regression function estimation. Annals of Statistics, 13, 1465-1481.

Haerdle, W. and Vieu, P. (1987). Non parametric kernel regression function estimation for ϕ-mixing observations, Part I: Optimal squared error estimation. To appear.

Hart, J. D. (1987). Kernel regression estimation with time series errors. To appear.

Hart, J. D. and Wehrly, T. E. (1986). Kernel regression using repeated measurements data. Journal of the American Statistical Association, 81, 1080-1088.

Marron, J. S. (1987). Partitioned cross-validation. Econometric Reviews, 6, 271-284.

Marron, J. S. (1988). Automatic smoothing parameter selection: A survey. Empirical Economics, 13, 187-208.

Mueller, H. G. (1988). Nonparametric Analysis of Longitudinal Data, Lecture Notes in Statistics, No. 46, Spring-Verlag, Berlin.

Nadaraya, E. A. (1964). On estimating regression. Theory of Probability and its Applications, 9, 141-142.

Rice, J. (1984). Bandwidth choice for nonparametric regression. Annals of Statistics, 12, 1215-1230.

Serfling, R. (1980). Approximation Theorems of Mathematical Statistics, Wiley, New York.

Vieu, P. and Hart, J. (1989). Nonparametric regression under dependence: A class of asymptotically optimal data-driven bandwidths. To appear.

Watson, G. S. (1964). Smooth regression analysis. Sankhya, Series A, 26, 359-372.

Whittle, P. (1960). Bounds for the moments of linear and quadratic forms in independent variables. Theory of Probability and its Applications, 5, 302-305.

Table 1: The sample MSE of $\hat{h}_{CV(\ell)}/h_M$ and $\hat{h}_{PCV(g)}/h_M$ for the positively correlated observations with a large amount of sample variability.

Ratios		Variance	Bias-square	MSE
h_A/h_M		0.067989	0.000839	0.068828
$\hat{h}_{CV(\ell)}/h_M$	ℓ Value			
	0	0.015217	0.340901	0.356118
	1	0.094015	0.157793	0.251808
	2	0.129529	0.066091	0.195620
	3	0.142950	0.035485	0.178436
	4	0.144861	0.022526	0.167387
	5	0.150511	0.016509	0.167020
	6	0.152457	0.013700	0.166157
	7	0.153662	0.011340	0.165001
	8	0.152041	0.010281	0.162322
	9	0.154379	0.009622	0.164000
	10	0.148788	0.008326	0.157113
	11	0.148136	0.008082	0.156217
	12	0.145901	0.007954	0.153855
	13	0.142961	0.005565	0.148526
	14	0.143305	0.004422	0.147727
$\hat{h}_{PCV(g)}/h_M$	g Value			
	1	0.014969	0.342923	0.357892
	2	0.051444	0.285233	0.336677
	3	0.079858	0.186568	0.266425
	4	0.082228	0.127764	0.209992
	5	0.075802	0.091330	0.167132
	6	0.072712	0.071990	0.144702
	7	0.065812	0.059694	0.125507
	8	0.063662	0.055051	0.118712
	9	0.059589	0.049922	0.109511
	10	0.056709	0.050466	0.107175
	11	0.054767	0.049093	0.103862
	12	0.054734	0.046613	0.101347
	13	0.052097	0.046844	0.098941
	14	0.046694	0.048023	0.094718
	15	0.045878	0.048859	0.094738