KERNEL SMOOTHING OF DATA WITH CORRELATED ERRORS

N. S. Altman
Biometrics Unit
Cornell University

This work was supported by a Natural Sciences and Engineering Research Council of Canada postgraduate scholarship, and Hatch Grant 151410 NYF. Much of this work was completed while the author was in the Department of Statistics at Stanford University. The author would like to thank Iain Johnstone for his valuable guidance in supervising the dissertation from which this article was taken. Discussions with Brad Efron, Jerome Friedman, and Peter Lewis also contributed substantially to the ideas in this article.

ABSTRACT

Kernel smoothing is a common method of estimating the mean function in the nonparametric regression model

$$y = f(x) + \varepsilon$$

where $f(x)$ is a smooth deterministic mean function, and ε is an error process with mean zero. In this paper, the mean square error of kernel estimators is computed for processes with correlated errors, and the estimators are shown to be consistent when the sequence of error processes converges to a mixing sequence. The standard techniques for bandwidth selection, such as cross-validation and generalized cross-validation, are shown to perform very badly when the errors are correlated. Standard selection techniques are shown to favor undersmoothing when the correlations are predominantly positive, and oversmoothing when negative. However, the selection criteria can be adjusted to correct for the effect of correlation.

In simulations, the standard selection criteria are shown to behave as predicted. The corrected criteria are shown to be very effective when the correlation function is known. Estimates of correlation based on the data are shown, by simulation, to be sufficiently good for correcting the selection criteria, particularly if the signal to noise ratio is small.

Keywords: mean squared error; kernel regression; autocorrelation; bandwidth; cross-validation; generalized cross-validation.

1.INTRODUCTION

Nonparametric regression techniques have become increasingly popular as tools for data analysis. Because these techniques impose few assumptions about the shape of the mean function, they are extremely flexible tools for uncovering nonlinear relationships between variables.

Many techniques include a smoothing parameter, the bandwidth, which controls the smoothness, bias and variability of the estimate. Various techniques have been developed for determining suitable values of the bandwidth from the data, when the errors are independently and identically distributed (i.i.d.) with finite variance. The purpose of this paper is to explore the properties of a class of nonparameteric regression techniques, kernel smoothers, and the use of the model selection techniques, when the errors are not independent, but instead come from a stationary correlated process.

Figures 1a and 1b show realizations of the process $y = cos(3.15\pi x) + \epsilon$ when the errors come from, respectively, a Gaussian white noise process with unit variance, and an AR(1) process with the same variance and $\rho = .5$. The Gaussian process used in Figure 1a was used to generate the shocks for the AR(1) process in Figure 1b, so the resulting sample paths are very similar. Figures 1c and 1d show kernel estimates of the mean function for this data when the bandwidth was chosen using cross-validation (CV). For the realization with i.i.d. errors, the estimate is quite smooth and captures the main features of the mean function. For the realization with correlated errors, the estimate is far too rough. However, when bandwidth is chosen to minimize the average squared error loss, (Figures 1e and 1f), kernel estimation works almost as well for the correlated error process as for the independent error process.

This paper addresses some of the issues raised by this example. How good are kernel and nearest neighbor smoothers for nonparametric regression estimation when the errors are correlated? How is the performance of standard model selection techniques affected by correlation of the errors? Can better model selection techniques be devised for use with correlated data?

Kernel smoothing is a common method of estimating the mean function in the nonparametric regression model

$$y = f(x) + \epsilon \tag{1}$$

where $f(x)$ is a smooth deterministic mean function, and ε is an error process with mean zero. In this paper, the kernel estimators of Priestley and Chao (1972) are used. These have the form

$$\hat{f}_{\lambda,n}(x) = \sum_{j=0}^{n} w_{\lambda,n}(x,j) y_{n,j} \tag{2}$$

where the weights are

$$w_{\lambda,n}(x,j) = \frac{K(\frac{x-x_{n,j}}{\lambda})}{n\lambda}. \tag{3}$$

K is called the kernel function, λ is a smoothing parameter, called the bandwidth, and n is the sample size.

Only kernels with the following properties are considered:

A) K is symmetric about 0.

B) K has support only on the interval $(-\frac{1}{2}, \frac{1}{2})$.

C) K is Lipschitz continuous of order $\alpha > 0$.

K is called a kernel of order p if the first $p-1$ moments of K are 0, and the p^{th} moment,

$$s_K = \int x^p K(x) dx \tag{4}$$

is not zero. The squared norm of K

$$W_K = \int K^2(x) dx \tag{5}$$

is also needed.

2.MODELS

For sample size n, the observations $y_{n,1} \dots y_{n,n}$ are assumed to be generated by the nonparametric regression model (1), with observations taken at regularly spaced design points $\frac{i}{n}$. The errors are assumed to come from a stationary process with covariance function

$$E(\varepsilon_{n,i}, \varepsilon_{n,j}) = \sigma^2 \rho_n(|i-j|). \tag{6}$$

where the variance, σ^2, is independent of n and $\rho_n(k)$ is a correlation function depending on n. The variance matrix of the errors will be denoted by Σ_n. This formulation of the covariance function allows the autocorrelation among the errors to vary both with the distance between the design points, and the sample size.

An important special case of this model is the process with $\rho_n(i) = \rho(\frac{i}{n})$ where $\rho(x)$ is continuous. Then the error process is a realization of a continuous parameter process on $[0, 1]$. This process has been discussed by Hart and Wehrly (1986) and Parzen (1959, 1961). An important result from these papers is that, if only a single realization of the process has been observed, there are no consistent linear estimators of the mean function as the design points are sampled more and more densely on the unit interval. Parzen's results show that the only unbiased linear estimator of $f(x)$ is y_x (with variance σ^2). Hart and Wehrly show that kernel estimators converge to random variables under this model, and that considerable improvement (in terms of mean squared error) can be made by using kernel estimators with $\lambda > 0$.

A second important special case is the process with $\rho_n(i) = \rho(i)$. In this case, the error process is constant, regardless of how close together the design points become. Models intermediate between the continuous parameter process and this process are also allowed by model (6).

The results in this paper require absolute summability conditions, D and E on the sequence of correlation functions either:

D) $\sum_i^{\frac{k}{2}} |\rho_n(i)|$ converges as n and $k \to \infty$.

This condition, which is common in time series analysis, ensures that observations sufficiently far apart are essentially uncorrelated. The following condition is also needed:

E) $\sum_{i=1}^{\frac{k}{2}} i\rho_n(i) = o(k)$ as n and $k \to \infty$.

Under condition D, the sum of the correlations, S_ρ, is well-defined. Let $S_{\rho_n}(k) = \sum_{j=1}^{\frac{k}{2}} \rho_n(j)$.

$$S_\rho = lim_{n,k\rightarrow\infty} S_{\rho_n}(k) \tag{7}$$

3.MEAN SQUARED ERROR

When the errors are assumed to have finite second moments, the MSE, defined by

$$MSE(x, \lambda, n) = E(\hat{f}_{\lambda,n}(x) - f(x))^2 \tag{8}$$

is often used as a goodness of fit criterion and as a means of assessing the asymptotic properties of the estimators. The optimal smoothing parameter is often considered to be the one which minimizes the MSE totalled, or equivalently, averaged, over the design points.

Let Σ_n denote the covariance of the observation, $u(\bullet)$ denote the column vector $u(j)$ and $u'(\bullet)$ the transpose of $u(\bullet)$. Simple algebra gives us

$$MSE(x,\lambda,n) = \left(w'_{\lambda,n}(x,\bullet)f(\bullet) - f(x)\right)^2 + w'_{\lambda,n}(x,\bullet)\Sigma_n\, w_{\lambda,n}(x,\bullet). \tag{9}$$

Notice that the bias,$w'_{\lambda,n}(x,\bullet)f(\bullet)-f(x)$ depends on the sample size only via the selection of design points and is not affected by the correlation structure. For mean functions with at least p derivatives, and kernels of order p, Gasser and Müller (1979) computed the asymptotic form of the bias (when the design points become dense on the interval) to be

$$w'_{\lambda,n}(x,\bullet)f(\bullet) - f(x) = (-\lambda)^p s_K f^{(p)}(x)/p! + o(\lambda^p) + o(\frac{1}{n\lambda}) \tag{10}$$

when $\lambda \rightarrow 0$ and $n\lambda \rightarrow \infty$ and $\frac{\lambda}{2} < x < 1 - \frac{\lambda}{2}$. The estimators are asymptotically unbiased under these conditions.

Correlation of the errors can affect the variance term, $w'_{\lambda,n}(x,\bullet)\Sigma_n\, w_{\lambda,n}(x,\bullet)$ very strongly. When the errors are i.i.d., the variance term is $\sigma^2 \parallel w_{\lambda,n}(x,\bullet) \parallel^2$, where $\parallel \bullet \parallel$ is Euclidean norm. As the design points become dense on the interval, $\parallel w_{\lambda,n}(x,\bullet) \parallel^2 \rightarrow \frac{W_K}{n\lambda}$, so the variance function decreases as $O(\frac{1}{n\lambda})$ regardless of the shape of the kernel. When the errors are correlated, the behavior of the variance term as a function of the bandwidth depends on both the correlation function and the kernel.

