Statistical Science
2001, Vol. 16, No. 2, 134–153

Nonparametric Regression with Correlated Errors

Jean Opsomer, Yuedong Wang and Yuhong Yang

Abstract. Nonparametric regression techniques are often sensitive to the presence of correlation in the errors. The practical consequences of this sensitivity are explained, including the breakdown of several popular data-driven smoothing parameter selection methods. We review the existing literature in kernel regression, smoothing splines and wavelet regression under correlation, both for short-range and long-range dependence. Extensions to random design, higher dimensional models and adaptive estimation are discussed.

Key words and phrases: Kernel regression, splines, wavelet regression, adaptive estimation, smoothing parameter selection.

1. INTRODUCTION

Nonparametric regression is a rapidly growing and exciting branch of statistics, both because of recent theoretical developments and more widespread use of fast and inexpensive computers. In nonparametric regression problems, the researcher is most often interested in estimating the mean function $\mathrm{E}(Y|\mathbf{X}) = f(\mathbf{X})$ from a set of observations $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$, where the $\mathbf{X}_i$ can be either univariate or multivariate. Many competing methods are currently available, including kernel-based methods, regression splines, smoothing splines and wavelet and Fourier series expansions. The bulk of the literature in these areas has focused on the case in which an unknown mean function is "masked" by a certain amount of white noise, and the goal of the regression is to "remove" the white noise and uncover the function. More recently, a number of authors have begun to look at the situation where the noise is no longer white and instead contains a certain amount of "structure" in the form of correlation. The focus of this article is to look at the problem of estimating the mean function $f(\cdot)$ in the presence of correlation, not that of estimating the correlation function itself. In this context, our goals are (1) to explain some of the difficulties associated with the presence of correlation in nonparametric regression, (2) to provide an overview of the nonparameteric regression literature that deals with the correlated errors case and (3) to discuss some new developments in this area. Much of the literature in nonparametric regression relies on asymptotic arguments to clarify the probabilistic behavior of the proposed methods. The same approach will be used here, but we attempt to provide intuition into the results as well.

In this article, we will be looking at the following statistical model:

$$Y_i = f(\mathbf{X}_i) + \varepsilon_i, \tag{1}$$

where $f(\cdot)$ is an unknown, smooth function, and the error vector $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)$ has variance–covariance matrix $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{C}$.

Researchers in different areas of nonparametric regression have addressed different versions of this problem. For instance, the $\mathbf{X}_i$ are either assumed to be random, or fixed within some domain (to avoid confusion, we will write lower case $\mathbf{x}_i$ when the covariates are fixed, and upper case $\mathbf{X}_i$ when these are random variables). The specification of the correlation can also vary significantly: the correlation matrix $\mathbf{C}$ is considered completely known in some of these areas, known up to a finite number of parameters, or only assumed to be stationary but otherwise left completely unspecified in other areas. Another issue concerns whether the errors are assumed to be *short-range dependent*, where the correlation decreases rapidly as the distance between two observations increases, or *long-range dependent* (short-range/long-range dependency will

Jean Opsomer is Associate Professor, Department of Statistics, Iowa State University, Snedecor Hall, Ames, Iowa 50011 (*e-mail: jopsomer@iastate.edu*). *Yuedong Wang is Associate Professor, University of California, Santa Barbara and Yuhong Yang is Associate Professor, Iowa State University.*

be defined more exactly in Section 3.1). When discussing the various methods proposed in the smoothing literature, we will point out the major differences in assumptions between these areas.

Section 2 explains the practical difficulties associated with estimating $f(\cdot)$ under model (1). In Section 3, we review the existing literature on this topic in several areas of nonparametric regression. Section 4 describes some extensions of existing results as well as new developments. This last section is more technical than the previous ones, and nonspecialists might want to skip it on first reading.

2. PROBLEMS WITH CORRELATION

A number of problems, some quite fundamental, occur when nonparametric regression is attempted in the presence of correlated errors. Indeed, in the most general setting where no parametric shape is assumed for the mean nor the correlation function, the model is essentially unidentifiable, so that it is theoretically impossible to estimate either function separately. In most practical applications, however, the researcher has some idea what represents a reasonable type of fit to the data at hand, and he will use that expectation to decide what is an "acceptable" or "unacceptable" function estimate.

For all nonparametric regression techniques, the shape and smoothness of the estimated function depends to a large extent on the specific value chosen for a "smoothing parameter," defined differently for each technique. In order to avoid having to select a value for the smoothing parameter by trial and error, several types of data-driven selection methods have been developed to assist researchers in this task. However, the presence of correlation between the errors, if ignored, causes the commonly used automatic tuning parameter selection methods, such as cross-validation or plug-in, to break down.

This breakdown of automated methods, as well as a possible solution to it, are illustrated by a simple simulated example in Figure 1. For 200 equally spaced observations and a low-order polynomial mean function $[f(x) = 1 - 6x + 36x^2 - 53x^3 + 22x^5]$, four progressively more correlated sets of errors were generated from the same vector of independent noise, and added to the mean function. The errors are normally distributed with variance $\sigma^2 = 0.5$ and correlation following an AR(1) process (autocorrelation of order 1), $\text{corr}(\varepsilon_i, \varepsilon_j) = \exp(-\alpha|x_i - x_j|)$. Figure 1 shows four local linear regression fits for these datasets. For each dataset, two bandwidth selection methods were used: cross-validation (CV) and a correlation-corrected method called (CDPI), further discussed in Section 4.1. Table 1 summarizes the bandwidths selected for the four datasets under both methods.

TABLE 1
*Summary of bandwidth selection for simulated data in Figure 1**

Correlation level	Autocorrelation	CV	CDPI
Independent	0	0.15	0.13
$\alpha = 400$	0.14	0.10	0.12
$\alpha = 200$	0.37	0.02	0.12
$\alpha = 100$	0.61	0.01	0.11

*Autocorrelation refers to the correlation between adjacent observations.

Table 1 and Figure 1 clearly show that as the correlation increases, the bandwidth selected by cross-validation becomes smaller and smaller, and the fits become progessively more undersmoothed. The bandwidths selected by CDPI, a method that accounts for the presence of correlation, are much more stable and result in virtually the same fit for all four cases.

This type of undersmoothing behavior in the presence of correlated errors has been observed with most commonly used automated bandwidth selection methods. At its most conceptual level, it is caused by the fact that the bandwidth selection method "perceives" all the structure in the data to be due to the mean function, and attempts to incorporate that information into its estimate of the trend. When the data are uncorrelated, this "perception" is valid, but it breaks down in the presence of correlation. Unlike in this simulated example, in practice it is very often not known what portions of the behavior of the observations *should* be attributed to signal or to noise for a given dataset. The choice of bandwith selection approach should therefore be dictated by an understanding of the nature of the data.

The previous example showed that correlation can cause the data-driven bandwidth selection methods to break down. Selecting the bandwidth "visually" or by trial and error can also be misleading, however. Indeed, even if data are independent, a wrong choice of the smoothing parameter can induce spurious serial correlation in the residuals. Conversely, a wrong choice of the smoothing parameter can lead to an estimated correlation that does not reflect the true correlation in the random error. Two simple simulations using smoothing splines illustrate these facts (see Figure 2). In the first simulation, 100 observations are generated from the model $Y_i = \sin(2\pi i/100) + \varepsilon_i$, $i = 1, \ldots, 100$, where ε_i's are independent and identically distributed normal random variables with mean zero and standard

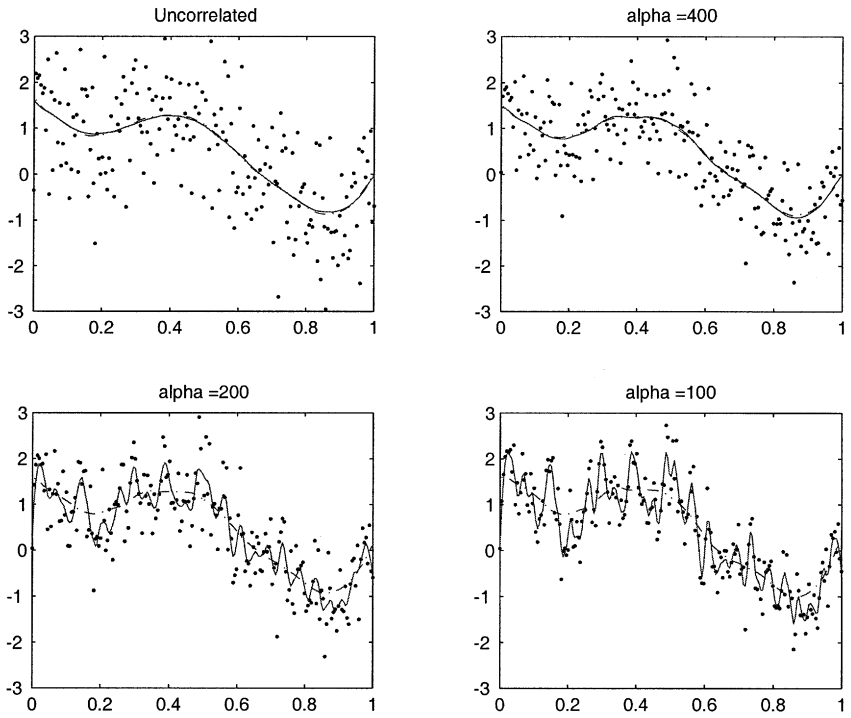

FIG. 1. *Simulated data with four levels of* AR(1) *correlation, fitted with local linear regression;* (−) *represents fit obtained with bandwidth selected by cross-validation,* (– ·) *fit obtained with bandwidth selected by CPDI.*

deviation 1. The S-Plus function `smooth.spline` is used to fit the data with smoothing parameter set at 0.01 [Figure 2(a)]. In the second simulation, 100 observations are generated according to model (1) with mean zero and errors following a first-order autoregressive process [AR(1)] with autocorrelation 0.5 and standard deviation 1. The S-Plus function `smooth.spline` is again used to fit the data with smoothing parameter selected by the *generalized cross-validation* (GCV) method [Figure 2(c)]. The estimated autocorrelation function (ACF) for the first plot looks autoregressive [Figure 2(b)], while that for the second plot appears independent [Figure 2(d)]. In both cases, this conclusion about the error structure is erroneous and the mean function is incorrectly estimated.

