The Annals of Statistics
1991, Vol. 19, No. 4, 1906–1918

COMPARISON OF TWO BANDWIDTH SELECTORS WITH DEPENDENT ERRORS[1]

BY C.-K. CHU AND J. S. MARRON

National Tsing Hua University and University of North Carolina, Chapel Hill

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-$(2l + 1)$-out" version of cross-validation and partitioned cross-validation, which adjust for the effect of dependence 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.

1. Introduction. Nonparametric regression is a smoothing method for recovering the regression function from noisy data. It has been well established as a powerful and useful data-analytic tool. See the monographs by Eubank (1988), Mueller (1988) and Härdle (1990) for a large variety of interesting examples where applications of this method have yielded analyses 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 kernel function is a given function to calculate the weights assigned to the observations. 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 independent observations, cross-validation is an attractive data-based method for choosing the bandwidth, although it suffers from considerable sample noise. See Härdle, 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

Received October 1989; revised December 1990.

[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-87-01201. The research of the first author was also partially supported by the National Science Council under contract NSC-80-0208-M007-32.

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

Key words and phrases. Cross-validation, partitioned cross-validation, autoregressive moving average process, bandwidth selector, nonparametric regression.

produce small bandwidths which result in rough kernel estimates of the regression function. On the other hand, if the observations are negatively correlated, then cross-validation will produce large bandwidths which result in oversmooth kernel estimates of the regression function. See Hart and Wehrly (1986), Chiu (1989), Diggle and Hutchinson (1989) and Hart (1991) for a detailed discussion of the effect of dependence on bandwidth selection.

For dependent observations, a central limit theorem (CLT) for the cross-validated bandwidth is given in Section 3 which quantifies the effect of dependence 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 Härdle, Hall and Marron (1988) for the case of 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-$(2l + 1)$-out" version of cross-validation. See Härdle and Vieu (1987), Vieu and Hart (1989) and Gyorfi, Härdle, Sarda and Vieu (1989) for earlier results on applications of this method to various settings involving mixing data. Section 3 contains a CLT for the modified cross-validated bandwidth, for each $l \geq 0$. This CLT shows clearly how the effect of dependence on cross-validation is alleviated as the value of l is increased. The value of l does not appear in the rate of convergence.

There are other possibilities for overcoming the effect of dependence on bandwidth selection. 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 gth observation. For 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 of this bandwidth 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 mean of this bandwidth 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 optimum. Essentially, PCV does not work well because it is too effective at removing the dependence.

When 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 MCV and PCV. The asymptotic behaviors of bandwidth estimates produced by MCV and PCV are given in Section 3. Based on the

linear structure of ARMA processes, the results of this paper could be extended directly to a more general dependence structure. Section 4 contains simulation results which give additional insight into what the theoretical results mean. Finally, sketches of the 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 $j = 1, 2, \ldots, n$,

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

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, that is, $x_j = j/n$, ε_j are an unknown causal ARMA process [see Definitions 3.1.2 and 3.1.3 of Brockwell and Davis (1987) for this process] and Y_j are noisy observations of the regression function m at the design points x_j. In this model, the design points become closer together as the sample size increases, but the error process remains the same.

To estimate the regression function m, 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

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

where $K_h(\cdot) = h^{-1}K(\cdot/h)$ [if the denominator is 0, take $\hat{m}(x) = 0$]. See Chu and Marron (1990) for the comparison of this estimator to other types of kernel estimators.

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

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

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

$$\mathrm{CV}_l(h) = n^{-1}\sum_{j=1}^{n}\left(\hat{m}_j(x_j) - Y_j\right)^2 W(x_j).$$

Here $\hat{m}_j(x_j)$ is a "leave-$(2l + 1)$-out" version of $\hat{m}(x_j)$, that is, the observations (x_{j+i}, Y_{j+i}), $-l \leq i \leq l$, are left out in constructing $\hat{m}(x_j)$. For the