In Theorem 1, $MSE(x,\lambda,n)$is computed. It is shown that, under fairly general conditions on the kernel and the correlation function, $w'_{\lambda,n}(x,\bullet)\Sigma_n\, w_{\lambda,n}(x,\bullet) \approx \sigma^2 \frac{W_K}{n\lambda}(1+2S_\rho)$ where S_ρ was defined by equation (7).

Theorem 1: Suppose the mean function f, has p derivatives and, the kernel function K is of order p and satisfies conditions A – C. Suppose the design points are equally spaced on the interval of estimation, and the errors are stationary and satisfy conditions D and E.

Then for $\frac{\lambda}{2} \leq x \leq 1 - \frac{\lambda}{2}$

$$MSE(x,\lambda,n) = \left(\lambda^p f^{(p)}(x) s_K/p!\right)^2 + \sigma^2 \frac{W_K}{n\lambda}(1+2S_\rho) + o(\frac{1}{n\lambda}) + o(\lambda^{2p}). \tag{11}$$

The proof is in the Appendix section A.1.

At boundary points (points distance $q\lambda$ away from an endpoint, with $q < \frac{1}{2}$,) the kernel function is truncated, introducing considerable bias into the regression estimate. Modifications to the kernel function to maintain the same order of bias and variance (in the i.i.d case) have been suggested (Gasser and Müller 1979; Rice, 1984b). If the "boundary kernel," $K^q(x)$, satisfies conditions (A.1) – (A.3) in the appendix, then

$$w'_{\lambda,n}(x,\bullet)\Sigma_n\, w_{\lambda,n}(x,\bullet) = \sigma^2 \frac{W_{K^q}}{n\lambda}(1+2S_\rho) + o(\frac{1}{n\lambda}).$$

Corollary (1.1):Under the conditions of Theorem 1, the asymptotically optimal bandwidth at x is

$$\lambda_n = \left(\frac{(p!)^2\sigma^2 W_K(1+2S_\rho)}{p s_K^2 (f^{(p)}(x))^2}\right)^{\frac{1}{2p+1}} n^{-\frac{1}{2p+1}}. \tag{12}$$

Corollary (1.2): Let z_x be the process generated by

$$z_x = f(x) + u_x$$

where the errors u_x are i.i.d. with variance $\sigma^2(1+2S_\rho)$, and z has the same mean function as y. Under the conditions of Theorem 1, asymptotically, as $\lambda \to 0$ and $n\lambda \to \infty$,

$$\frac{MSE_y(x,\lambda,n)}{MSE_z(x,\lambda,n)} \to 1.$$

Corollary (1.3): Under the conditions of Theorem 1, kernel estimators are consistent (as $\lambda \to 0$ and $n\lambda \to \infty$).

The corollaries follow simply from Theorem 1.

Corollaries 1.1 and 1.2 provide simple means of comparing the correlated process with variance σ^2 to the i.i.d. process with the same variance, σ^2. Let $v_x = f(x) + \eta_x$ where η_x is i.i.d. with variance σ^2. Then, if $S_\rho > 0 (< 0)$, the optimal bandwidth for z (and hence for y) is greater (smaller) than the optimal bandwidth for v, and the MSE achieved at the optimal bandwidth is also greater (smaller).

Corollary 1.2 also shows that the results of Gasser and Müller (1979) about the shapes of optimal kernels continue to hold when the errors are correlated.

4. SELECTING A SMOOTHING PARAMETER

For a given, finite set of observations, choice of an effective smoothing parameter is of considerable interest. A "good" value of the smoothing parameter will result in a small value of $MSE(x, \lambda, n)$.

Several criteria based on the data have been used for bandwidth selection. Those most commonly used are CV, (Allen 1974; Geisser 1975; Stone 1974), generalized cross-validation, GCV (Craven and Wahba 1979), and Mallows C_L (Mallows 1973). The properties of these criteria, including convergence of the smoothing parameter chosen by one of the selection criteria to the truly optimal value, and the asymptotic equivalence of the the criteria have been explored in some detail by the authors above, as well as by others, (for example: Efron 1986; Härdle, Hall and Marron 1987; Härdle and Marron 1985; Li 1984, 1985; Stone 1977).

Mallows' C_L, CV and GCV can all be viewed as estimators of expected squared prediction error (ESPE) based on a correction to the observed squared residual. The errors in the new observations are independent of errors in the original observations, so :

$$\begin{aligned} ESPE(x, \lambda, n) &= E\left(y_{new}(x) - \hat{f}_{\lambda,n}(x)\right)^2 \\ &= \sigma^2 + MSE(x, \lambda, n). \end{aligned} \tag{13}$$

The squared residual, $r^2(i, \lambda, n) = \left(y_{n,i} - \hat{f}_{\lambda,n}(\frac{i}{n})\right)^2$ is biased as an estimator of $ESPE(\frac{i}{n}, \lambda, n)$, as it has expectation

$$E(r^2(i, \lambda, n)) = \sigma^2 + MSE(\frac{i}{n}, \lambda, n) - 2\sigma^2 w_{\lambda,n}(\frac{i}{n}, i) - V_2(\frac{i}{n}, \lambda, n). \tag{14}$$

The term $2\sigma^2 w_{\lambda,n}(\frac{i}{n}, i)$ arises because $y_{n,i}$ is both a term in the estimator, $\hat{f}_{\lambda,n}(\frac{i}{n})$, and the estimate of $y_{new}(\frac{i}{n})$. The additional variance term,

$$V_2(\frac{i}{n}, \lambda, n) = 2\sigma^2 \sum_{j \neq 0} w_{\lambda,n}(\frac{i}{n}, i+j)\rho_n(j). \tag{15}$$

arises because of the correlation between $\varepsilon_{n,i}$ and the other errors.

Remark: Note that in the region $\frac{\lambda}{2} < \frac{i}{n} < 1 - \frac{\lambda}{2}$ the weights do not depend on i and so both $w_{\lambda,n}(\frac{i}{n}, i)$ and $V_2(\frac{i}{n}, \lambda, n)$ are independent of i.

Mallow's C_L is defined by

$$r^2_{CL}(i, \lambda, n) = r^2(i, \lambda, n) + 2\hat{\sigma}^2 w_{\lambda,n}(\frac{i}{n}, i) \tag{16}$$

where $\hat{\sigma}^2$ is some unbiased estimator of σ^2. (In Mallows' original paper, the criterion is divided by $\hat{\sigma}^2$.) For bandwidth selection, the criterion is usually totalled over all the design points. However, the theoretical computations in this section are done pointwise.

CV and GCV are criteria which are asymptotically equivalent to C_L (when the errors are i.i.d. and the design points are equally spaced on the interval) but which do not require an estimate of σ^2. The CV criterion is

$$r^2_{CV}(i, \lambda, n) = \frac{r^2(i, \lambda, n)}{\left(1 - w_{\lambda,n}(\frac{i}{n}, i)\right)^2}. \tag{17}$$

CV is based on a simple heuristic for estimating ESPE: Estimate $y_{new}(i)$ by $y_{n,i}$ and leave $y_{n,i}$ out of the estimator.

Generalized cross-validation was proposed by Craven and Wahba (1978) as an adjustment to cross-validation that is more nearly unbiased for ESPE in the case of unequally spaced points, if the design points are considered to be fixed. The GCV criterion is

$$r^2_{GCV}(i, \lambda, n) = \frac{r^2(i, \lambda, n)}{\left(1 - \frac{1}{n} tr W_{\lambda,n}\right)^2} \tag{18}$$

where $W_{\lambda,n}$ is the matrix $[w_{\lambda,n}(\frac{i}{n}, j)]$ and $tr W_{\lambda,n} = \sum_{i=0}^{n} w_{\lambda,n}(\frac{i}{n}, i)$. If λ is small, $\frac{tr W_{\lambda,n}}{n} \approx w_{\lambda,n}(\frac{i}{n}, i) \approx \frac{K(0)}{n\lambda}$ so CV and GCV differ very little. A Taylor series expansion of (17) or (18) gives the asymptotic equivalence of the 3 criteria:

$$E(r^2_{(G)CV}(i, \lambda, n) = \sigma^2 + MSE(\frac{i}{n}, \lambda, n) - V_2(\frac{i}{n}, \lambda, n) + o(\lambda^{2p}) + o(\frac{1}{n\lambda}). \tag{19}$$

So, asymptotically, C_L , CV, and GCV have the same expectation for equally spaced design points. Theorem 2 describes the behavior of this expectation.

Lemma 2.1: If the kernel function satisfies conditions A–C, and the correlation function satisfies conditions D and E, then for $\frac{\lambda}{2} < x < 1 - \frac{\lambda}{2}$,

$$V_2(\frac{i}{n}, \lambda, n) = 4\sigma^2 \frac{K(0)}{n\lambda} S_\rho + o(\frac{1}{n\lambda}).$$

The proof is in the Appendix section A.2.