As mentioned above, the researcher will often have some idea on what represents an appropriate cut-off between the short-term behavior induced by the correlation and the long-term behavior of primary interest. Establishing that cut-off for a specific dataset can be done by trying out different values for the smoothing parameter and picking the one that results in an appropriate fit. In fact, the presence of correlation in the residuals can sometimes be used to assist in the visual selection of an appropriate bandwidth, when the data are independent. Consider the examples from Figure 2 again. In 2(a) and 2(b), the positive serial correlation was induced by the oversmoothed fit. A smaller bandwidth will result in a better fit to the data, and remove the correlation in the residuals. Figure 3(a) and 3(b) show the fit produced by the GCV-selected bandwidth and the corresponding autocorrelation function of the residuals. Because the data are truly independent here, the GCV smoothing parameter selection method works correctly.

In 2(c) and 2(d), this situation is reversed. If the "pattern" found by GCV in Figure 2(c) is not an acceptable fit (depending on the application), a larger smoothing parameter value has to be set by hand, or by using one of the smoothing parameter selection methods that account for correlation, as will be presented below. Figure 3(c) and 3(d) display the fit and the autocorrelation function for the smoothing parameter value selected by the *extended GML* method (see Section 4.2) with an assumed AR(1) error process. The residuals from this new fit are now correlated. If the researcher prefers this new fit, he should assume that the model errors were also correlated. It should be noted that in such situations, standard residual diagnostic tests for nonparametric goodness-of-fit, such as those in Hart (1997), will fail. If the error structure is correctly modeled, statistical inference is still possible, however. For instance, Figure 3(c) shows a 95% Bayesian confidence interval [see Wang (1998b)], indicating that the slight upward trend is most

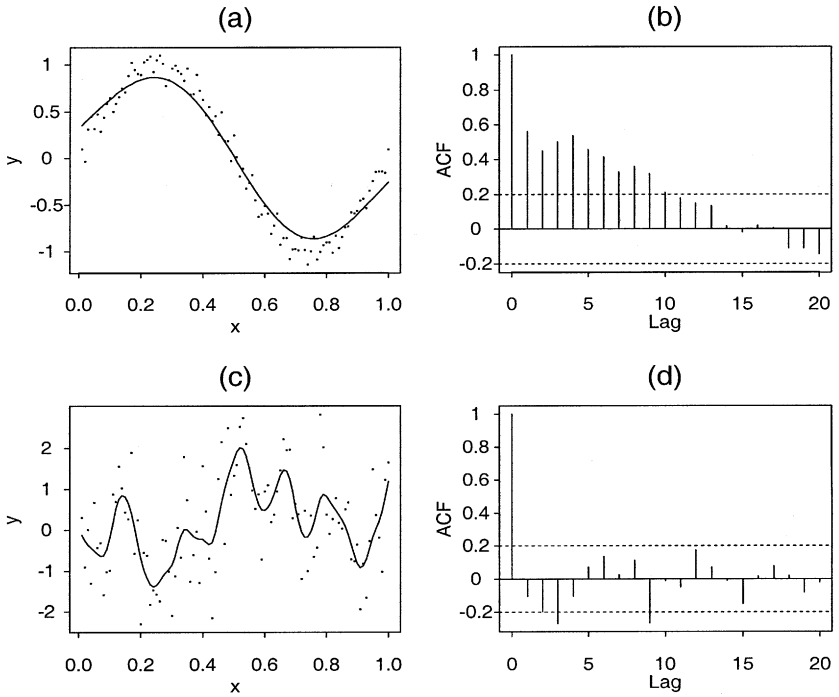

FIG. 2. *Simulation* 1 (*independent data*): (a) *spline fit using smoothing parameter value of* 0.01; (b) *autocorrelation function of residuals. Simulation* 2 (*autoregressive*): (c) *spline fit using GCV smoothing parameter selection*; (d) *autocorrelation function of residuals.*

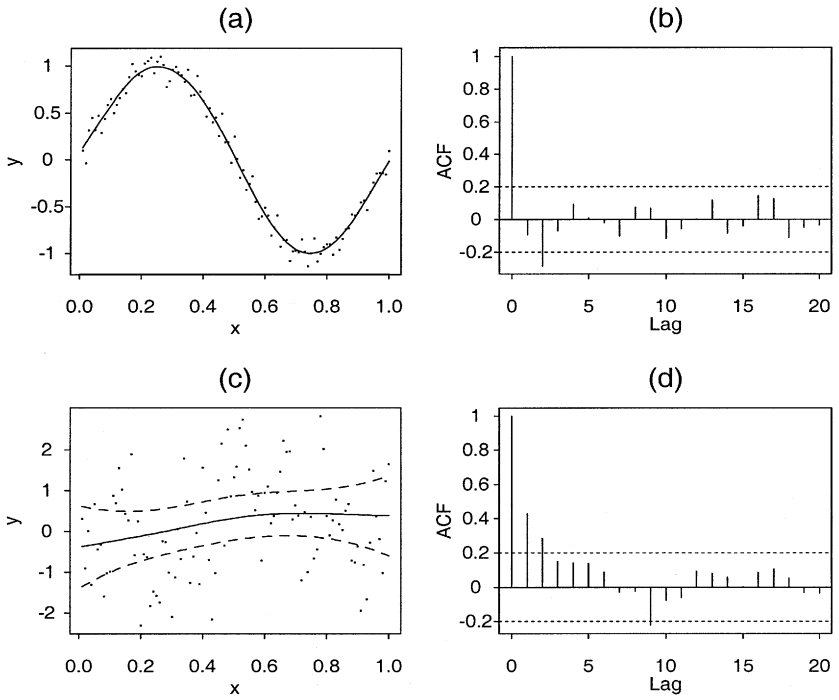

FIG. 3. *Simulation* 1 *again* (*independent data*): (a) *Spline fit* (*using GVC*); (b) *Autocorrelation function of residuals. Simulation* 2 (*autoregressive*): (c) *Spline fit* (*and* 95% *Bayesian confidence interval, using extended GML*); (d) *Autocorrelation function of residuals.*

likely not significant, the correct conclusion in this case.

3. RESULTS TO DATE

3.1 Kernel-based Methods

In this section, we consider kernel regression estimation of the mean function for data assumed to follow model (1), with

$$\begin{aligned} &\mathrm{E}(\varepsilon_i) = 0, \qquad \mathrm{Var}(\varepsilon_i) = \sigma^2, \\ &\mathrm{Corr}(\varepsilon_i, \varepsilon_j) = \rho_n(\mathbf{X}_i - \mathbf{X}_j) \end{aligned} \tag{2}$$

and σ^2 unknown and $\rho_n(\cdot)$ an unknown, stationary correlation function. The dependence of ρ on n is indicated by the subscript, because consistency properties of the estimators will depend on the behavior of the correlation function as n increases. In this section, we only consider univariate fixed $\mathbf{x}_i = x_i$ in a bounded interval $[a, b]$, and for simplicity, we let $[a, b] = [0, 1]$. The researchers who have studied the properties of kernel-based estimators of the function $f(\cdot)$ have focused on the *time series* case, in which the design points are fixed and equally spaced $x_i \equiv i/n$, so that

$$Y_i = f\left(\frac{i}{n}\right) + \varepsilon_i,$$

$$\mathrm{Corr}(\varepsilon_i, \varepsilon_j) = \rho_n\left(\frac{i}{n} - \frac{j}{n}\right).$$

We consider the simplest situation, in which the correlation function is taken to be $\rho_n(t/n) = \rho(t)$ for all n, but otherwise $\rho(\cdot)$ is left unspecified. Note that this implies that the correlation between two fixed locations decreases as $n \to \infty$, because a fixed value for $t = |i - j|$ corresponds to a decreasing distance between observations.

We also assume that the errors are short-range dependent. The error process is said to be short-range dependent if for some $c > 0$ and $\gamma > 1$, the spectral density $H(\omega) = \frac{\sigma^2}{2\pi}\sum_{k=-\infty}^{\infty}\rho(k)e^{-i\omega}$ of the errors satisfies

$$H(\omega) \sim c\omega^{-(1-\gamma)} \quad \text{as } \omega \to 0$$

(see, e.g., Cox, 1984). In that case, $\rho(j)$ is of order $|j|^{-\gamma}$ (see, e.g., Adenstedt, 1974). When the correlation decreases at order $|j|^{-\gamma}$ for some dependency index $0 < \gamma \leq 1$, the errors are said to have a long-range dependence. Long-range dependence substantially increases the difficulty in estimating the mean function and will be discussed separately in Section 3.3.