Nadaraya–Watson estimator, $\hat{m}_j(x_j)$ are defined by

$$\hat{m}_j(x_j) = \frac{(n-2l-1)^{-1}\Sigma_{i:|i-j|>l} K_h(x_j - x_i)Y_i}{(n-2l-1)^{-1}\Sigma_{i:|i-j|>l} K_h(x_j - x_i)}.$$

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

For any $g \geq 1$, PCV involves splitting the observations into g subgroups by taking every gth observation, calculating the ordinary cross-validation score $\mathrm{CV}_{0,k}(h)$ of the kth subgroup of observations separately, for $k = 1, 2, \ldots, g$, and minimizing the average of these ordinary cross-validation score

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

If n is not a multiple of g, then the observations Y_j, $j \leq g[n/g]$, are applied to construct $\mathrm{CV}^*(h)$ and the rest of the observations are dropped out in constructing $\mathrm{CV}^*(h)$. The notation $[x]$ denotes the largest integer which is less than or equal to x. Here for simplicity of notation, assume n is a multiple of g. The amount of dependence inherent to each $\mathrm{CV}_{0,k}(h)$ is reduced as g is increased. The minimizer of $\mathrm{CV}^*(h)$ is denoted by $\hat{h}^*_{\mathrm{CV}}$. Since $\hat{h}^*_{\mathrm{CV}}$ is appropriate for the sample size n/g, the partitioned cross-validated bandwidth $\hat{h}_{\mathrm{PCV}(g)}$ is defined to be the rescaled $\hat{h}^*_{\mathrm{CV}}$, $h_{\mathrm{PCV}(g)} = g^{-1/5}\hat{h}^*_{\mathrm{CV}}$. This scale factor is explained in Section 5. When $g = 1$, PCV is ordinary cross-validation.

3. Results. In this section, we shall study the asymptotic behaviors of $\hat{h}_{\mathrm{MCV}(l)}$ for any $l \geq 0$, and $\hat{h}_{\mathrm{PCV}(g)}$ for any $g \geq 1$. For these, using the regression model (2.1) and the kernel estimator (2.2), we impose the following assumptions:

(A.1) The regression function $m(x)$ supported on the interval $[0, 1]$ has a uniformly continuous and square integrable second derivative $m''(x)$ on the interval $(0, 1)$.

(A.2) The kernel function K is a square integrable and symmetric probability density function with support contained in the interval $[-1, 1]$. Also, K has a Hölder continuous of order 1 second derivative. The function f is said to be Hölder continuous of order 1 if there is a constant b such that $|f(s) - f(t)| \leq b \cdot |s - t|$ for any s and t in the domain of f.

(A.3) The weight function W is bounded, Hölder continuous of order 1 and supported on a subinterval of the interval $(0, 1)$.

(A.4) The regression errors ε_j are obtained from e_j by application of a casual linear filter [see Definition 3.1.3 of Brockwell and Davis (1987) for the filter]. Here e_j are independent and identically distributed (iid) random variables with mean 0 and all finite moments.

(A.5) The autocovariance function $\gamma(\cdot)$ of ε_j satisfies $0 < \Sigma_{k=-\infty}^{\infty}\gamma(k) < \infty$.

(A.6) The total number of observations in this regression setting is n, with $n \to \infty$. The "leave-$(2l + 1)$-out" version of MCV is applied. The number of subgroups of PCV is g. The number of observations of each subgroup of PCV is $\eta = n/g$.

(A.7) For any $l \geq 0$, the minimizer of $CV_l(h)$ is searched on the interval $H_n = [\alpha n^{-1/5}, \beta n^{-1/5}]$, for $n = 1, 2, \ldots$. For any $g \geq 1$, the minimizer of $CV^*(h)$ is searched on the interval $H_{n,g} = [\alpha \eta^{-1/5}, \beta \eta^{-1/5}]$, for $\eta = 1, 2, \ldots$. Here the constant α is arbitrarily small and β is arbitrarily large.