Theorem 2: Under the conditions of Theorem 1, for $\frac{\lambda}{2} < x < 1 - \frac{\lambda}{2}$

$$E(r_C^2(i, \lambda, n)) = (\lambda^p f^{(p)}(x) \frac{s_K}{p!})^2 + \sigma^2 \frac{W_K}{n\lambda}(1 + 2S_\rho) - 4\sigma^2 \frac{K(0)}{n\lambda} S_\rho \\ + o(\frac{1}{n\lambda}) + o(\lambda^{2p}). \tag{20}$$

where C is one of C_L , CV or GCV.

Proof: The result follows simply from Theorem 1, and Lemma 2.1.

For points x, distance $q\lambda$ from the endpoints, K^q may be substituted for K in the variance terms. The bias term depends on the order of the boundary kernel, but is ***the same as*** the bias term of $MSE(\frac{i}{n}, \lambda, n)$

Härdle, Hall and Marron, (1987), show that, for i.i.d. errors, the bandwidths selected by minimizing squared error, MSE, and the bandwidth selection criteria are asymptotically equivalent in the sense that the ratio of selected bandwidths tends in probability to 1. Corollary 1.2 suggests that in the correlated case, the bandwidth selected by minimising MSE should tend to the bandwidth which is optimal for an i.i.d.process with variance $\sigma^2(1 + 2S_\rho)$. However, Theorem 2 suggests that, when $1 + 2S_\rho(1 - 2K(0)/W_K) > 0$ the bandwidth chosen by the selection criteria should tend to the bandwidth which is optimal for an i.i.d.process with variance $\sigma^2(1 + 2S_\rho(1 - 2\frac{K(0)}{W_K}))$. So, for kernels with $2K(0) > W_K$, when $S_\rho < 0$ (> 0), the criteria tend to choose bandwidths that are too large (small). If $2K(0) > W_K$ (which is the case for all kernels in common use) and S_ρ is sufficiently large, $1 + 2S_\rho(1 - 2K(0)/W_K)$ can be negative. In this case, the expectation of r_C^2 is strictly increasing with λ, and, asymptotically, the bandwidth selection criteria should tend to select interpolation. Hart (1986) shows, under conditions similar to these, that with probability tending to 1, CV picks arbitrarily small bandwidths.

5.CORRECTING FOR CORRELATION

In this section, two methods are suggested for correcting the selection criteria when the correlation function is known. The direct method adjusts the ***criteria*** to make them more nearly unbiased for ESPE. The indirect method transforms the ***residuals*** to produce transformed residuals which are less correlated.

If the correlation function, ρ_n, is known, with corresponding correlation matrix, R_n, Mallow's C_L can be corrected to be an unbiased estimator of ESPE.

From equation (16) an appropriate adjustment for Mallow's C_L criterion is

$$r^2_{CL,\rho}(i,\lambda,n) = r^2(i,\lambda,n) + 2\hat{\sigma}^2 \sum_{j=-[\frac{n\lambda}{2}]}^{[\frac{n\lambda}{2}]} w_{\lambda,n}(\frac{i}{n}, i+j)\rho_n(j). \tag{21}$$

Corresponding adjustments for CV and GCV are intended to match the low order terms in the Taylor series expansion in equation (19) to the adjusted C_L criterion in equation (21). One way to do this is to set

$$r^2_{CV,\rho}(i,\lambda,n) = \frac{r^2(i,\lambda,n)}{(1-\sum_{j=-[\frac{n\lambda}{2}]}^{[\frac{n\lambda}{2}]} w_{\lambda,n}(\frac{i}{n}, i+j)\rho_n(j))^2} \tag{22}$$

and

$$r^2_{GCV,\rho}(i,\lambda,n) = \frac{r^2(i,\lambda,n)}{(1-\frac{1}{n}\mathrm{tr}W_{\lambda,n}R_n)^2} \tag{23}$$

We will call this the direct method of correcting for correlation, and denote the corresponding bandwidth selection criteria by CV_ρ and GCV_ρ respectively.

Another approach to the problem when the correlation matrix is known, is to compute the transformed residuals: $r_{\rho^{-1}}(\bullet,\lambda,n) = R_n^{-\frac{1}{2}} r(\bullet,\lambda,n)$. This has been used with some success in the context of spline smoothing with normal AR(1) errors, (Diggle 1985; Diggle and Hutchinson 1985; Engle, Granger, Rice, Weiss, 1986). The goodness of fit criterion is then the total weighted MSE,

$$TSE_{\rho^{-1}}(\lambda,n) = E(\hat{f}_{\lambda,n}(\bullet) - f(\bullet))' R_n^{-1}(\hat{f}_{\lambda,n}(\bullet) - f(\bullet)). \tag{24}$$

The totalled C_L criterion based on the transformed residuals,

$$\sum r^2_{CL,\rho^{-1}} = \sum_{i=0}^{n} r^2_{\rho^{-1}}(i,\lambda,n) + 2\hat{\sigma}^2 \text{ tr } W_{\lambda,n}, \tag{25}$$

is then asymptotically unbiased for the expected value of the prediction sum of squares for the transformed residuals.

The totalled CV and GCV criteria based on the transformed residuals can also be readily defined. They are

$$\sum r^2_{CV,\rho^{-1}}(\lambda,n) = \sum_{i=0}^{n} \frac{r^2_{\rho^{-1}}(i,\lambda,n)}{(1-w_{\lambda,n}(\frac{i}{n},i))^2}. \tag{26}$$

and

$$\sum r^2_{GCV,\rho^{-1}}(\lambda,n) = \frac{\sum_{i=0}^{n} r^2_{\rho^{-1}}(i,\lambda,n)}{(1-\frac{1}{n}\text{ tr } W_{\lambda,n})^2}. \tag{27}$$

We will call this the indirect method of correcting for correlation, and denote the corresponding bandwidth selection criteria by $\sum CV_{\rho^{-1}}$ and $\sum GCV_{\rho^{-1}}$ respectively. Although these criteria are based on a somewhat different goodness of fit statistic than the others, the simulations described in section 7 show that they lead to estimators with comparable unweighted TSE.

6.ESTIMATING THE CORRELATION FUNCTION

Usually the correlation function is unknown and must be estimated from the data. In the context of growth curves, where many curves were expected to have similar error structures, Hart and Wehrly (1986) successfully used the standard method of moments (MM) estimators for the correlations, averaged over the curves. Estimating the autocorrelations from a single realization is more challenging.

A simple approach is to smooth the data, compute the low-order sample autocorrelations of the residuals and fit an autoregressive-moving average (ARMA) model (Box and Jenkins 1976). A more detailed look at these methods is in Altman (1988).

Theorem 3 below, shows that the MM estimator of $\rho_n(s)$ is consistent under mild regularity conditions on the errors.

Theorem 3: Suppose the data and kernel satisfy the conditions of Theorem 1 and the mean function has p^{th} derivative which is Lipschitz of order $\gamma > 0$. Suppose, as well, the errors satisfy the following regularity condition:

$$\begin{array}{ll} \text{F) } \epsilon_{n,t} = \sum_{j=-\infty}^{\infty} \psi_j z_{t-j} & \text{with } \sum_{j=-\infty}^{\infty} |\psi_j| < \infty \\ \quad z_t \text{ i.i.d.} & \text{with } E(z_t) = 0 \\ & \quad E(z_t^2) = \sigma^2 \\ & \quad E(z_t^4) < \infty \end{array}$$

For fixed s, n and λ, define the method of moments estimator of $\rho(s)$ by:

$$\hat{\rho}_n(s,\lambda) = \frac{\sum_{i=[\frac{n\lambda}{2}]}^{n+1-[\frac{n\lambda}{2}]-s} r(i,\lambda,n) r(i+s,\lambda,n)}{\sum_{i=[\frac{n\lambda}{2}]}^{n+1-[\frac{n\lambda}{2}]-s} r^2(i,\lambda,n)}. \tag{28}$$

Then $\hat{\rho}_{\lambda,n}(s)$ is asymptotically normal and, as $\lambda \to 0$ and $n\lambda \to \infty$,

$$E(\hat{\rho}_{\lambda,n}(s)) = \frac{\rho(s) + \lambda^{2p}\left(\frac{s_K}{p!}\right)^2 \frac{\int f^{(p)}(x)^2 dx}{\sigma^2} + \frac{(1+2S_\rho)}{n\lambda}(W_K - 2K(0))}{1 + \lambda^{2p}\left(\frac{s_K}{p!}\right)^2 \frac{\int f^{(p)}(x)^2 dx}{\sigma^2} + \frac{(1+2S_\rho)}{n\lambda}(W_K - 2K(0))}$$
$$+ o(\lambda^{2p}) + o(\frac{1}{n\lambda}) + o(\frac{s}{n}). \tag{29}$$

and

$$Var(\hat{\rho}_{\lambda,n}(s)) = o(\frac{1}{n\lambda})$$

.

The proof is in Altman (1989).

It is easy to show from Theorem 3 that, if $2K(0) \geq W_K$, then $\hat{\rho}_n(1,\lambda)$ has bias which is increasing in λ, and the signal to noise ratio, $\frac{\int f^{(p)}(x)^2 dx}{\sigma^2}$, and is decreasing in S_ρ.