The function $f(\cdot)$ can be fitted by kernel regression or local polynomial regression. Following the literature in this area, we discuss estimation by kernel regression, and for simplicity, we consider the Priestley–Chao kernel estimator (Priestley and Chao, 1972). The estimator of $f(\cdot)$ at a point $x \in [0, 1]$ is defined as

$$\hat{f}(x) = \mathbf{s}_{x;h}^T\mathbf{Y} = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)Y_i$$

for some kernel function K and bandwidth h, where ^{T} denotes transpose, $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ and

$$\mathbf{s}_{x;h} = \frac{1}{nh}\left(K\left(\frac{x_1 - x}{h}\right), \ldots, K\left(\frac{x_n - x}{h}\right)\right)^T.$$

The mean squared error of $\hat{f}(x)$ is

$$\begin{aligned} &\mathrm{MSE}(\hat{f}(x); h) \\ &\quad = \mathrm{E}\left(\mathbf{s}_{x;h}^T\mathbf{Y} - f(x)\right)^2 \\ &\quad = \left(\mathbf{s}_{x;h}^T\mathrm{E}(\mathbf{Y}) - f(x)\right)^2 + \sigma^2\mathbf{s}_{x;h}^T\mathbf{C}\mathbf{s}_{x;h}, \end{aligned} \tag{3}$$

where $\mathbf{C}$ is the unknown correlation matrix of $\mathbf{Y}$.

Before we can study the asymptotic behavior of $\hat{f}(x)$, a number of assumptions on the statistical model and the components of the estimator are needed.

(AS.I) The kernel K is compactly supported, bounded and continuous. We assume that $\int K(u)\,du = 1$, $\int uK(u)\,du = 0$ and $\int u^2K(u)\,du \neq 0$.

(AS.II) The 2nd derivative function, $f''(\cdot)$, of $f(\cdot)$ is bounded and continuous.

(AS.III) As $n \to \infty$, $h \to 0$ and $nh \to \infty$.

The first term of the MSE in (3) represents the squared bias of $\hat{f}(x)$, and does *not* depend on the dependency structure of the errors. Under the assumptions (AS.I)–(AS.III), the bias can be asymptotically approximated by

$$\mathbf{s}_{x;h}^T\mathrm{E}(\mathbf{Y}) - f(x) = h^2\frac{\mu_2(K)}{2}f''(x) + o(h^2),$$

with $\mu_r(G) = \int u^rG(u)\,du$ for any function $G(\cdot)$. The effect of the correlation structure on the variance part of the MSE is potentially severe, however. If

(R.I) $\lim_{n\to\infty}\sum_{k=1}^{n}|\rho(k)| < \infty$, so that $R = \sum_{k=1}^{\infty}\rho(k)$ exists;

(R.II) $\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n}k|\rho(k)| = 0$,

the variance component of the MSE can be approximated asymptotically by

$$\sigma^2\mathbf{s}_{x;h}^T\mathbf{C}\mathbf{s}_{x;h} = \frac{1}{nh}\mu_0(K^2)\sigma^2(1 + 2R) + o\left(\frac{1}{nh}\right),$$

(Altman, 1990). The assumptions (R.I) and (R.II), common in time series analysis, ensure that observations sufficiently far apart are essentially uncorrelated. When the observations are uncorrelated,

$R = 0$ so that this result reduces to the one usually reported for kernel regression with independent errors. Note also that $\frac{1}{2\pi}\sigma^2(1+2R) = H(0)$, the spectral density at $\omega = 0$. This fact will be useful in developing bandwidth selection methods that incorporate the effect of the correlation (see Section 4.1).

Let AMSE denote the asymptotic approximation to the MSE in (3),

$$\text{AMSE}(\hat{f}(x);h) = h^4\frac{\mu_2(K)^2}{4}f''(x)^2 + \frac{1}{nh}\mu_0(K^2)\sigma^2(1+2R), \tag{4}$$

which is minimized for

$$h_{\text{opt}} = \left(\frac{\mu_0(K^2)\sigma^2(1+2R)}{n\,\mu_2(K)^2 f''(x)^2}\right)^{1/5},$$

the asymptotically optimal bandwidth for estimating $f(x)$. The effect of the correlation sum R on this optimal bandwidth is easily seen. Note first that the optimal *rate* for the bandwidth, $h_{\text{opt}} \propto n^{-1/5}$, is the same as that for the independent errors case. The exact value of h_{opt} depends of R, however. If $R > 0$, implying that the error correlation is positive, then the variance of $\hat{f}(x)$ will be larger than in the corresponding uncorrelated case. The AMSE is therefore minimized by a value for the bandwidth h that is *larger* than in the uncorrelated case. Conversely, if $R < 0$, the AMSE-optimal bandwidth is smaller than in the uncorrelated case, but in practice, positive correlation is much more often encountered.

Positive correlation has the additional perverse effect of making automated bandwidth selection methods pick *smaller* bandwidths, as illustrated in Figure 1. This behavior is explained in Altman (1990) and Hart (1991) for cross-validation and briefly reviewed here. As a global measure of goodness-of-fit for $\hat{f}(\cdot)$, we consider the mean average squared error (MASE),

$$\text{MASE}(h) = \frac{1}{n}\sum_{i=1}^{n}\text{E}\left(\hat{f}\left(\frac{i}{n}\right) - f\left(\frac{i}{n}\right)\right)^2, \tag{5}$$

which is equal to (3) averaged over the observed locations. Let $\hat{f}_{(-i)}$ denote the kernel regression estimate computed on the dataset with the ith observation removed. The cross-validation criterion for choosing the bandwidth is

$$\text{CV}(h) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}_{(-i)}\left(\frac{i}{n}\right) - Y_i\right)^2, \tag{6}$$

so that

$$\text{E}(\text{CV}(h)) \approx \text{MASE}(h) + \sigma^2 - \frac{2}{n}\sum_{i=1}^{n}\text{Cov}\left(\hat{f}_{(-i)}\left(\frac{i}{n}\right), \varepsilon_i\right).$$

The latter covariance term is in addition to the correlation effect already included in the MASE. It can be shown that asymptotically, $\text{Cov}(\hat{f}_{(i)}(\frac{i}{n}), \varepsilon_i) \approx T_c/nh$ for some constant T_c. For severe correlation, this additional term dominates CV(h) and leads to a fit that nearly interpolates the data, as shown in Figure 1. This results holds not only for cross-validation but also for related measures of fit such as generalized cross-validation (GCV) and Mallows' criterion (Chiu, 1989).

One approach to solve this problem is to model the correlation parametrically, and several such methods were proposed independently by Chiu (1989), Altman (1990) and Hart (1991). While their specific implementations varied, each author chose to estimate the correlation function parametrically and to use this estimate to adjust the bandwidth selection criterion. The estimation of the correlation function is of course complicated by the fact that the errors in (1) are unobserved. Chiu (1989) attempts to bypass that problem by estimating the correlation function in the frequency domain while down-weighting the low frequency periodogram components. Hart (1991) attempts to remove most of the trend by differencing, and then also estimates the correlation function in the frequency domain. In contrast to the previous authors, Altman (1990) proposes performing an initial "pilot" regression to estimate the mean function and calculate residuals, and then fits a low-order autoregressive process to these residuals. Hart (1994) describes a further refinement to this parametrically modelled correlation approach. He introduces *time series cross-validation* as a new goodness-of-fit criterion that can be *jointly* minimized over the set of parameters for the correlation function (a pth-order autoregressive process in this case) and the bandwidth parameter. These approaches appear to work well in practice. Even when the parametric part of the model is misspecified, they provide a significant improvement over the fits computed under the assumption of independent errors, as the simulation experiments in Altman (1990) and Hart (1991) show.

However, when performing nonparametric regression, it is sometimes desirable to completely avoid such parametric assumptions. Several methods have been proposed that pursue completely nonparametric approaches. Chu and Marron (1991) propose two new cross-validation based criteria that estimate the MASE-optimal bandwidth without specifying the correlation function. In *modified cross-validation* (MCV), the kernel regression values $\hat{f}_{(-i)}$ in (6) are computed by leaving out the $2l+1$ observations $i-l, i-l+1, \ldots, i+l-1, i+l$

surrounding the ith observation. Because the correlation is assumed to be short-range, proper choice of l will greatly decrease the effect of the terms $\mathrm{Cov}(\hat{f}_{(-i)}(\frac{i}{n}), \varepsilon_i)$ in the CV criterion. In *partitioned cross-validation* (PCV), the observations are partitioned into g subgroups by taking every gth observations. Within each subgroup, the observations are further apart and hence are assumed less correlated. Cross-validation is performed for each subgroup, and the bandwidth estimate for all the observations is a simple function of the average of the subgroup optimal bandwidths. The drawback of both MCV and PCV is that the values of l and g need to be selected with some care.