Under the above assumptions, it is shown briefly in Section 5 that $d_M(h)$ can be asymptotically expressed as

$$d_M(h) = V(nh)^{-1} + B_2 h^4 + o\big((nh)^{-1} + h^4\big), \tag{3.1}$$

where

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

$$B_2 = \frac{1}{4}\left(\int u^2 K\right)^2 \int (m'')^2 W.$$

Here and throughout this paper, the notation $\int$ denotes $\int du$. For the components of MASE, $V(nh)^{-1}$ and $B_2 h^4$ 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

$$h_M = Cn^{-1/5}(1 + o(1)), \tag{3.2}$$

where

$$C = \left[\frac{V}{4B_2}\right]^{1/5} = \left[\left(\sum_{k=-\infty}^{\infty} \gamma(k)\right) \int K^2 \int W \left(\int u^2 K\right)^{-2} \left(\int (m'')^2 W\right)^{-1}\right]^{1/5}.$$

We now quantify the effects of dependence on MCV for each $l \geq 0$ and PCV for each $g \geq 1$, through the following limiting distributions which are shown in Section 5. Let the coefficients $V_{\mathrm{MCV}(l)}$ and $C_{\mathrm{MCV}(l)}$ be the coefficients V and C with $[\Sigma_{k=-\infty}^{\infty} \gamma(k) \int K^2]$ replaced by $[\Sigma_{k=-\infty}^{\infty} \gamma(k) \int K^2 - 4K(0) \Sigma_{k>l} \gamma(k)]$ in each case. Let the coefficients $V_{\mathrm{PCV}(g)}$ and $C_{\mathrm{PCV}(g)}$ be the coefficients V and C with $[\Sigma_{k=-\infty}^{\infty} \gamma(k) \int K^2]$ replaced by $[\Sigma_{k=-\infty}^{\infty} \gamma(gk) \int K^2 - 4K(0) \Sigma_{k>0} \gamma(gk)]$ in each case.

THEOREM 1. *Under the above assumptions, if* $B_2 > 0$, $V_{\mathrm{MCV}(l)} > 0$ *for* $\hat{h}_{\mathrm{MCV}(l)}$ *and* $V_{\mathrm{PCV}(g)} > 0$ *for* $\hat{h}_{\mathrm{PCV}(g)}$, *then*

$$n^{1/10}\left[\hat{h}_{\mathrm{MCV}(l)}/h_M - C_{\mathrm{MCV}(l)}/C\right] \Rightarrow N(0, \mathrm{VAR}_{\mathrm{MCV}(l)}), \tag{3.3}$$

$$g^{2/5} n^{1/10}\left[\hat{h}_{\mathrm{PCV}(g)}/h_M - C_{\mathrm{PCV}(g)}/C\right] \Rightarrow N(0, \mathrm{VAR}_{\mathrm{PCV}(g)}), \tag{3.4}$$

where

$$\mathrm{VAR}_{\mathrm{MCV}(l)} = \left[\frac{C_{\mathrm{MCV}(l)}}{C}\right]^{-7}\left[\sum_{k=-\infty}^{\infty}\gamma(k)\right]^{1/5} C_M,$$

$$\mathrm{VAR}_{\mathrm{PCV}(g)} = C_g\left[\sum_{k=-\infty}^{\infty}\gamma(k)\right]^{-2/5} C_M$$

and where

$$C_M = \left(\frac{8}{25}\right)\frac{\int(K^*(K-L)-(K-L))^2\int W^2}{\left[(\int K^2)^9(\int W)^9(\int u^2K)^2\left(\int(m'')^2W\right)\right]^{1/5}},$$

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

Here $L(u) = -uK'(u)$, $*$ means convolution and K' denotes the first derivative of K.

REMARK 3.1. If $V_{\mathrm{MCV}(l)} \le 0$, then $\hat{h}_{\mathrm{MCV}(l)}$ is at the left end of H_n asymptotically. If $V_{\mathrm{PCV}(g)} \le 0$, then $\hat{h}_{\mathrm{PCV}(g)}$ is also at the left end of $H_{n,g}$ asymptotically. Theorem 2 of Hart (1991) gave a similar condition such that, with probability tending to 1, cross-validation picks arbitrarily small bandwidths in the case of positively correlated data. If $B_2 = 0$, then, asymptotically, h_M is at the right end of H_n, and $\hat{h}_{\mathrm{MCV}(l)}$ and $\hat{h}_{\mathrm{PCV}(g)}$ are at the right or the left ends of H_n and $H_{n,g}$ respectively, depending on the values of $V_{\mathrm{MCV}(l)}$ and $V_{\mathrm{PCV}(g)}$.