7.SIMULATION RESULTS

A simulation study was carried out to test the theoretical results of this paper. For each experiment, three second order kernels and a fourth order kernel were used. The definitions of the kernels are displayed in Table 1. The fourth order"spline" kernel is a truncated version of the effective kernel for the cubic smoothing spline with equally spaced design points (Silverman 1984, 1985).

Table 1. Kernels

Kernel	equation	$K(0)$	W_K
uniform	1	1	1
triangular	$4(\frac{1}{2} - \lvert x \rvert)$	2	1.333
quadratic	$6(\frac{1}{4} - x^2)$	1.5	1.2
spline	$8exp(-\lvert x \rvert/16\sqrt{2})sin(\lvert x \rvert/16\sqrt{2} + \pi/4)$	$4\sqrt{2}$	$3\sqrt{2}$

At the boundary, the kernels were adjusted by reweighting to $w^*_{\lambda,n}(\frac{i}{n},j) = w_{\lambda,n}(\frac{i}{n},j)/\sum w_{\lambda,n}(\frac{i}{n},k)$. While reweighting is not as good as the use of a boundary kernel for bias reduction, (Gasser and Müller 1979, Rice 1984b) the added bias is the same for MSE and for the bandwidth selection criteria.

A small preliminary study showed, empirically, that, for the sample sizes and the mean function used in this study, and for i.i.d. errors, the optimal bandwidths for the uniform, triangular and quadratic kernels were about the same. The optimal bandwidth for the spline kernel was about three times as large so, when bandwidth λ was used for the other kernels, bandwidth 3λ was used with the spline kernel.

The mean function for all three experiments was

$$f(x) = \cos(3.15\pi x) \text{ for } x \text{ in } [0,1].$$

The errors were generated from a Gaussian $AR(1)$ process, with $\rho(1) = -.9, -.6, -.3, 0, .1, .2, .3, .6, .9$. At each value of $\rho(1)$, 50 realizations of the process were generated. As has been noticed by others (Wendelberger 1987), when the errors are correlated the bandwidth selection criteria often have many local minima for a given realization. So the minimizing bandwidths were selected from the 15 values: .02, .04, .06, .08, .10, .12, .16, .20, .24, .28, .32, .38, .42, .46, and .50. With each value of $\rho(1)$, three combinations of sample size and variance were used, $n = 128$ and $\sigma^2 = .01$ and 1.0, and $n = 256$ and $\sigma^2 = 1.0$. Each set of simulations was repeated 50 times.

Table 2 is a summary of the numerical results for $n = 128$ for the uniform and spline equivalent kernels. Results for $n = 256$ were similar. Results for the triangular and quadratic kernels were similar to those for the uniform kernel. Results for CV and GCV were practically identical (since the design points are equally spaced) and so only GCV is listed here. The average squared error (ASE), rather than the total squared error, is used here to standardize for sample size.

The entries in the table are

min ASE	The actual minimum of the average squared error for the realization
min GCV	The value of the ASE at the bandwidth minimizing (standard) $\sum GCV$
min $GCV_{\rho-1}$	The value of the ASE at the bandwidth minimizing $\sum GCV_{\rho-1}$
min GCV_{ρ}	The value of the ASE at the bandwidth minimizing $\sum GCV_{\rho}$
min $GCV_{\tilde{\rho}-1}$	The value of the ASE at the bandwidth minimizing $\sum GCV_{\tilde{\rho}-1}$
min $GCV_{\tilde{\rho}}$	The value of the ASE at the bandwidth minimizing $\sum GCV_{\tilde{\rho}}$

The minimum ASE is tabulated as a raw value. (For $\sigma^2 = 1.0$, the ASE has been multiplied by 100, and for $\sigma^2 = .01$, by 10000.) The other entries are expressed as a multiple of ASE. The percentage is estimated for each realization, and the median over the 50 realizations is reported. Bracketing each median, in smaller print, are the quartiles. The last two entries are corrected criteria using $\tilde{\rho}$, an estimate of the correlation computed from the residuals.

Figure 2 contains plots of TSE, $\sum GCV - n\sigma^2$, $\sum GCV_\rho - n\sigma^2$, and $\sum GCV_{\rho-1} - n\sigma^2$ versus bandwidth for the uniform kernel with $n = 128$ and $\sigma^2 = 1.0$, where $\sum GCV$, $\sum GCV_\rho$, and $\sum GCV_{\rho-1}$ are, respectively, the sum of the uncorrected, and directly and indirectly corrected GCV criteria defined by (18), (23), and (27). (The plots for the other kernels, and other values of n and σ^2 are similar, as are the plots for CV.) The minimum is marked on each curve with an asterisk.

The large effect of correlation on both the magnitude of the TSE and the location of the minimum can clearly be seen. Also notable is that the TSE curves are quite flat near their minima for large values of $\rho(1)$. The large differences between the TSE curves is due solely to the effect of the correlation on the variance term.

Correlation has a large effect on the performance of the bandwidth selection criteria. The differences in TSE at the truly optimizing bandwidth, and the bandwidths picked by CV and GCV are notable. As predicted by Theorem 1, the criteria pick bandwidth too large when

$\rho(1)$ is negative, and too small when $\rho(1)$ is positive. For $\rho(1) \geq .3$, the criteria tend to pick interpolation, no matter what the variance or sample size. However, when either direct and indirect corrections were made to the bandwidth selection criteria using the true correlation coefficients, the corrected criteria performed well.

As can be seen in Figure 2, $\sum GCV_\rho$ has much the same shape as the TSE curve, while $\sum GCV_{\rho^{-1}}$ has a somewhat different shape. However, all three curves have minima that are very close together for negative values of $\rho(1)$. For positive $\rho(1)$, $\sum GCV_{\rho^{-1}}$ has a minimum at bandwidths that are much smaller than the optimal bandwidths for unweighted TSE. However, as the TSE curve is quite flat near its minimum, the resulting estimator has similar TSE. This can be seen in Table 2.

7.1 Performance of Corrected Bandwidth Selection Criteria with the Estimated Correlation Function

As a final step, the correlations were estimated from the data. The estimated correlations were then used to compute $\sum GCV_\rho$ and $\sum GCV_{\rho^{-1}}$ which were used to select the bandwidth for the final smooth.

In each case, the same kernel was used to estimate the correlations and the mean function. The correlations were estimated from the residuals from the kernel smooth with bandwidth .25. Three estimators of correlation were tried, $\hat{\rho}_{n,\lambda}(s)$, $\hat{\rho}^M_{n,\lambda}(s)$, a MM estimator based on the centered residuals $r(i,\lambda,n) - \frac{1}{n}\sum r(j,\lambda,n)$, $\hat{\rho}^M_{n,\lambda}(s)$, and $\tilde{\rho}_{n,\lambda}(s)$ a "running" version of $\hat{\rho}^M(s,\lambda)$ using 1/4 of the data. However, as mean of the residuals was close to zero for these trials, all three estimators produced very similar results. Only the results for $\tilde{\rho}_{n,\lambda}(s)$ are presented here.

While estimates of individual correlation coefficients are consistent, the sum of the estimated coefficients is not necessarily a good estimate of S_ρ (since the sample periodogram is not a consistent estimator of the the spectrum). An approach which appears, empirically, to have some promise, is to compute a few, low-order autocorrelations, and then to fit a low order autoregressive-moving-average (ARMA) model (Box and Jenkins 1976). For these simulations, the first two correlations were estimated directly and constrained by truncation to lie in the region of stationarity for an AR(2) process. The correlations at higher lags were generated from the AR(2) process with these two values at the first 2 lags.

The direct correction was based on the AR(2) estimate of the correlation function. Since the errors are actually AR(1), this procedure introduces additional error in estimating the correlation. However, in a real data situation, the order of the error process will not be known, so this method provides a fair assessment of how estimating the correlations will perform in practise.

For computational simplicity, the indirect correction was based on the partially differenced residuals, $r(i+1,\lambda,n)-\tilde{\rho}_{n,\lambda}(1)r(i,\lambda,n)$, which is appropriate when the errors are AR(1). This means that the simulation method is somewhat biased in favor of the indirect correction. This becomes apparent when the signal to noise ratio is large, due to the large error in estimating the correlations in this case.

The results of Theorem 3 show that, if $2K(0) \geq W_K$, then the bias in estimating the correlation is increasing in the signal to noise ratio, and is decreasing in S_ρ. As predicted, the estimates of correlation were more accurate when the signal to noise ratio was larger, and the correlations were smaller. The simulation results for estimating the correlation function are reported in full in Altman 1989. In general, the estimates of correlation were good when $\sigma^2 = 1.0$ and poor when $\sigma^2 = .01$. However, even when $\sigma^2 = .01$, the estimates appear to be adequate to correct the bandwidth selection criteria.

From Table 2, when $\sigma^2 = 1.0$, GCV_ρ and $GCV_{\rho^{-1}}$ both perform well. When the correlations are computed from the data, the corrections are, in general, not as good, but are an improvement over ordinary GCV. However, for large negative ρ, the estimated correlations perform somewhat better than the true values, possibly indicating that the corrections suggested by the asymptotic expansion are too extreme for samples of this size.