Herrmann, Gasser and Kneip (1992) also propose a fully nonparametric method for estimating the MASE-optimal bandwidth, but replace the CV-based criteria by a *plug-in* approach. This type of bandwidth selection has been shown to have a number of theoretical and practical advantages over CV (Härdle, Hall and Marron, 1988, 1992). Plug-in bandwidth selection is performed by estimating the unknown quantities in the AMSE (4), replacing them by estimators (hence the name "plug-in") and minimizing the resulting estimated AMSE with respect to the bandwidth h. The estimation of the bias component $B(x;h)^2$ is completely analogous to that in the uncorrelated case. The variance component $\sigma(1+2R)$ is estimated by a summation over second-order differences of lagged residuals.

More recently, Hall, Lahiri and Polzehl (1995) extended the results of Chu and Marron (1991) in a number of useful directions. Their theoretical results apply to kernel regression as well as local linear regression. They also explicitly consider the long-range dependence case, where assumptions (R.I) and (R.II) are no longer required. They discuss bandwidth selection through MCV and compare it with a bootstrap-based approach which estimates the MASE in (5) directly through resampling of "blocks" of residuals from a pilot smooth. As was the case for Chu and Marron (1991), both approaches are fully nonparametric but require the choice of other tuning parameters.

In Section 4.1, we will introduce a new type of fully nonparametric plug-in method that is applicable to both one- and two-dimensional covariates $\mathbf{X}_i$ following either a fixed or a random design.

3.2 Polynomial Splines

In this section, we consider model (1) with fixed design points x_i in $\chi = [0,1]$, and as usually done in the smoothing spline literature, we assume that f is a function with certain smoothness properties. More precisely, assume f belongs to the Sobolev space

$$W_2^m = \Big\{ f\colon f^{(v)} \text{ absolutely continuous}, \quad v = 0, \ldots, m-1, \int_0^1 (f^{(m)}(x))^2\,dx < \infty \Big\}. \tag{7}$$

For a given variance–covariance matrix $\mathbf{C}$, the smoothing spline estimate $\hat{f}$ is the minimizer of the following penalized weighted least-square objective function

$$\min_{f \in W_2^m} \Big\{ \frac{1}{n}(\mathbf{Y}-\mathbf{f})^T \mathbf{C}^{-1}(\mathbf{Y}-\mathbf{f}) + \lambda \int_0^1 (f^{(m)}(x))^2\,dx \Big\}, \tag{8}$$

where $\mathbf{f} = (f(x_1), \ldots, f(x_n))^T$, and λ is the smoothing parameter controlling the trade-off between the goodness-of-fit measured by weighted least-squares and the roughness of the estimate measured by $\int_0^1 (f^{(m)}(x))^2\,dx$.

Unlike in the previous section, smoothing spline research to date always assumes that the $\mathbf{C}$ is parametrically specified. The sample size n is kept fixed when studying the statistical properties of the smoothing splines estimator $\hat{f}$ under correlated errors, so that only finite-sample properties have been studied. Hence, the effect of short-range or long-range dependence of the errors on the asymptotic behavior of the estimator has not been explicitly considered and remains an open research question. We discuss the main finite-sample properties of $\hat{f}$ under correlation in this section and in Section 4.2.

Let

$$\phi_\nu(x) = x^{\nu-1}/(\nu-1)!, \qquad \nu = 1, \ldots, m,$$

$$R(x,z) = \int_0^1 (x-u)_+^{m-1}(z-u)_+^{m-1}\,du/((m-1)!)^2,$$

where $(x)_+ = x$ for $x \geq 0$ and $(x)_+ = 0$ otherwise. Denote $\mathbf{T}_{n\times m} = \{\phi_\nu(x_i)\}_{i=1\,\nu=1}^{n\ m}$ and $\mathbf{\Sigma}_{n\times n} = \{R(x_i,x_j)\}_{i=1\,j=1}^{n\ n}$. Let $\mathbf{T} = (\mathbf{Q}_1\ \mathbf{Q}_2)(\mathbf{R}^T\ 0^T)^T$ be the QR decomposition of $\mathbf{T}$.

Kimeldorf and Wahba (1971) showed that the solution to (8) is

$$\hat{f}(x) = \sum_{\nu=1}^m d_\nu \phi_\nu(x) + \sum_{i=1}^n c_i R(x_i, x),$$

where $\mathbf{c} = (c_1, \ldots, c_n)^T$ and $\mathbf{d} = (d_1, \ldots, d_m)^T$ are solutions to

$$(\mathbf{\Sigma} + n\lambda\mathbf{C})\mathbf{c} + \mathbf{T}\mathbf{d} = \mathbf{Y}, \qquad \mathbf{T}^T\mathbf{c} = 0. \tag{9}$$

At the design points, $\hat{\mathbf{f}} = (\hat{f}(x_1), \ldots, \hat{f}(x_n))^T = \mathbf{AY}$, where

$$\mathbf{A} = \mathbf{I} - n\lambda\mathbf{C}\mathbf{Q}_2\big(\mathbf{Q}_2^T(\boldsymbol{\Sigma} + n\lambda\mathbf{C})\mathbf{Q}_2\big)^{-1}\mathbf{Q}_2^T$$

is the "hat" matrix. This "hat" matrix is often used to define degrees of freedom and to construct confidence intervals (Wahba, 1990). Note that $\mathbf{A}$ may be asymmetric, in contrast to the independent error case.

So far the smoothing parameter λ has been fixed. Good choices of λ are crucial to the performance of spline estimates (Wahba, 1990). Much research has been devoted to developing data-driven methods for selecting λ when observations are independent. Several methods have been proposed and among them the CV (cross-validation), GCV (generalized cross-validation), GML (generalized maximum likelihood) and UBR (unbiased risk) methods are popular choices (Wahba, 1990). The CV and GCV methods are well known for their optimal properties (Wahba, 1990). The GML is very stable and efficient for small-to-moderate sample sizes. When the dispersion parameter is known, for example, for binary and Poisson data, the UBR method works better. All these methods tend to underestimate smoothing parameters when data are correlated, for the reasons discussed in Section 2. In kernel regression, the correlation was often assumed to be short-range but otherwise unspecified, and bandwidth selection adjustments were proposed based on that assumption. In the smoothing spline literature, several authors have considered specific parametric assumptions on the correlation function.

Diggle and Hutchinson (1989) assumed the random errors are generated by an autoregressive process. In the following we discuss their method in a more general setting where we assume $\mathbf{C}$ is determined by a set of parameters $\boldsymbol{\alpha}$. Denote $\boldsymbol{\tau} = (\lambda, \boldsymbol{\alpha})$. Define $a = \operatorname{tr} \mathbf{A}$ as the effective degrees of freedom taken up in estimating f. Replacing the function f by its smoothing spline estimate $\hat{f}$, Diggle and Hutchinson (1989) proposed to estimate all parameters using the penalized profile likelihood (we use the negative log likelihood here)

$$\min_{\boldsymbol{\tau},\,\sigma^2}\big\{(\mathbf{Y}-\hat{\mathbf{f}})^T\mathbf{C}^{-1}(\mathbf{Y}-\hat{\mathbf{f}})/\sigma^2 + \ln|\mathbf{C}|/2 + n\ \ln\ \sigma^2 + \phi(n,a)\big\}, \tag{10}$$

where ϕ is a penalty function that is increasing in a. It is easy to see that $\hat{\sigma}^2 = (\mathbf{Y}-\hat{\mathbf{f}})^T\mathbf{C}^{-1}(\mathbf{Y}-\hat{\mathbf{f}})/n$, reducing (10) to

$$\min_{\boldsymbol{\tau}}\big\{n\ln(\mathbf{Y}-\hat{\mathbf{f}})^T\mathbf{C}^{-1}(\mathbf{Y}-\hat{\mathbf{f}}) + \ln|\mathbf{C}|/2 + \phi(n,a)\big\}.$$

Two forms of penalty have been compared in Diggle and Hutchinson (1989),

$$\phi(n,a) = -2n\ \ln(1-a/n),$$
$$\phi(n,a) = a\ \ln\ n,$$

that are analogs of AIC (Akaike information criterion) and BIC (Bayesian information criterion). When observations are independent, the first penalty gives a method approximating the GCV solution and the second penalty gives a new method which does not reduce to any existing methods. Simulation results in Diggle and Hutchinson (1989) suggest that the second penalty function works better than the first. However, Diggle and Hutchinson (1989) commented that using the second penalty gives results which are significantly inferior to those obtained by GCV when $\mathbf{C}$ is known (including independent data as the special case $\mathbf{C} = I$). More research is necessary to find properties of this method. Diggle and Hutchinson (1989) have developed an efficient $O(n)$ smoothing parameter selection algorithm in the special case of AR(1) error structures.

For independent observations, Wahba (1978) showed that a polynomial spline of degree $2m-1$ can be obtained by signal extraction. Denote $W(x)$ as a zero-mean Wiener process. Suppose that f is generated by the stochastic differential equation

$$d^m f(x)/dx^m = (n\lambda)^{-1/2}\sigma\ dW(x)/dx \tag{11}$$

with initial conditions

$$\mathbf{z}_0 = \big(f(0), f^{(1)}(0), \ldots, f^{(m-1)}(0)\big)^T \sim N(0, a\mathbf{I}_m). \tag{12}$$

Let $\hat{f}(x;a) = \mathrm{E}(f(x)|\mathbf{Y}, a)$ represent the signal extraction estimate of f. Then $\lim_{a\to\infty}\hat{f}(x;a)$ equals the smoothing spline estimate.