REMARK 3.2. The rates of convergence of h_M, $\hat{h}_{\mathrm{MCV}(l)}$ and $\hat{h}_{\mathrm{PCV}(g)}$ are of the same order as those given in Härdle, 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 $C_{\mathrm{MCV}(l)}/C$ and $C_{\mathrm{PCV}(g)}/C$ are equal to 1 for any values of l and g. However, when the regression errors ε_j are a casual ARMA process, these two ratios have different values. For MCV, if l increases, then $C_{\mathrm{MCV}(l)}/C$ converges to 1 at a polynomial rate [see Exercise 3.11 of Brockwell and Davis (1987) for this convergence rate]. This means MCV would produce a nearly asymptotically unbiased bandwidth with respect to h_M whenever l is moderately large. For PCV, if g increases, then $C_{\mathrm{PCV}(g)}/C$ converges to $[\gamma(0)/\Sigma_{k=-\infty}^{\infty}\gamma(k)]^{1/5}$ at a polynomial rate. This means PCV would produce an 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 reducing this bias is to split the observa-

tions into g subgroups by taking every gth cluster. Each cluster is composed of ζ consecutive observations. Thus PCV would be able to reflect the dependence structure of the data through that of each cluster. In this case, if g increases, then $C_{\mathrm{PCV}(g)}/C$ converges to $[\Sigma_{|k|<\zeta}\gamma(k)/\Sigma_{k=-\infty}^{\infty}\gamma(k)]^{1/5}$. A drawback of this approach is that it will often require too many observations.

REMARK 3.4. Consider ε_j as observations on a weakly stationary process which has a spectral density $f(\lambda)$, $-\pi \leq \lambda \leq \pi$. Based on the results of Section 2.1 of Olshen (1967), ε_j are a moving average $\varepsilon_j = \Sigma_{i=-\infty}^{\infty}\psi_i e_{j-i}$, where $\Sigma_{i=-\infty}^{\infty}\psi_i^2 < \infty$ and e_j are orthonormal random variables, that is, $E[e_i e_j] = \delta_{ij}$. In this case, Theorem 1 and Remarks 3.1 through 3.3 still hold if ψ_i satisfy $\Sigma_{k=-\infty}^{\infty}|k\gamma(k)| < \infty$ and e_j have the same properties as they did in (A.4). Here $\gamma(k)$ are defined by $\gamma(k) = E(e_1^2)\Sigma_{i=-\infty}^{\infty}\psi_i\psi_{i+k}$ for all integers k.

REMARK 3.5. The bandwidth selectors of Chiu (1989) and Hart (1991) are modifications of Mallow's criterion. Based on the parametric model of the spectral density of ε_j, Hart and Chiu proposed methods to estimate the unknown factors of this criterion. If their estimates of the unknown factors converge to the true values with a rate $n^{-1/2}$ (or somewhat slower than $n^{-1/2}$), then the limiting distributions of their bandwidth estimates are the same as that of $\hat{h}_{\mathrm{MCV}(l)}$ for the case that l is sufficiently large to get rid of the dependence effect. This is based on the asymptotic equivalence of cross-validation and Mallows' criterion. See Härdle, Hall and Marron (1988) for this, in the case of independent observations. However, if the underlying model of the spectral density of ε_j is incorrect, then their methods can not be expected to perform as well as MCV.