When $\sigma^2 = .01$, the criteria corrected with the true value of the correlations perform quite well, and are improvements over uncorrected GCV. Despite the large positive bias in estimating the correlations, $GCV_{\rho^{-1}}$ performs very well, even for negative correlations. However, due to the large upwards bias of $\tilde{\rho}(2,.25,128)$, GCV_ρ performs poorly when the correlation is negative, and is not as good as the uncorrected criterion in this case.

Of course, in real use, the correlations are never known. The real choice is between using the uncorrected criteria, or estimating the correlations and using the estimates to make direct corrections to the selection criteria. This study suggests that, despite the poor quality of the correlation estimates, corrections to the selection criteria considerably improve the choice of bandwidth, and thus the quality of the estimate of the mean function.

The simulation results show that even very rough estimates of the correlation function based on the data can be quite effective in correcting the bandwidth selection criteria for the effects of correlation. The noisier the data, the better the estimate of the correlation function, and the more important the effects of choosing the correct bandwidth for the smooth.

8. EXAMPLE: SEA SURFACE TEMPERATURE DATA

The methodology of the previous section was applied to a data set of 4380 sea surface temperatures collected daily at Granite Canyon, California (Breaker and Lewis, 1988). This is a complicated data set, with larger events, called El Niño episodes, superposed on asymmetric seasonal and shorter term periodic effects. The correlation function of the errors was estimated from the data. The estimated correlations were used to compute GCV_ρ and the data was smoothed with the selected bandwidth.

The uniform and quadratic kernels were used. The bandwidths were constrained to lie between 3 days and half a year and GCV_ρ was computed at 50 evenly spaced bandwidths in this interval. No attempt was made to estimate, or adjust for, any seasonal effects. The correlations were estimated using a running MM estimate based on the residuals from a smooth with bandwidth of 73 days. The first two autocorrelations were computed directly, and the rest estimated from these according to the law of an AR(2) process. (A more sophisticated approach would be to estimate several autocorrelations, and fit a low order ARMA model.)

Using the uniform kernel, the first two estimated correlations were 0.85 and 0.69. GCV_ρ selected 87 days, or about a quarter of a year. Using the quadratic kernel, the first two estimated correlations were 0.82 and .65. Despite the difference in the estimated correlations, the bandwidth selected by GCV_ρ was very similar – 93 days. For both kernels, ordinary GCV selected the smallest bandwidth allowed by the program – three days – which produced almost no smoothing.

Figure 3 is a plot of the raw data, the quadratic smooth and the residuals from the smooth. The quadratic smooth is somewhat smoother than running means. Both smooths show the long term trends in the data that are obscured by the variability in the raw data. The seasonal effects in the data are clearly seen in the smooth. The larger peaks in 1973 and 1977 are El Niño events. There was also an El Niño in 1979, which consists of several sharp peaks, and does not appear in the smooth. The residuals appear to have a stationary mean, but there is some suggestion that the variance varies with the mean temperature.

9.CONCLUSIONS

The increasing popularity of smoothing techniques is evident from the explosion of papers on the topic in statistical journals, (see for example, the bibliography by Collomb, 1985) and the increasing use of data smooths in the applied literature (for example, Engle et al 1986; Williams et al 1985). Smoothing is also used as a computational subprocedure in other data analytic techniques such as projection pursuit (Friedman and Stuetzle 1981; Huber 1985) and the alternating conditional expectations (ACE) method of computing the maximal correlation (Brieman and Friedman 1985).

This paper shows that kernel regression performs well for data with correlated errors, as long as the correlations are sufficiently short-term. If an appropriate bandwidth must be chosen from the data, the bandwidth selection criteria must be suitably corrected for the correlation. MM estimates of the correlation, based on residuals from a preliminary smooth, may be used for this purpose.

The evidence from the simulations done here, as well as those reported by Hart (1989), suggest that corrections to the bandwidth selection criteria based on even rough estimates of the correlation function are considerable improvements on the standard criteria. Therefore, the bandwidth for the preliminary smooth need not be too accurate. Equation (20) and the discussion following, show that for kernels with $2K(0) > W_K$, overestimation of the correlations is less serious than underestimation (which may lead to interpolation). Therefore, it is preferable for the preliminary estimate to oversmooth the data. When the signal to noise ratio is very small, the bias in estimating correlation is mainly due to the correlation between the observations and the estimates. In this case, the bandwidth for the preliminary smooth should be chosen very large – use of $y_{n,i} - \bar{y}_n$, corresponding to a very large λ may be preferred. When the signal to noise ratio is very small, a preliminary estimate of a bandwidth which mildly oversmooths can readily be obtained by inspection.

Smoothing data with correlated errors is a useful technique for estimating an unknown, nonlinear mean function. If a rough estimate of the correlation function can be made, bandwidth selection can be done automatically. If no estimate of the correlations is available, an effective bandwidth can be estimated subjectively by viewing smooths computed at several different bandwidths. In either case, smoothing offers considerable variance reduction compared to interpolation, and is more flexible than parametric techniques.

APPENDIX: PROOF OF THEOREMS

A.1 Proof of Theorem 1

Lemma A.1: If K satisfies condition C then

$$\left| ||w_{\lambda,n}(\frac{i}{n},\bullet)||^2 - \frac{\int K^2(x)dx}{n\lambda} \right| = o(\frac{1}{n\lambda}). \qquad (A.1)$$

Lemma A.2: If K satisfies conditions A–C then for $\frac{\lambda}{2} < \frac{i}{n} < 1 - \frac{\lambda}{2}$ and $0 < j \leq [\frac{n\lambda}{2}]$

$$\left| \sum_{i=1}^{[\frac{n\lambda}{2}]-j} w_{\lambda,n}(\frac{i}{n},i)(w_{\lambda,n}(\frac{i}{n},i+j) - w_{\lambda,n}(\frac{i}{n},i) \right| = o(\frac{1}{n\lambda}). \qquad (A.2)$$

Also, letting $s_j = 2[\frac{n\lambda}{2}] - j + 1$

$$\left| \sum_{i=s_j-[\frac{n\lambda}{2}]}^{[\frac{n\lambda}{2}]} w_{\lambda,n}(\frac{i}{n},i)(w_{\lambda,n}(\frac{i}{n},s_j - i) - w_{\lambda,n}(\frac{i}{n},i)) \right| = o(\frac{1}{n\lambda}). \qquad (A.3)$$

Lemma A.3: If the correlations satisfy condition E, as $\lambda \rightarrow 0$ and $n\lambda \rightarrow \infty$ then, letting $s_j = 2[\frac{n\lambda}{2}] - j + 1$,

$$\frac{1}{n\lambda} \sum_{j=1}^{[\frac{n\lambda}{2}]} j\, |\rho_n(j) - \rho_n(s_j)| = o(1) \qquad (A.4)$$

Lemma A.4: *If the kernel weights satisfy conditions A.1 – A.3 and the errors satisfy condition D and A.4 then for* $\frac{\lambda}{2} < \frac{i}{n} < 1 - \frac{\lambda}{2}$

$$\left| w'_{\lambda}(\frac{i}{n}, \bullet)\Sigma_n w_{\lambda,n}(\frac{i}{n}, \bullet) - \sigma^2 \frac{W_K}{n\lambda}(1 + 2S_\rho) \right| = o(\frac{1}{n\lambda})$$

where $W_K = \int K^2(x)dx$ *and* S_ρ *is defined by (7).*

proof of A.4: Let R^n_λ *be the* $(2[\frac{n\lambda}{2}] + 1) \times (2[\frac{n\lambda}{2}] + 1)$ *circulant matrix*

$$R^n_\lambda(i,j) = \begin{cases} \sigma^2 \rho(|i-j|) & |i-j| \le [\frac{n\lambda}{2}] \\ \sigma^2 \rho(2[\frac{n\lambda}{2}] + 1 - |i-j|) & |i-j| > [\frac{n\lambda}{2}]. \end{cases}$$

and let $s_j = 2[\frac{n\lambda}{2}] - j + 1$*. Direct computation gives*

$$\begin{aligned} &\left| w'_{\lambda,n}(\frac{i}{n}, \bullet) R^n_\lambda w_{\lambda,n}(\frac{i}{n}, \bullet) - \sigma^2 \frac{W_K}{n\lambda}(1 + 2S_\rho) \right| \\ &\qquad \le \sigma^2 \sum_{j=1}^{[\frac{n\lambda}{2}]} |\rho_n(j)| \sum_{i=1}^{[\frac{n\lambda}{2}]-j} \left| w_{\lambda,n}(\frac{i}{n}, i)\big(w_{\lambda,n}(\frac{i}{n}, i+j) - w_{\lambda,n}(\frac{i}{n}, i)\big) \right| \\ &\qquad\quad + \sigma^2 \sum_{j=1}^{[\frac{n\lambda}{2}]} |\rho_n(j)| \sum_{i=s_j - [\frac{n\lambda}{2}]}^{[\frac{n\lambda}{2}]} \left| w_{\lambda,n}(\frac{i}{n}, i)(w_{\lambda,n}(\frac{i}{n}, s_j - i) - w_{\lambda,n}(\frac{i}{n}, i)) \right| + o(\frac{1}{n\lambda}) \\ &\qquad = o(\frac{1}{n\lambda}) \end{aligned}$$

Let $L = max_x K(x)$*.* R^n_λ *and the* $(2[\frac{n\lambda}{2}]+1) \times (2[\frac{n\lambda}{2}]+1)$ *variance matrix for the design points receiving positive weight are the same on the first* $[\frac{n\lambda}{2}]+1$ *diagonals, and the correlations satisfy condition E, so*

$$\begin{aligned} \left| w'_{\lambda}(\frac{i}{n}, \bullet)\Sigma_n w_{\lambda,n}(\frac{i}{n}, \bullet) - w'_{\lambda,n}(\frac{i}{n}, \bullet) R^n_\lambda w_{\lambda,n}(\frac{i}{n}, \bullet) \right| &\le 2 \left(\frac{L}{n\lambda} \right)^2 \sum_{j=1}^{[\frac{n\lambda}{2}]} j \, |\rho_n(j) - \rho_n(s_j)| \\ &= o(\frac{1}{n\lambda}) \end{aligned}$$

The proof of Theorem 1, follows from equation (10) and lemma A.4.