Kohn, Ansley and Wong (1992) used this signal extraction approach to derive a method for spline smoothing with autoregressive moving average errors. Assuming that observations are equally spaced ($x_i = i/n$), Kohn, Ansley and Wong (1992) considered model (1) with signal f generated by the stochastic model (11) and (12) and the errors ε_i generated by a discrete time stationary ARMA(p,q) model

$$\varepsilon_i = \phi_1\varepsilon_{i-1} + \cdots + \phi_p\varepsilon_{i-p} + e_i + \psi_1 e_{i-1} + \cdots + \psi_q e_{i-q}, \tag{13}$$

where $e_i \overset{\text{iid}}{\sim} N(0,\sigma^2)$ and are independent of the Wiener process $W(x)$ in (11).

Denote $\mathbf{z}_i = (f(x_i), f^{(1)}(x_i), \ldots, f^{(m-1)}(x_i))^T$. The stochastic model (11) and (12) can be written in a state space form as

$$\mathbf{z}_i = \mathbf{F}_i \mathbf{z}_{i-1} + \mathbf{u}_i, \qquad i = 1, \ldots, n,$$

where $\mathbf{F}_i$ is an upper triangular $m \times m$ matrix having ones on the diagonal and (j, k)th element $(x_i - x_{i-1})^{k-j}/(k-j)!$ for $k > j$. The perturbation $\mathbf{u}_i \sim N(0, \sigma^2 \mathbf{U}_i/n\lambda)$, where $\mathbf{U}_i$ is a $m \times m$ matrix with (j, k)th element being $(x_i - x_{i-1})^{2m-i-k+1}/[(2m-i-j+1)!(m-k)!(m-j)!]$.

For the ARMA(p, q) model (13), let $m' = \max(p, q+1)$ and

$$\mathbf{G} = \begin{pmatrix} \phi_1 & | & \\ \vdots & | & \\ \phi_{m'-1} & | & \mathbf{I}_{m'-1} \\ - & - & - \\ \phi_{m'} & | & 0^T \end{pmatrix}.$$

Consider the following state space model:

$$\mathbf{w}_i = \mathbf{G}\mathbf{w}_{i-1} + \mathbf{v}_i, \qquad i = 1, \ldots, n, \tag{14}$$

where $\mathbf{w}_i$ is a m' vector, and $\mathbf{v}_i = (e_i, \psi_1 e_i, \ldots, \psi_{m'-1} e_i)^T$. Substituting repeatedly from the bottom row of the system, it is easy to see that the first element in $\mathbf{w}_i$ is identical to the ARMA model defined in (13). Therefore the ARMA (p, q) model can be represented in a state space form (14).

Combining the two state space representations for the signal and the random error, the original model (1) can be represented by the following state space model:

$$\mathbf{Y}_i = \mathbf{h}^T \mathbf{x}_i,$$
$$\mathbf{x}_i = \mathbf{H}_i \mathbf{x}_{i-1} + \mathbf{a}_i,$$

where

$$\mathbf{x}_i = \begin{pmatrix} \mathbf{z}_i \\ \mathbf{w}_i \end{pmatrix}, \quad \mathbf{a}_i = \begin{pmatrix} \mathbf{u}_i \\ \mathbf{v}_i \end{pmatrix}, \quad \mathbf{H}_i = \begin{pmatrix} \mathbf{F}_i & 0 \\ 0 & \mathbf{G} \end{pmatrix}.$$

Here $\mathbf{h}$ is a $m + m'$ vector with 1 in the first and the $(m+1)$st positions and zeros elsewhere. Due to this state space representation, filtering and smoothing algorithms can be used to calculate the estimate of the function. Kohn, Ansley and Wong (1992) also derived algorithms to calculate GML and GCV estimates of all parameters $\boldsymbol{\tau} = (\lambda, \phi_1, \ldots, \phi_p, \psi_1, \ldots, \psi_q)$ and σ^2.

3.3 Long-range Dependence

In Section 2, we have seen that even under an AR(1) correlation structure, a simple type of short-range dependence, familiar nonparametric procedures intended for uncorrelated errors behave rather poorly. Several methods have been proposed in Sections 3.1 and 3.2 to handle the dependences. When the correlation $\rho(t) = \mathrm{Corr}(\varepsilon_i, \varepsilon_{i+t})$ decreases more slowly in t, regression estimation becomes even harder. In this section, we review the theoretical results on the effects of long-range dependent stationary Gaussian errors.

Estimation under long-range dependence has attracted more and more attention in recent years. In many scientific research fields, such as astronomy, physics, geoscience, hydrology and signal processing, the observational errors sometimes reveal long-range dependence. Künsch, Beran and Hampel (1993) wrote, "Perhaps most unbelievable to many is the observation that high-quality measurement series from astronomy, physics, chemistry, generally regarded as prototypes of 'i.i.d.' observations, are not independent but long-range correlated."

Minimax risks have been widely considered for evaluating performance of an estimator of a function assumed to be in a target class. Let $\mathscr{F}$ be a nonparametric (infinite-dimensional) class of functions on $[0, 1]^d$, where d is the number of predictors, and as before, let $\mathbf{C}$ denote the covariance matrix of the errors. Let $\|u - v\| = (\int (u-v)^2 \, d\mathbf{x})^{1/2}$ be the L_2 distance between two functions u and v. The minimax risk for estimating the regression function f under the squared L_2 loss is

$$R(\mathscr{F}; \mathbf{C}; n) = \min_{\hat{f}} \max_{f \in \mathscr{F}} E\|f - \hat{f}\|^2,$$

where $\hat{f}$ denotes any estimator based on $(\mathbf{X}_i, Y_i)_{i=1}^n$ and the expectation is taken with respect to the true regression function f. The minimax risk measures how well one can estimate f uniformly over the function class $\mathscr{F}$. Due to the difficulty in evaluating $R(\mathscr{F}; \mathbf{C}; n)$ in general, its convergence rate to zero as a function of the sample size n is often considered. An estimator with risk converging at the minimax risk rate uniformly over $\mathscr{F}$ is said to be minimax-rate optimal.

For short-range dependence, it has been shown that the minimax risk rate remains unchanged compared to the case with independent errors (see, e.g., Bierens, 1983; Collomb and Härdle, 1986; Johnstone and Silverman, 1997; Wang, 1996; Yang, 1997). However, fundamental differences show up when the errors become long-range dependent. For that case, a number of results have been obtained for parametric estimation of f (i.e., f is assumed to have a known parametric form) and also for the estimation of dependence parameters. For a review of the results, see Beran (1992, 1994). We focus here on nonparametric estimation of f.

For long-range dependent errors, results on minimax rates of convergence are obtained for univariate regression with equally spaced (fixed) design in

Hall and Hart (1990), Wang (1996), and Johnstone and Silverman (1997). The model being considered is again model (1), with $x_i = i/n$, $\mathrm{Corr}(\varepsilon_i, \varepsilon_j) \sim c|i-j|^{-\gamma}$ for some $0 < \gamma < 1$, and f is assumed to be smooth with the αth derivative bounded for some $\alpha \geq 1$ (the results of Wang (1996) and Johnstone and Silverman (1997) are in more general forms). The minimax rate of convergence under squared L_2 loss for estimating f is shown to be of order

$$n^{-2\alpha\gamma/(2\alpha+\gamma)}. \tag{15}$$

In contrast, the rate of convergence under independent or short-range dependent errors is $n^{-2\alpha/(2\alpha+1)}$. This shows the damaging effect of long-range dependence on the convergence rate.

Suppose now that $\mathbf{X}_i$, $i \geq 1$ are random variables, specified to be i.i.d. independent of the ε_i's. We consider the case where the dependence between the errors depends only on the orders of observations (the correlation has nothing to do with the values of the $\mathbf{X}_i$'s). Given $\mathscr{F}$, the distribution of $\mathbf{X}_i$, $1 \leq i \leq n$, and a general dependence among the errors (not necessarily stationary), Yang (1997) shows that the minimax risk rate for estimating f is determined by the maximum of two rates: the rate of convergence for the class $\mathscr{F}$ under i.i.d. errors and the rate of convergence for the estimation of the mean value of the regression function, $\mu = Ef(\mathbf{X})$, under the dependence model. The first rate is determined by the largeness of the target class $\mathscr{F}$ and the second rate is determined by severity of the dependence among the errors. As a consequence, the minimax rate may well remain unchanged if the dependence is not severe enough relative to the largeness of the target function class. A similar result was obtained independently by Efromovich (1999) in a univariate regression setting. It is also shown in Yang (1997) that dependence among the errors, as long as its form is known, generally does not hurt prediction of the next response.

When Yang's result is applied to the class of regression functions with the αth derivative bounded, one has the minimax rate of convergence

$$n^{-\min(2\alpha/(2\alpha+1),\,\gamma)}. \tag{16}$$

For a given long-range dependence index γ, the rate of convergence gets damaged only when α is relatively large, that is, $\alpha > \gamma/(2(1-\gamma))$. Note that the rate of convergence in (16) is always faster compared to the rate given in (15).

3.4 Wavelet Estimation

In this subsection, we review wavelet methods for regression on $[0, 1]$ under dependence focusing on the long-range dependence case.