REMARK 3.6. A possible approach for practical choice of the value of l is based on an analogue of the mean square error [see Marron (1987) for the discussion of this idea]. Using the asymptotic variance and the asymptotic mean of $\hat{h}_{\mathrm{MCV}(l)}/h_M$ given in (3.3), the asymptotic mean square error (AMSE) of this ratio is defined by

$$\mathrm{AMSE}\left(\hat{h}_{\mathrm{MCV}(l)}/h_M\right) = n^{-1/5}\mathrm{VAR}_{\mathrm{MCV}(l)} + \left[C_{\mathrm{MCV}(l)}/C - 1\right]^2. \tag{3.5}$$

Theoretically, if there is a value l_0 which minimizes (3.5) over $l \geq 0$, then l_0 is taken as the optimal value of l in the sense of AMSE. For positively correlated data, the value of l_0 is as large as possible. In other cases, the value of l_0 may depend on the unknown factors m and $\gamma(\cdot)$. In practice, we may plug estimates of the unknown factors into (3.5) to get an estimate $\hat{l}_0$ of l_0. However, the performance of $\hat{l}_0$ derived by this approach needs further study. The same ideas apply to g. Based on this plug-in AMSE approach, it is also possible to choose between MCV and PCV, depending on which gives the smallest estimated AMSE. The performance of this approach also needs further study.

4. Simulations. To investigate the practical implications of the asymptotic results for $\hat{h}_{\mathrm{MCV}(l)}$ and $\hat{h}_{\mathrm{PCV}(g)}$ 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 $0 \le x \le 1$. The kernel function was $K(x) = (15/8)(1 - 4x^2)^2$ for $-1/2 \le x \le 1/2$. The weight function was $W(x) = 5/3$ for $1/5 \le x \le 4/5$. The same functions m and K were also used in Rice (1984) and Härdle, Hall and Marron (1988). The regression errors ε_j were an AR(1) process, that is, $\varepsilon_j = \phi\varepsilon_{j-1} + e_j$, where e_j were pseudo iid normal random variables $N(0, \sigma^2)$ and ε_1 was $N(0, \sigma^2/(1 - \phi^2))$. The AR(1) parameters were $\phi = 0.6$ and $\sigma = 0.0071$, although discussion of others is given in Section 6.2 of Chu (1989). For the given m, K, W and ϕ, based on (3.2), this value of σ made h_M roughly equal to $1/2$. The reason for choosing $h_M = 1/2$ in this simulation study is that, given any h in the neighborhood of h_M, there were still several observations used by the kernel estimates in $\mathrm{CV}_l(h)$ and $\mathrm{CV}^*(h)$ even when large values of l and g were considered. In this case, the characteristics of MCV and PCV are more clear. For this combination of ϕ and σ, 1000 independent sets of data were generated. For MCV, the values of l were $0, 1, 2, \ldots, 14$. For PCV, the values of g were $1, 2, \ldots, 15$. For each data set, the values of $d_A(h)$ given in (5.4), $\mathrm{CV}_l(h)$, and $\mathrm{CV}^*(h)$ were calculated on an equally spaced logarithmic grid of 11 values. The endpoints of the grid were chosen to contain essentially all the bandwidths of interest. Here the value of $d_M(h)$ was empirically approximated by averaging $d_A(h)$ over the 1000 pseudo data sets, for each given value of h. The minimizers h_M, $\hat{h}_{\mathrm{MCV}(l)}$ and $\hat{h}^*_{\mathrm{CV}}$ of $d_M(h)$, $\mathrm{CV}_l(h)$ and $\mathrm{CV}^*(h)$, respectively, were calculated. After evaluation on the grid, a one-step interpolation improvement was done, with the results taken as the selected bandwidths. If these functions had multiple minimizers on the grid, the algorithm chose the smallest one, respectively (the choice could be made arbitrarily).

The sample variances, the sample bias-squares, and the sample mean square errors (MSE) of the ratios $\hat{h}/h_M$ were summarized. Here $\hat{h}$ denotes $\hat{h}_{\mathrm{MCV}(l)}$ or $\hat{h}_{\mathrm{PCV}(g)}$ in each respective case. The sample bias-square of $\hat{h}/h_M$ was taken as the square of the average of the 1000 values of $\hat{h}/h_M - 1$. The sum of the sample variance and the sample bias-square was taken as the sample MSE. The numeric results are given in Table 1.