A.2 Proof of Lemma 2.1

Proof: Let $S_{|\rho|} = \sum_{i=1}^{\infty} |\rho(i)|$. Pick $\varepsilon > 0$. Let $x_\varepsilon = inf\{x > 0, |K(x) - K(0)| > \frac{\varepsilon}{2S_{|\rho|}}$.

Let $L = max_x K(x)$. Pick N_ε so that $n > N_\varepsilon$ implies $\sum_{j=[n\lambda_\varepsilon]+1}^{\infty} |\rho_n(j)| < \frac{\varepsilon}{4L}$. Then, for $n\lambda > 1$,

$$n\lambda|\frac{K(0)}{n\lambda}S_\rho - \sum_{j=1}^{[\frac{n\lambda}{2}]} w_{\lambda,n}(\frac{i}{n}, i+j)\rho(j)| \leq 2L \sum_{j=[n\lambda x_\varepsilon]+1}^{\infty} |\rho_n(j)| + \sum_{j=1}^{[n\lambda x_\varepsilon]} |K(0) - K(\frac{j}{n\lambda})||\rho_n(j)|$$
$$\leq \varepsilon.$$

References

Allen, D.M. (1974) "The relationship between variable selection and data augmentation and a method for prediction," *Technometrics,* **16** 1307-1325.

Altman, N. S. (1989) "Estimation Error Correlation in Nonparametric Regression," Biometrics Unit Memo BU-982-M, Cornell University.

Becker, R.A., Chambers, J.M. (1984) S **An Interactive Environment for Data Analysis and Graphics**, Wadsworth Statistics/Probability Series.

Box, G.E.P. and Jenkins, G.M. (1976) **Time Series Analysis: Forecasting and Control**, San Francisco, CA: Holden-Day.

Breaker, L.C. and Lewis, P.A.W. (1988) "On the Detection of a 40 to 50 Day Oscillation in Sea Surface Temperature along the Central California Coast," Estuarine Coastal and Shelf Science, **26** 395-408.

Breiman, L., Friedman, J.H. (1985) "Estimating Optimal Transformations for Multiple Regression and Correlation," *Journal of the American Statistical Association,* **80** 580-619.

Collomb G. (1985) "Nonparametric Regression: An Up-to-Date Bibliography," *Statistics* **16** 309-324.

Craven, P. and Wahba, G. (1979) "Smoothing Noisy Data with Spline Functions," *Numer. Math.* **31** 377-403.

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*, Ser. B, **47** 1-52.

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

Efron, B. (1986) "How Biased is the Apparent Error Rate of a Prediction Rule?" *Journal of the American Statistical Association* **81** 461-470.

Engle, R. Granger, C.W.J. Rice, J. Weiss, A. (1986) "Semiparametric Estimates of the Relation Between Weather and Electricity Sales," *Journal of the American Statistical Association*

81 310-320.

Friedman, J.H., Stuetzle, W. (1981) "Projection Pursuit Regression," ***Journal of the American Statistical Association*** **76** 817-823.

Gasser, T., Müller, H-G. (1979) "Kernel estimation of regression functions," ***Smoothing Techniques for Curve Estimation*** 23-67 **Lecture Notes in Math.** 757,Berlin: Springer-Verlag.

Geisser, S. (1975) "The predictive sample reuse method with applications," ***Journal of the American Statistical Association*** **70** 320-328.

Greblicki, W., Krzyzak, A., Pawlak, M. (1984) "Distribution-Free Pointwise Consistency of Kernel Regression Estimate," ***Annals of Statistics,*** **12** 1570-1575.

Härdle, W. and Kelly, G. (1987) "Nonparametric Kernel Regression Estimation - Optimal Choice of Bandwidth," ***Statistics***, **18**, 21-35.

Härdle, W. Hall, P., Marron, J.S. (1988) "How Far are Automatically Chosen Regression Smoothing Parameters from their Optimum?" ***Journal of the American Statistical Association*** **83** 86-95.

Härdle,W. and Marron, J.S. (1985) "Asymptotic nonequivalence of some bandwidth selectors in nonparametric regression," ***Biometrika*** **72** 481-484.

Hart, J.D. (1987) "Kernel Regression with Time Series Errors," (in press)

Hart, J.D. and Wehrly, T.E. (1986) "Kernel Regression Estimation Using Repeated Measurements Data," ***Journal of the American Statistical Association*** **81** 1080-1088.

Huber, P. J.(1985) "Projection Pursuit," ***Annals of Statistics,*** **13** 435-475.

Li, K-C. (1985) "Consistency for Cross-Validated Nearest Neighbor Estimates in Nonparametric Regression," ***Annals of Statistics,*** **12** 230-240.

– (1986) "Asymptotic Optimality of C_L and Generalized Cross-Validation in Ridge Regression with Application to Spline Smoothing," ***Annals of Statistics,*** **14** 1101-1112.

Mallows, C.L. (1973) "Some Comments on C_P," ***Technometrics*** **15** 661-675.

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

Parzen, E. (1959) "Statistical Inference on Time Series by Hilbert Space Methods, I," Technical Report No. 23 (NR-C42-993), Department of Statistics, Stanford University .

– (1961) "Regression Analysis of Continuous Parameter Time Series," Fourth Berkeley Symposium.

Priestley, M.B. and Chao, M.T. (1972) "Non-parametric function fitting," *Journal of the Royal Statistical Society*, Ser. B,**34** 385-392.

Rice, J. (1984a) "Bandwidth Choice for Nonparametric Regression," *Annals of Statistics,* **12** 1215-1230.

– (1984b) "Boundary Modification for Kernel Regression," *Communications in Statistics* 13 893-900.

Silverman, B.W. (1984) "Spline Smoothing: The Equivalent Variable Kernel Method," *Annals of Statistics,* **12** 898-916.

– (1985) "Some Aspects of the Spline Smoothing Approach to Non-parametric Regression Curve Fitting," *Journal of the Royal Statistical Society*, Ser. B, **47** 1-52.

Stone, M. (1974) "Cross-validatory choice and assessment of statistical predictions," *Journal of the Royal Statistical Society*, Ser. B, **36** 111-147.

– (1977) "An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion," *Journal of the Royal Statistical Society*, Ser. B, **39** 44-47.

Wahba, G. (1978) "Improper Priors, Spline Smoothing and the Problem of Guarding Against Model Errors in Regression," *Journal of the Royal Statistical Society*, Ser. B,**40** 364-372.

– (1984) "Cross-Validated Spline Methods for the Estimation of Multivariate Functions from Data on Functionals," *Proceedings 50th Anniversary Conference Iowa State Statistical Laboratory*, eds. H.A. David and H.T. David, The Iowa State University Press.

Watson, G.S. (1964) "Smooth Regression Analysis," *Sankhyā, Series A* **26** 359-372.