Orthogonal series expansion is a commonly used method for function estimation. Compared with trigonometrics or Legendre polynomials, orthogonal wavelet bases have been shown to have desirable local properties that lead to optimal performance in statistical estimation for a rich collection of function classes. In the wavelet expansion of a function, the coefficients are organized in different levels called *multiresolutions*. The coefficients are estimated based on orthogonal wavelet transformation. Then one needs to decide which estimated coefficients are above the noise level and thus need to be kept in the wavelet expansion. This is usually done using a thresholding rule based on statistical considerations. Nason (1996) reported that methods intended for uncorrelated errors do not work well for correlated data. Johnstone and Silverman (1997) point out that for independent errors, one can use the same threshold for all the coefficients while for dependent errors, the variances of the empirical wavelet coefficients depend on the level but are the same within each level. Accordingly, level-dependent thresholdings are proposed. Their procedure is briefly described as follows.

Consider the regression model (1) with $n = 2^J$ for some integer J. Let $\mathscr{W}$ be a periodic discrete wavelet transform operator [for examples of wavelet bases and fast $O(n)$ algorithms; see, e.g., Donoho and Johnstone, 1998]. Let $w_{j,k} = (\mathscr{W}\mathbf{Y})_{j,k}$, $j = 0, \ldots, j-1$, $k = 1, \ldots, 2^j$ be the wavelet transform of the data $Y = (Y_1, \ldots, Y_n)^T$. Let $\mathbf{Z} = \mathscr{W}\varepsilon$ be the wavelet transform of the errors. Let λ_j be the threshold to be applied to the estimated coefficients at level j. Then define $\hat{\boldsymbol{\theta}} = (\hat{\theta}_{j,k})$, $j = 0, \ldots, J-1$, $k = 1, \ldots, 2^j$ by

$$\hat{\theta}_{j,k} = \eta(w_{j,k}, \hat{\sigma}_j\lambda_j),$$

where η is a threshold function and $\hat{\sigma}_j$ is an estimate of the standard deviation of $w_{j,k}$. The final estimator of the regression function is

$$\hat{\mathbf{f}} = \mathscr{W}^T\hat{\boldsymbol{\theta}},$$

where $\mathscr{W}^T$ is the inverse transform of $\mathscr{W}$. Earlier work of Donoho and Johnstone (see, e.g., 1998) suggest soft (S) or hard (H) thresholding as follows:

$$\eta_{\mathrm{S}}(w_{j,k}, \hat{\sigma}_j\lambda_j) = \mathrm{sgn}(w_{j,k})(|w_{j,k}| - \hat{\sigma}_j\lambda_j)_+$$

$$\eta_{\mathrm{H}}(w_{j,k}, \hat{\sigma}_j\lambda_j) = w_{j,k}I_{\{|w_{j,k}| \geq \hat{\sigma}_j\lambda_j\}}.$$

A suggested choice of $\hat{\sigma}_j$ is

$$\hat{\sigma}_j^2 = \mathrm{MAD}\{w_{j,k}, k = 1, \ldots, 2^j\}/0.6745,$$

where MAD means the median absolute deviation and the constant 0.6745 is derived to work for the Gaussian errors. To choose the λ_j's, Johnstone

and Silverman (1997) suggest a method based on Stein unbiased risk estimation (SURE) as described below. For a given m-dimensional vector $\mathbf{v} = (v_1, \ldots, v_m)$ (m will be taken to be 2^j at level j) and $\hat{\sigma}$, define a function

$$\widehat{U}(t) = \hat{\sigma} m + \sum_{k=1}^{m} \left\{ (v_k^2 \wedge t^2) - 2\hat{\sigma}^2 I_{\{|v_k| \le t\}} \right\}.$$

Then define

$$\hat{t}(\mathbf{v}) = \arg_{0 \le t \le \hat{\sigma}\sqrt{2 \log m}} \min\{\widehat{U}(t)\}.$$

By comparing $s_m^2 = m^{-1} \Sigma_{k=1}^m v_K^2 - 1$ with a threshold β_m, let

$$\tilde{t}(\mathbf{v}) = \begin{cases} \sqrt{2 \log m}, & s_m^2 \le \beta_m, \\ \hat{t}(\mathbf{v}), & s_m^2 > \beta_m. \end{cases}$$

Now choose the level-dependent threshold by

$$\lambda_j = \begin{cases} 0, & j \le L, \\ \tilde{t}(w_j/\hat{\sigma}_j), & L+1 \le j \le J-1, \end{cases}$$

where L is an integer specified by a user as the primary resolution level, below which signal clearly dominates over noise. From Johnstone and Silverman (1997), the regression estimator $\hat{f}$ produced by the above procedure converges at the minimax rate of convergence simultaneously over a rich collection of classes of smooth functions (Besov) without causing the need to know the long-range dependence index nor how smooth the regression function is. The adaptation is therefore over both the regression function classes and over the dependence parameter as well (see next subsection for a review of adaptive estimation for nonparametric regression).

3.5 Adaptive Estimation

Many nonparametric procedures have tuning parameters, for example, the bandwidth h for local polynomial and the smoothing parameter λ for the smoothing spline, as considered earlier. The optimal choices of such tuning parameters in general depend on certain characteristics of the unknown regression function and therefore are unknown to us. Various methods (e.g., AIC, BIC, cross-validation and other related model selection criteria) have been proposed to give a choice of a tuning parameter automatically based on the data so that the final estimator performs as well as (or nearly as well as) the estimator based on the optimal choice. This is the task of adaptive estimation. In this subsection, we briefly review the ideas of adaptive estimation in the context of nonparametric regression. This will provide a background for some of the results in the next section.

Basically there are two types of results on adaptive estimation, namely adaptive estimation with respect to a collection of estimation procedures (procedure-wise adaptation) and adaptive estimation with respect to a collection of target classes (target-oriented adaptation). For the first case, one is given a collection of procedures and the goal of adaptation is to have a final procedure performing close to the best one in the collection. For example, the collection might be a kernel procedure with all valid choices of the bandwidth h. Another collection may be a list of wavelet estimators based on different choices of wavelet bases. In general, one may have completely different estimation procedures in the collection for greater flexibility. This is desirable in applications where it is rather unclear beforehand which procedures are appropriate. For instance, for the case of high-dimensional regression estimation, one faces the so-called curse of dimensionality in the sense that the traditional function estimation methods (such as histogram and series expansion) would have exponentially many parameters in d to be estimated. This cannot be done accurately based on a moderate sample size. For this situation, a solution is to try different parsimonious models. Because one does not know which parsimonious characterization works best for the underlying unknown regression function, adaptation capability is desired.

For the second type of adaptation, one is given a collection of target function classes; that is, the true unknown function is assumed to be in one of the classes (without one's knowing which one it is). A goal of adaptation is to have an estimator with the capability to perform optimally simultaneously for the target classes; that is, the estimator automatically converges optimally at the minimax rate of the class that contains the true function.

The two types of adaptations are closely related. If each procedure in a collection is designed optimally for a specific function class in a collection, then the procedurewise adaptation implies the target-oriented adaptation. In this sense the procedurewise adaptation is more general than the target-oriented adaptation. In applications, one may encounter a mixture of both types of adaptation at the same time. On one hand, you have several plausible procedures that you wish to try, and on the other hand, you may have a few plausible characteristics (e.g., monotonicity, additivity) of the regression function you want to explore. For this situation, you may derive optimal (or at least reasonably good) estimators for each characteristic respectively, and then add them to the original collection of procedures. Then adaptation

with respect to the collection of procedures is the desired property.

A large number of results have been obtained on adaptive estimation with independent errors (see Yang, 2000, for references). More recently, for nonparametric regression with i.i.d. Gaussian errors, Yang (2000) shows that under very mild conditions, for any collection of uniformly bounded function classes, minimax-rate adaptive estimators exist. More generally, for any given collection of regression procedures, a single adaptive procedure can be constructed by combining them. The new procedure pays only a small price for adaptation; that is, the risk of the combined procedure is bounded above by the risk of each procedure plus a small penalty term that is asymptotically negligible for nonparametric estimation.

For nonparametric regression with dependence, adaptation with respect to both the unknown characteristics of the regression function and with respect to the unknown dependence is of interest. A success in this direction is the wavelet estimator based on Stein unbiased risk estimation proposed by Johnstone and Silverman (1997), as discussed in Section 3.4.

4. NEW DEVELOPMENTS

In the remainder of the article, we describe several new development areas related to smoothing in the presence of correlation. As mentioned at the beginning of the article, this discussion will be at a somewhat higher technical level than the previous material.

4.1 Kernel Regression Extensions

Opsomer (1997) introduces recent research that extends existing methodological results for kernel-based regression estimators under short-range dependence in several directions. The approach is fully nonparametric, uses local linear regression and implements a plug-in bandwidth estimator. The range of applications is extended to include random design, univariate and bivariate observations, and additive models. We will review some of the main findings here. Full details and proofs are available in Opsomer (1995). The method discussed below was used in Figure 1 to correct for the presence of correlation in the simulated example.