Since the data are positively correlated here, most ordinary cross-validated bandwidths $\hat{h}_{\mathrm{MCV}(0)}$ and $\hat{h}_{\mathrm{PCV}(1)}$ were at the left ends of the bandwidth selection intervals. Hence the sample variances of $\hat{h}_{\mathrm{MCV}(0)}/h_M$ and $\hat{h}_{\mathrm{PCV}(1)}/h_M$ were very small. As the values of l and g increased, the magnitude of both effects of dependence on MCV and PCV decreased and the values of $\hat{h}$ moved away from the left ends of the bandwidth selection intervals. Hence the sample variances of $\hat{h}/h_M$ increased. When the values of l and g were large enough, the characteristics of MCV and PCV appeared, respectively. The bias-squares for $\hat{h}_{\mathrm{MCV}(l)}/h_M$ decreased to 0 as l increased further. However, the bias-squares for $\hat{h}_{\mathrm{PCV}(g)}/h_M$ converged to a nonzero constant as g increased further. In

TABLE 1

The sample variance, bias-square and MSE of $\hat{h}_{\mathrm{MCV}(l)}/h_M$ *and* $\hat{h}_{\mathrm{PCV}(g)}/h_M$ *for the positively correlated data*

Ratios		Variance	Bias-square	MSE
$\hat{h}_{\mathrm{MCV}(l)}/h_M$	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
$\hat{h}_{\mathrm{PCV}(g)}/h_M$	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

contrast to the bias-squares, the variances for $\hat{h}_{\mathrm{MCV}(l)}/h_M$ stayed essentially the same for all l and the variances of $\hat{h}_{\mathrm{PCV}(g)}/h_M$ decreased. In this example, the empirically best values of l and g were nearly the biggest we tried, but note there is little practical difference between these and more intuitively appealing smaller values. Here PCV has smaller overall MSE than MCV (i.e., PCV's effect of reducing variability in the selected bandwidth was stronger than MCV's bias reduction), but this is not generally true. See Section 6.2 of Chu (1989) for examples where the opposite is the case.

5. Sketches of proofs. The following notation and results will be used in this section. Let the notation $X_n = o_u(\rho_n)$ mean that, as $n \to \infty$, $|X_n/\rho_n| \to 0$

almost surely, and uniformly on H_n or $H_{n,g}$ if ρ_n involves $h \in H_n$ or $h \in H_{n,g}$ respectively. For all integers i and j, let Z_i be iid random variables with mean 0 and all finite moments, and a_i and b_{ij} be real numbers such that $\Sigma_{i=-\infty}^{\infty}|a_i| < \infty$ and $\Sigma_{i=-\infty}^{\infty}\Sigma_{j=-\infty}^{\infty}|b_{ij}| < \infty$. Using Theorem 2 of Whittle (1960) and Theorem A of Section 1.4 of Serfling (1980), then, for all positive integers k, we have

$$E\left[\left(\sum_{i=-\infty}^{\infty} a_i Z_i\right)^{2k}\right] \le c_1\left(\sum_{i=-\infty}^{\infty} a_i^2\right)^k, \tag{5.1}$$

$$E\left[\left(\sum_{i=-\infty, i\neq j}^{\infty}\sum_{j=-\infty}^{\infty} b_{ij} Z_i Z_j\right)^{2k}\right] \le c_2\left(\sum_{i=-\infty}^{\infty}\sum_{j=-\infty}^{\infty} b_{ij}^2\right)^k, \tag{5.2}$$

where c_1 and c_2 are constants involving k and moments of Z. For the ARMA process of ε_j given in (A.4), by Theorems 3.1.1 and 3.1.3 of Brockwell and Davis (1987), ε_j can be uniquely expressed as a moving average $\varepsilon_j = \Sigma_{i=0}^{\infty}\psi_i e_{j-i}$, for $j = 1, 2, \ldots, n$, where ψ_i are real numbers with $\Sigma_{i=0}^{\infty}|\psi_i| < \infty$. Combining this with (5.1), (5.2), Fubini's theorem, Minkowski's inequality and Theorem A of Section 1.4 of Serfling (1980), then, for all positive integers k, we have

$$E\left[\left(\sum_{j=1}^{n} \varepsilon_j\right)^{2k}\right] = O(n^k). \tag{5.3}$$