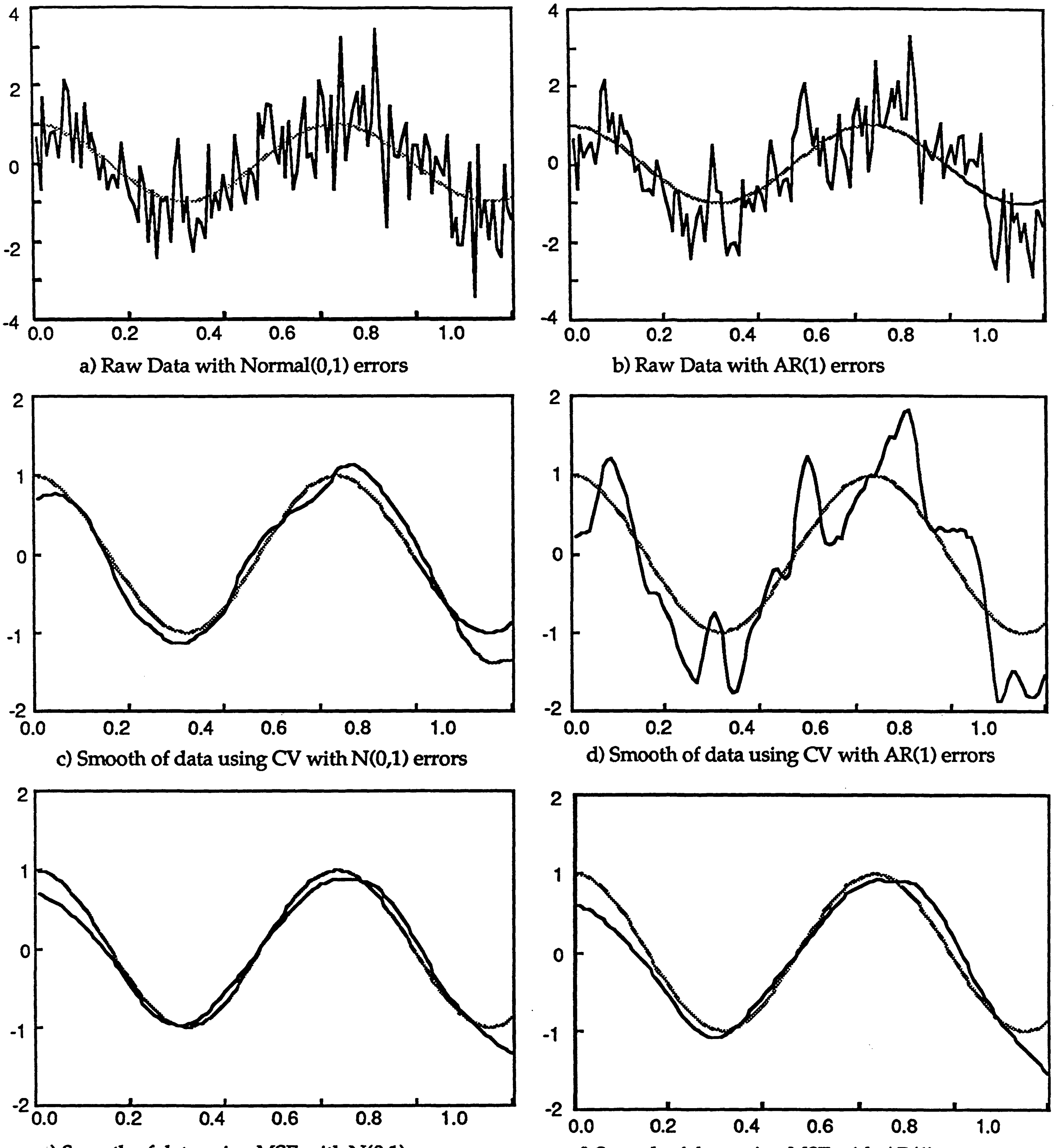
Wendelberger, J. G. (1987) "Multiple Minima of the Generalized Cross Validation Function:

Paint Attribute Data", *Technical Report, Mathematics Department General Motors Research Laboratories.*

Williams, P.T., Wood, P.D., Vranizan, K.M., Albers, J.J., Garay, S.C., Taylor, C.B. (1985) "Coffee Intake and Elevated Cholesterol and Apolipoprotein B Levels in Men" *Journal of the American Medical Association,* 253 1407-1411.

	$\sigma^2 = 1.0$						$\sigma^2 = .01$					
	Uniform Kernel			Spline Kernel			Uniform Kernel			Spline Kernel		
				$\rho(1) = .9$								
min ASE	21.0	35.3	57.3	21.2	35.1	58.3	38.4	63.8	86.7	34.6	59.2	78.9
min GCV	1.60	2.24	3.72	1.60	2.26	3.71	1.11	1.23	1.46	1.22	1.37	1.75
min $GCV_{\rho^{-1}}$	1.09	1.39	2.01	1.06	1.21	1.92	1.01	1.09	1.24	1.09	1.23	1.65
min GCV_{ρ}	1.05	1.17	1.48	1.01	1.07	1.50	1.00	1.04	1.08	1.00	1.04	1.14
min $GCV_{\tilde{\rho}^{-1}}$	1.25	1.63	2.37	1.27	1.55	2.78	1.01	1.09	1.24	1.09	1.23	1.65
min $GCV_{\tilde{\rho}}$	1.06	1.22	1.63	1.26	1.70	2.16	1.02	1.09	1.24	1.22	1.48	1.73
				$\rho(1) = .3$								
min ASE	4.98	7.18	10.7	3.94	5.59	9.62	15.9	19.0	25.4	10.8	12.4	16.8
min GCV	2.31	3.91	6.49	2.73	4.79	7.65	1.83	2.19	2.69	1.60	2.27	2.90
min $GCV_{\rho^{-1}}$	1.04	1.13	1.71	1.00	1.09	1.49	1.00	1.07	1.31	1.00	1.08	1.29
min GCV_{ρ}	1.00	1.14	1.56	1.02	1.18	1.54	1.00	1.04	1.24	1.05	1.26	1.59
min $GCV_{\tilde{\rho}^{-1}}$	1.05	1.32	1.84	1.01	1.17	1.93	1.00	1.07	1.31	1.00	1.08	1.29
min $GCV_{\tilde{\rho}}$	1.03	1.13	1.52	1.03	1.20	1.91	1.64	1.80	2.05	4.65	5.51	6.76
				$\rho(1) = 0$								
min ASE	3.46	5.48	7.45	2.74	4.45	6.36	11.1	13.5	18.1	6.76	9.71	12.1
min GCV	1.00	1.11	1.46	1.02	1.09	1.26	1.16	1.24	1.47	1.00	1.11	1.32
min $GCV_{\rho^{-1}}$	1.00	1.08	1.39	1.01	1.05	1.18	1.00	1.02	1.24	1.00	1.04	1.23
min GCV_{ρ}	1.00	1.16	1.46	1.02	1.09	1.26	1.00	1.08	1.27	1.00	1.11	1.32
min $GCV_{\tilde{\rho}^{-1}}$	1.00	1.09	1.39	1.00	1.08	1.37	1.00	1.02	1.24	1.00	1.04	1.23
min $GCV_{\tilde{\rho}}$	1.00	1.09	1.44	1.07	1.18	1.71	2.07	2.41	2.69	5.00	6.47	8.53
				$\rho(1) = -.3$								
min ASE	2.51	3.16	4.25	1.85	2.41	3.26	7.27	8.92	11.0	4.21	5.68	6.87
min GCV	1.04	1.20	1.46	1.01	1.09	1.36	1.34	1.47	1.66	1.00	1.00	1.08
min $GCV_{\rho^{-1}}$	1.04	1.16	1.39	1.00	1.07	1.20	1.00	1.00	1.13	1.00	1.00	1.05
min GCV_{ρ}	1.06	1.23	1.74	1.01	1.08	1.21	1.00	1.02	1.29	1.00	1.11	1.24
min $GCV_{\tilde{\rho}^{-1}}$	1.02	1.10	1.35	1.00	1.09	1.20	1.00	1.00	1.13	1.00	1.00	1.05
min $GCV_{\tilde{\rho}}$	1.00	1.16	1.36	1.01	1.13	1.47	2.90	3.34	3.89	6.34	7.56	9.68
				$\rho(1) = -.9$								
min ASE	0.59	0.85	1.00	0.26	0.31	0.48	2.67	3.34	4.06	0.73	0.93	1.26
min GCV	1.00	1.45	2.59	1.66	2.61	3.21	3.17	3.64	4.08	1.61	2.33	4.21
min $GCV_{\rho^{-1}}$	1.00	1.15	1.49	1.00	1.02	1.26	1.00	1.40	1.50	1.00	1.04	1.18
min GCV_{ρ}	1.29	2.52	5.96	1.02	1.41	4.73	1.33	1.48	1.66	1.09	1.19	1.69
min $GCV_{\tilde{\rho}^{-1}}$	1.00	1.12	1.25	1.00	1.06	1.25	1.00	1.40	1.50	1.00	1.04	1.18
min $GCV_{\tilde{\rho}}$	1.03	1.37	1.62	1.00	1.11	2.00	2.28	2.98	4.16	12.7	15.1	20.0