We again assume that the data are generated by model (1), where $\mathbf{X}_i$ are random and can be either scalars or bivariate vectors. Hence, the $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ are a set of random vectors in $\mathbb{R}^{d+1}$, with $d = 1$ or 2. The model errors ε_i are assumed to have the moment properties (2). In this general setting, we refer to model (1) as the *general* (G) model. We also consider two special cases: a model with univariate X_i as in Section 3.1, referred to as

$$\text{simple model (S1)}: \quad Y_i = f(x_i) + \varepsilon_i,$$

and the bivariate model, in which $f(\cdot)$ is assumed to be additive, that is,

$$\text{additive model (A2)}: \quad Y_i = \mu + f_1(X_{1i}) + f_2(X_{2i}) + \varepsilon_i.$$

We define the expected correlation funtion

$$c_n(\mathbf{x}) = n\mathrm{E}(\rho_n(\mathbf{X}_i - x)), \tag{17}$$

and let $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_n]^T$ represent the $n \times d$ matrix of covariates. Also, when $d = 2$, let $\mathbf{X}_{[k]} = (X_{k1}, \ldots, X_{kn})^T$ represent the kth column of $\mathbf{X}$, $k = 1, 2$. Let $\chi = [0, 1]^d$ represent the support of $\mathbf{X}_i$ and g its density function, with g_k the marginal density corresponding to X_{ki} for $k = 1, 2$. As in Section 3.1, let K represent a univariate kernel function. In order to simplify notation for model G, we restrict our attention to (tensor) product kernels $K(u_1) \times \cdots \times K(u_d)$, with corresponding bandwidth matrix $\mathbf{H} = \mathrm{diag}\{h_1, \ldots, h_d\}$.

For model A2, we need some additional notation. Let $\mathbf{T}_{12}^*$ represent the $n \times n$ matrix whose ijth element is

$$[\mathbf{T}_{12}^*]_{ij} = \frac{g(X_{1i}, X_{2j})}{g_1(X_{1i})g_2(X_{2j})} - \frac{1}{n},$$

and let $\mathbf{t}_i^T$, $\mathbf{v}_j$ represent the ith row and jth column of $(\mathbf{I} - \mathbf{T}_{12}^*)^{-1}$, respectively. Let

$$\mathbf{f}_1'' = \begin{bmatrix} \frac{d^2 f_1(X_{11})}{dx_1^2} \\ \vdots \\ \frac{d^2 f_1(X_{1n})}{dx_1^2} \end{bmatrix},$$

$$\mathrm{E}(f_1''(X_{1i})|\mathbf{X}_{[2]}) = \begin{bmatrix} \mathrm{E}(f_1''(X_{1i})|X_{21}) \\ \vdots \\ \mathrm{E}(f_1''(X_{1i})|X_{2n}) \end{bmatrix},$$

and analogously for $\mathbf{f}_2''$ and $\mathrm{E}(f_2''(X_{2i})|\mathbf{X}_{[1]})$.

The local linear estimator of $f(\cdot)$ at a point $\mathbf{x}$ for model G (and, with the obvious changes, model S1), is defined as $\hat{f}(\mathbf{x}) = \mathbf{s}_\mathbf{x}^T\mathbf{Y}$, with the vector $\mathbf{s}_\mathbf{x}^T$ defined as

$$\mathbf{s}_\mathbf{x}^T = \mathbf{e}_1^T(\mathbf{X}_\mathbf{x}^T\mathbf{W}_\mathbf{x}\mathbf{X}_\mathbf{x})^{-1}\mathbf{X}_\mathbf{x}^T\mathbf{W}_\mathbf{x}, \tag{18}$$

with $\mathbf{e}_1^T$ a row vector with 1 in its first position and 0's elsewhere, the weight matrix $\mathbf{W}_\mathbf{x} = \frac{1}{|\mathbf{H}|}\mathrm{diag}\{K(\mathbf{H}^{-1}(\mathbf{X}_1 - \mathbf{x})), \ldots, K(\mathbf{H}^{-1}(\mathbf{X}_n - \mathbf{x}))\}$ and

$$\mathbf{X}_\mathbf{x} = \begin{bmatrix} 1 & (\mathbf{X}_1 - \mathbf{x})^T \\ \vdots & \vdots \\ 1 & (\mathbf{X}_n - \mathbf{x})^T \end{bmatrix}.$$

For model A2, the estimator of $f(\cdot)$ at any location $\mathbf{x}$ can also be written as a linear smoother, but the expression is much more complicated and rarely directly used to compute the estimators. We assume here that the conditions guaranteeing convergence of the backfitting algorithm, further described in Opsomer and Ruppert (1997), are met. These conditions do not depend on the correlation structure of the errors.

Assumption (AS.I) from Section 3.1 is maintained, but the remaining assumptions on the statistical model and the estimator are replaced by the following:

(AS.II′) The density g is compactly supported, bounded and continuous, and $g(\mathbf{x}) > 0$ for all $\mathbf{x} \in \chi$,

(AS.III′) The second derivative(s) of $f(\cdot)$ are bounded and continuous.

(AS.IV′) As $n \to \infty$, $\mathbf{H} \to \mathbf{0}$ and $n|\mathbf{H}| \to \infty$.

In addition, we assume that the correlation function ρ_n is an element of a sequence $\{\rho_n\}$ with the following properties:

(R.I′) ρ_n is differentiable, $\int n|\rho_n(\mathbf{t}-\mathbf{x})|\,d\mathbf{t} = O(1)$, $\int n\rho_n(\mathbf{t}-\mathbf{x})^2\,d\mathbf{t} = o(1)$ for all $\mathbf{x}$.

(R.II′)

$$\exists\,\xi > 0\colon \int |\rho_n(\mathbf{t})| I_{(\|\mathbf{H}^{-1}\mathbf{t}\|>\xi)}\,d\mathbf{t} = o\left(\int |\rho_n(\mathbf{t})|\,d\mathbf{t}\right).$$

The properties require the effect of ρ_n to be "short-range" (relative to the bandwidth), but allow its functional form to be otherwise unspecified. These properties are generalizations of assumptions (R.I) and (R.II) in Section 3.1 to the random, multivariate design.

As noted in Section 3.1, the conditional bias of $\hat{f}(\mathbf{X}_i)$ is not affected by the presence of correlation in the errors. We therefore refer to Ruppert and Wand (1994) for the asymptotic bias of local polynomial estimators for models G and S1, and to Opsomer and Ruppert (1997) for the estimator of additive model A2. We construct asymptotic approximations to the conditional variance of $\hat{f}(\mathbf{X}_i)$ and to the conditional mean average squared error (MASE) of $\hat{f}$, defined in (5) for the fixed design case.

THEOREM 4.1. *The conditional variance of* $\hat{f}(\mathbf{X}_i)$ *for models* G *and* S1 *is*

$$\mathrm{Var}(\hat{f}(\mathbf{X}_i)|\mathbf{X}) = \sigma^2 \frac{1}{n|\mathbf{H}|}\frac{\mu_0(K^2)^d}{g(\mathbf{X}_i)}(1 + c_n(\mathbf{X}_i)) + o_p\left(\frac{1}{n|\mathbf{H}|}\right).$$

For model A2,

$$\mathrm{Var}(\hat{f}(\mathbf{X}_i)|\mathbf{X}) = \sigma^2\mu_0(K^2)\left(\frac{g_1(X_{1i})^{-1}(1+\mathrm{E}(c_n(\mathbf{X}_i|X_{1i})))}{nh_1} + \frac{g_2(X_{2i})^{-1}(1+\mathrm{E}(c_n(\mathbf{X}_i|X_{2i})))}{nh_2}\right) + o_p\left(\frac{1}{nh_1}+\frac{1}{nh_2}\right).$$

We let

$$R_n = n\int_{-1/2}^{1/2} \rho_n(\mathbf{t})\,d\mathbf{t}$$

and define $IC_n = \sigma^2(1+R_n)$, the *integrated covariance function*. We also define the second derivative regression functionals

$$\theta_{22}(k,l) = \mathrm{E}\left(\frac{\partial f(\mathbf{X}_i)}{\partial X_1}\frac{\partial f(\mathbf{X}_i)}{\partial X_2}\right)$$

with $k, l = 1, \ldots, d$ for models G and S1, and

$$\theta_{22}(1,1) = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{t}_i^T\mathbf{f}_1'' - \mathbf{v}_i^T\mathrm{E}\left(f_1''(X_{1i})|\mathbf{X}_{[2]}\right)\right)^2,$$

$$\theta_{22}(2,2) = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{v}_i^T\mathbf{f}_2'' - \mathbf{t}_i^T\mathrm{E}\left(f_2''(X_{2i})|\mathbf{X}_{[1]}\right)\right)^2,$$

$$\theta_{22}(1,2) = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{t}_i^T\mathbf{f}_1'' - \mathbf{v}_i^T\mathrm{E}\left(f_1''(X_{1i})|\mathbf{X}_{[2]}\right)\right) \times \left(\mathbf{v}_i^T\mathbf{f}_2'' - \mathbf{t}_i^T\mathrm{E}\left(m_2''(X_{2i})|\mathbf{X}_{[1]}\right)\right)$$

for model A2. Because of assumptions (AS.I) and (AS.II′)–(AS.IV′), these quantities are well defined and bounded, as long as the additive model has a unique solution.