For any $l \ge 0$ 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:

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

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

$$b_j = \left[n^{-1}\sum_{i=1}^{n} K_h(x_j - x_i)\big(m(x_i) - m(x_j)\big)\right]\Big/\left[n^{-1}\sum_{i=1}^{n} K_h(x_j - x_i)\right]$$

$$= \tfrac{1}{2}h^2 m''(x_j)\int u^2 K + o(h^2),$$

$$v_j = \left[n^{-1}\sum_{i=1}^{n} K_h(x_j - x_i)\varepsilon_i\right]\Big/\left[n^{-1}\sum_{i=1}^{n} K_h(x_j - x_i)\right]$$

$$= n^{-1}\sum_{i=1}^{n} K_h(x_j - x_i)\varepsilon_i + o_u\big((nh)^{-7/5}\big) = o_u\big((nh)^{-2/5}\big).$$

PROOF 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 above asymptotic results of b_j and v_j, through a straightforward calculation, then the proof of (3.1) is complete. □

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

$$\begin{aligned}\mathrm{CV}_l(h) &= n^{-1}\sum_{j=1}^{n} \varepsilon_j^2 W(x_j) + D(h) + d_M(h) - 2\,\mathrm{Cross}_l(h)\\ &\quad + \mathrm{Remainder}_l(h),\end{aligned} \tag{5.4}$$

where

$$\begin{aligned}D(h) &= d_A(h) - d_M(h),\\ d_A(h) &= n^{-1}\sum_{j=1}^{n} \big(\hat{m}(x_j) - m(x_j)\big)^2 W(x_j),\\ \mathrm{Cross}_l(h) &= n^{-1}\sum_{j=1}^{n} \varepsilon_j\big(\hat{m}_j(x_j) - m(x_j)\big)W(x_j),\\ \mathrm{Remainder}_l(h) &= n^{-1}\sum_{j=1}^{n} \big(\hat{m}_j(x_j) - \hat{m}(x_j)\big)\\ &\qquad \times\big(\hat{m}_j(x_j) + \hat{m}(x_j) - 2m(x_j)\big)W(x_j).\end{aligned}$$

Using (5.1) through (5.3), and the above asymptotic results of b_j and v_j, through a straightforward calculation, then, as $n \to \infty$,

$$D(h) = o_u(d_M(h)), \tag{5.5}$$

$$\mathrm{Cross}_l(h) = 2(nh)^{-1}\Big(\sum_{k>l}\gamma(k)\Big)K(0)\int W + o_u(d_M(h)), \tag{5.6}$$

$$\mathrm{Remainder}_l(h) = o_u(d_M(h)). \tag{5.7}$$

As $n \to \infty$, $V_{\mathrm{MCV}(l)} > 0$ and $B_2 > 0$, a consequence of (3.1) and (5.4) through (5.7) is that the minimizer of (5.4) can be asymptotically expressed as

$$\hat{h}_{\mathrm{MCV}(l)} = C_{\mathrm{MCV}(l)} n^{-1/5}(1 + o_u(1)).$$

Using the results of (5.4) through (5.7), through a straightforward calculation, then $\mathrm{CV}^*(h)$ can be asymptotically expressed as

$$\mathrm{CV}^*(h) = n^{-1}\sum_{j=1}^{n} \varepsilon_j^2 W(x_j) + V_{\mathrm{PCV}(g)}(\eta h)^{-1} + B_2 h^4 + o_u\big((\eta h)^{-1} + h^4\big).$$