Table 2:*Effects of using the direct and indirect methods of correcting the bandwidth selection criteria for correlation. Both the true and estimated correlation functions are used. The data are 128 points from* $y_i = cos(3.15\pi x) + \epsilon_i$ *where the design points are equally spaced on* $[0, 1]$ *and the errors are AR(1). Results for the uniform and spline kernels are shown. The entries are the medians over 50 realizations of the process with the quartiles flanking in smaller print. "Min ASE" is the median ASE at the bandwidth minimizing the ASE. It is displayed in natural units, multiplied by 100, for* $\sigma^2 = 1.0$, *or by 10000, for* $\sigma^2 = .01$. *The other entries are median ASE at the minimum value of the bandwidth selection criterion, expressed as a multiple of the actual minimum. "GCV" is the uncorrected GCV criterion, "*$GCV_{\rho^{-1}}$*" is the GCV using the indirect correction with the true value of the correlation function, and "*GCV_{ρ}*" is GCV using the direct correction with the true value of the correlation function. "*$GCV_{\tilde{\rho}^{-1}}$*" and "min* $GCV_{\tilde{\rho}}$*" are the corrected GCV criteria when the correlations have been estimated from residuals from a moderate bandwidth smooth.*



a) Raw Data with Normal(0,1) errors

b) Raw Data with AR(1) errors

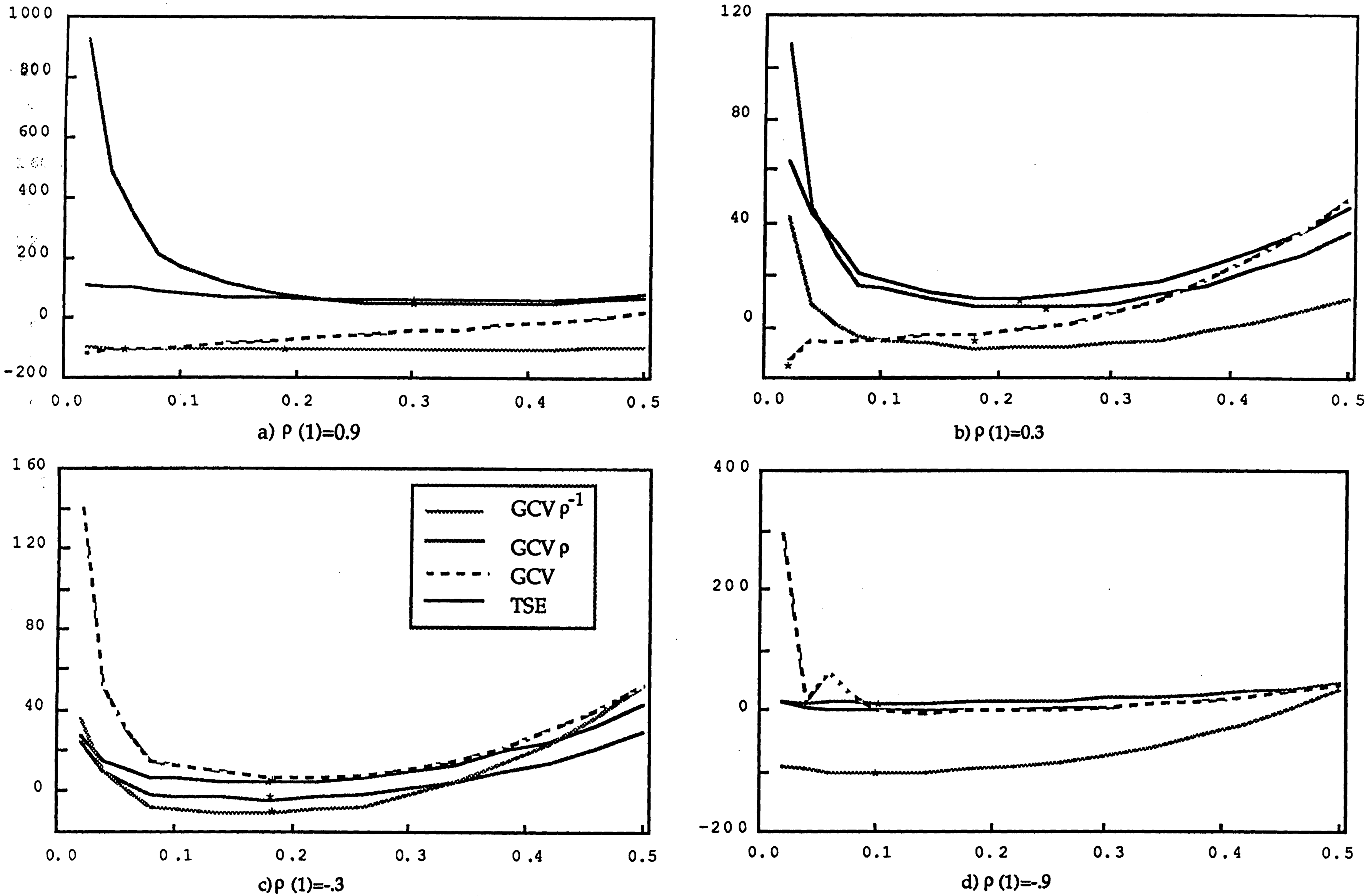
c) Smooth of data using CV with N(0,1) errors

d) Smooth of data using CV with AR(1) errors

e) Smooth of data using MSE with N(0,1) errors

f) Smooth of data using MSE with AR(1) errors

Figure 1: Raw and smoothed data for the mean function E(y)=cos(3.15πx). The error variance is 1.0. The bandwidth for the smooth was selected using ordinary cross-validation (c and d) or minimum MSE (e and f).



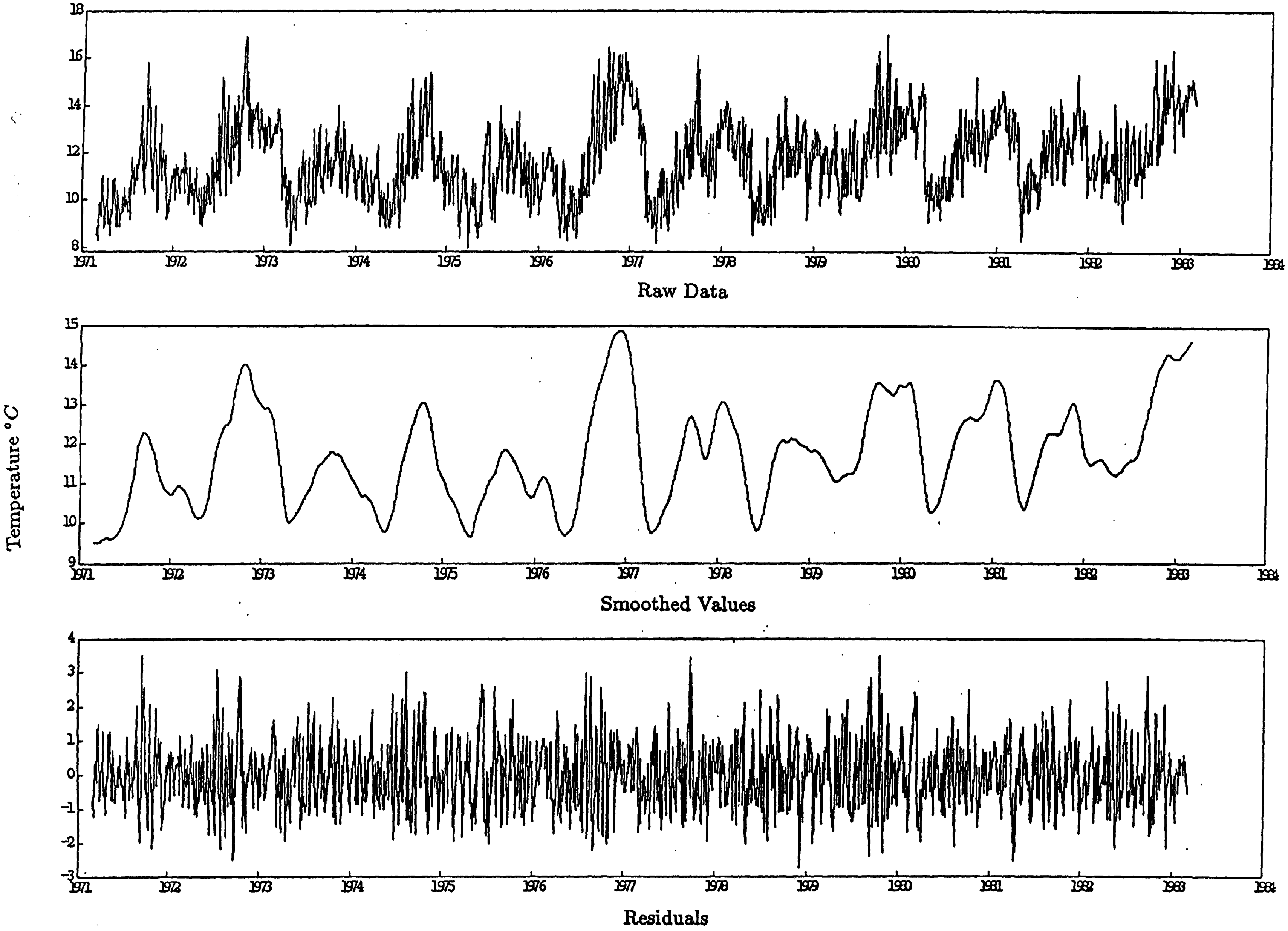
a) ρ (1)=0.9

b) ρ (1)=0.3

c) ρ (1)=-.3

d) ρ (1)=-.9

Figure 2: TSE, uncorrected and corrected GCV as a function of bandwidth for the uniform kernel with AR(1) errors.
$f(x)=\cos(3.15\pi x)$ σ^2 =1.0 n=128 various values of $\rho(1)$



Raw Data

Smoothed Values

Residuals

Figure 3: Sea Surface Temperatures at Granite Canyon, California.