This implies that, as $n \to \infty$, $V_{\mathrm{PCV}(g)} > 0$ and $B_2 > 0$, then the minimizer of $\mathrm{CV}^*(h)$ can be asymptotically expressed as

$$\hat{h}^*_{\mathrm{CV}} = C_{\mathrm{PCV}(g)}\eta^{-1/5}(1 + o_u(1)).$$

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

Using the linear expression of the ARMA process ε_j, asymptotic properties given above and Proposition 6.3.9 of Brockwell and Davis (1987), the proofs of (3.3) and (3.4) of Theorem 1 are essentially the same as the proofs of Theorems 1 and 2 of Härdle, Hall and Marron (1988) and Theorem 1 of Marron (1987), respectively. The only difference is that $\hat{h}_{\mathrm{MCV}(l)}$ should be close to $C_{\mathrm{MCV}(l)}n^{-1/5}$ and $\hat{h}_{\mathrm{PCV}(g)}$ close to $C_{\mathrm{PCV}(g)}n^{-1/5}$, not h_M. The proof of Theorem 1 is complete. □

Acknowledgments. Naomi Altman first suggested the use of PCV for dependent data. We gratefully thank the referees and the editor for their valuable comments which substantially improved the presentation.

REFERENCES

BROCKWELL, P. J. and DAVIS, R. A. (1987). *Time Series: Theory and Methods*. Springer, New York.

CHIU, S.-T. (1989). Bandwidth selection for kernel estimation with correlated noise. *Statist. Probab. Lett.* **8** 347–354.

CHU, C.-K. (1989). Some results in nonparametric regression. Ph.D. dissertation, Dept. Statistics, Univ. North Carolina, Chapel Hill.

CHU, C.-K. and MARRON, J. S. (1990). Choosing a kernel regression estimator. Unpublished manuscript.

DIGGLE, P. J. and HUTCHINSON, M. F. (1989). On spline smoothing with autocorrelated errors. *Austral. J. Statist.* **31** 166–182.

EUBANK, R. L. (1988). *Spline Smoothing and Nonparametric Regression*. North-Holland, Amsterdam.

GASSER, T. and MUELLER, H. G. (1979). Kernel estimation of regression functions. *Smoothing Techniques for Curve Estimation. Lecture Notes in Math.* **757** 23–68. Springer, New York.

GYORFI, L., HÄRDLE, W., SARDA, P. and VIEU, P. (1989). *Nonparametric Curve Estimation from Time Series. Lecture Notes in Statist.* **60**. Springer, New York.

HÄRDLE, W. (1990). *Applied Nonparametric Regression*. Cambridge Univ. Press.

HÄRDLE, W., HALL, P. and MARRON, J. S. (1988). How far are automatically chosen regression smoothing parameters from their optimum? *J. Amer. Statist. Assoc.* **83** 86–101.

HÄRDLE, W. and VIEU, P. (1987). Non parametric kernel regression function estimation for ϕ-mixing observations. I. Optimal squared error estimation. Unpublished manuscript.

HART, J. D. (1991). Kernel regression estimatión with time series errors. *J. Roy. Statist. Soc. Ser. B* **53** 173–187.

HART, J. D. and WEHRLY, T. E. (1986). Kernel regression using repeated measurements data. *J. Amer. Statist. Assoc.* **81** 1080–1088.

MARRON, J. S. (1987). Partitioned cross-validation. *Econometric Rev.* **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 Statist.* **46**. Springer, New York.

NADARAYA, E. A. (1964). On estimating regression. *Theory Probab. Appl.* **9** 141–142.

OLSHEN, R. A. (1967). Asymptotic properties of the periodogram of a discrete stationary process. *J. Appl. Probab.* **4** 508–528.

RICE, J. (1984). Bandwidth choice for nonparametric regression. *Ann. Statist.* **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. Unpublished manuscript.
WATSON, G. S. (1964). Smooth regression analysis. *Sankhyā Ser. A* **26** 359–372.
WHITTLE, P. (1960). Bounds for the moments of linear and quadratic forms in independent variables. *Theory Probab. Appl.* **5** 302–305.

INSTITUTE OF STATISTICS
NATIONAL TSING HUA UNIVERSITY
HSINCHU, TAIWAN 30043
REPUBLIC OF CHINA

DEPARTMENT OF STATISTICS
UNIVERSITY OF NORTH CAROLINA
AT CHAPEL HILL
CHAPEL HILL, NORTH CAROLINA 27514