THEOREM 4.2. *The conditional mean average squared error of* $\hat{f}$ *for model* G (S1) *is*

$$\mathrm{MASE}(\mathbf{H}|\mathbf{X}) = \left(\frac{\mu_2(K)}{2}\right)^2 \sum_{k=1}^{d}\sum_{l=1}^{d} h_k^2 h_l^2 \theta_{22}(k,l) + \frac{1}{n|\mathbf{H}|}\mu_0(K^2)^d IC_n + o_p\left(\sum_{k=1}^{d} h_k^4 + \frac{1}{n|\mathbf{H}|}\right).$$

For model A2,

$$\begin{aligned}\mathrm{MASE}(\mathbf{H}|\mathbf{X}) = {} & \left(\frac{\mu_2(K)}{2}\right)^2 \sum_{k=1}^{d}\sum_{l=1}^{d} h_k^2 h_l^2 \theta_{22}(k,l) \\ & + R(K)\sigma^2\left(\frac{1+\mathrm{E}(g_1(X_{1i})^{-1}c_n(\mathbf{X}_i))}{nh_1}\right. \\ & \left. + \frac{1+\mathrm{E}(g_2(X_{2i})^{-1}c_n(\mathbf{X}_i))}{nh_1}\right) \\ & + o_p\left(h_1^4 + h_2^4 + \frac{1}{nh_1} + \frac{1}{nh_2}\right).\end{aligned}$$

In this more general setting, the optimal rates of convergence for the bandwidths are again the same as those found for the independent errors case: $h_k = O_p(n^{-1/(4+d)})$ for models G and S1, and model A2 achieving the same rate as model S1. Similarly, if $R_n > 0$ for those models, the optimal bandwidths are larger than those for the uncorrelated case. For model A2, note that independent errors imply that $c_n(\mathbf{X}_i) = 0$, so that the result in Theorem 4.2 reduces to the MASE approximation derived in Opsomer and Ruppert (1997) for the independent errors case.

An interesting aspect of Theorem 4.2 is that the presence of correlated errors induces an adjustment in the MASE approximations for models G and S1 relative to the independent error case, which does *not* depend on the distribution of the observations. It is therefore easy to show that the MASE approximations for models G and S1 in Theorem 4.2 are valid for both random and fixed designs. For the additive model, the adjustment to the variance contains $c_n(\mathbf{X}_i)$ from (17), which depends on the covariate value and hence will vary depending on the design.

For the case $d = 1$, Opsomer (1995) develops a plug-in bandwidth selection method that generalizes the direct plug-in (DPI) bandwidth selection proposed by Ruppert, Sheather and Wand (1995) for the independent error univariate case and extended to the independent error additive model by Opsomer and Ruppert (1998). The method described here is therefore referred to as *correlation DPI* (CDPI), and was used as the correlation-adjusted bandwidth selection method in Figure 1. The estimation of the $\theta_{22}(k,l)$ in CDPI is analogous to that in DPI, and IC_n is estimated in the frequency domain by periodogram smoothing (Priestley, 1972) of the residuals of a pilot fit. CDPI behaves very much like DPI when the data are uncorrelated, but at least partly offsets the effect of the correlation when it is present, as illustrated in Figure 1. To illustrate this point on a real dataset, we will use the "drum roller" data analyzed by Laslett (1994) and Altman (1994) (the data are available on *Statlib*). As noted by both authors, the data appear to exhibit significant short-range correlation, so that analysis using a correlation-corrected method is warranted.

Figure 4 shows four fits to the data using both DPI (dash-dotted lines) and CDPI (solid lines): $n = 1150$ represents the full dataset, $n = 575$ uses every other observation, $n = 230$ every fifth and $n = 115$ every tenth. The remaining observations being located at increasing distance, it can be expected that the correlation should decrease with decreasing sample size. The plots in Figure 4 indeed exhibit this behavior, with the DPI fits nearly coinciding with the CDPI ones for the two smaller sample sizes. For the two larger sample sizes, CDPI manages to display an approximately "correct" shape for the mean function, while DPI results in a severely undersmoothed estimate.

4.2 Smoothing Spline ANOVA Models

All methods reviewed in Section 3.2 are developed for polynomial splines with special error structure. Some even require the design points to be equally spaced, so that their applications are limited to time series. In many applications, the data and/or the error structure are more complicated. Interesting examples are spatial, longitudinal and spatio-temporal data. The mean function of these kinds of data can be modeled in a unified fashion using the general spline models and the smoothing spline ANOVA models defined on arbitrary domains (Wahba, 1990). However, previous research on the general spline models and the smoothing spline ANOVA models assumed that the observations are independent. When data are correlated, which often is the case for spatial and longitudinal data, conventional methods for selecting smoothing parameters for these models face the same problems as illustrated in Section 2. Our goal in this section is to present extensions of the GML, GCV and UBR methods for smoothing spline ANOVA (SS ANOVA) models when observations are correlated.

Consider model (1) with fixed design points $\mathbf{x}_i = (x_i, \ldots, x_{di}) \in \chi = \chi_1 \otimes \cdots \otimes \chi_d$, where χ_i are measurable spaces of rather general form. We assume that f belongs to a subspace of tensor products of *reproducing kernel Hilbert spaces* (RKHS). More precisely, the model space $\mathscr{H}$ of a SS ANOVA model contains elements

$$\begin{aligned}f(\mathbf{x}) = {} & \mu + \sum_{j\in J_1} f_j(x_j) \\ & + \sum_{(j_1,j_2)\in J_2} f_{j_1,j_2}(x_{j_1}, x_{j_2}) \\ & + \cdots + \sum_{(j_1,\ldots,j_d)\in J_d} f_{j_1,\ldots,j_d}(x_{j_1},\ldots,x_{j_d}),\end{aligned} \tag{19}$$

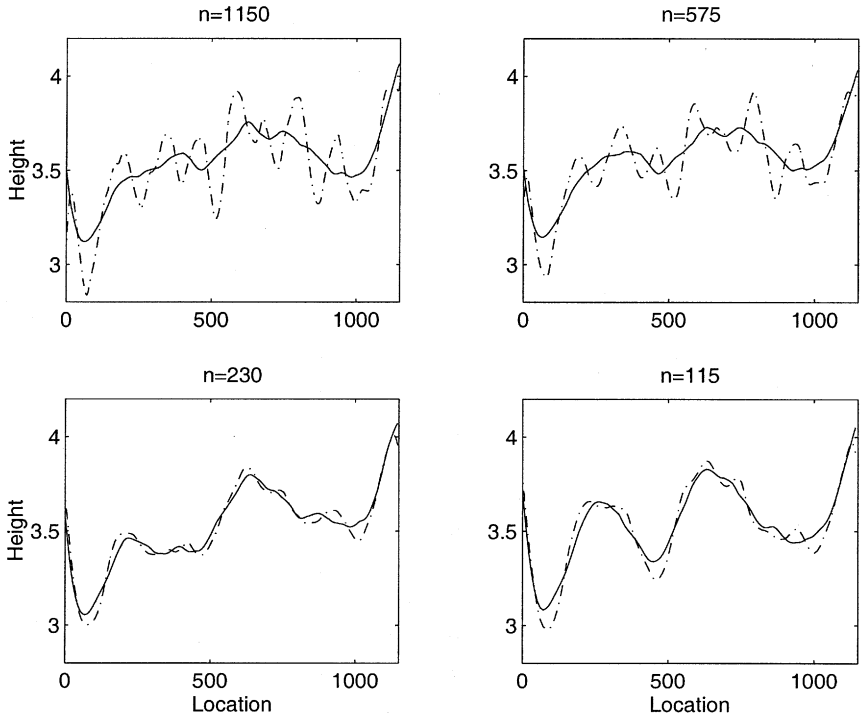

FIG. 4. *DPI*(– ·) *and DPI* (–) *fits to the drum roller data for four different sample sizes.*

where $\mathbf{x} = (x_1, \ldots, x_d) \in \chi$, $x_k \in \chi_k$, and J_k is a subset of the set of all k-tuples $\{(j_1, \ldots, j_k)\colon 1 \leq j_1 < \cdots < j_k \leq d\}$ for $k = 1, \ldots, d$. Identifiability conditions are imposed such that each term in the sums is integrated to zero with respect to any one of its arguments. Each term in the first sum is called a *main effect*, each term in the second sum is called a *two-factor interaction*, and so on. Similar to the analysis of variance, higher-order interactions are often eliminated from the model space to relieve the curse of dimensionality. See Aronszajn (1950) for details about RKHS and Wahba (1990), Gu and Wahba (1993) and Wahba et al. (1995) for details about SS ANOVA models. After a subspace is chosen as the model space, we can regroup and write it in the form

$$\mathscr{H} = \mathscr{H}_0 \oplus \sum_{j=1}^{p} \mathscr{H}_j,$$

where $\mathscr{H}_0$ is a finite-dimensional space containing functions which are not going to be penalized, and $\mathscr{H}_j$'s are subspaces which contain "smooth" elements in the decomposition (19).

Again, suppose that $\mathbf{C}$ is known up to a set of parameters $\boldsymbol{\alpha}$. No specific structure is assumed for $\mathbf{C}$, therefore it is not limited to the autoregressive or any special type of error structure. In practice, if the error structure is unknown, different structures may be fitted to select a final model. See Wang (1998a) for an example.

To illustrate potential applications of the SS ANOVA models with correlated errors, consider spatio-temporal data. Denote $\chi_1 = [0, 1]$ as the time domain and $\chi_2 = R^2$ as the spatial domain (latitude and longitude). Polynomial splines are often used to model temporal data and thin plate splines are often used to model spatial data. Thus the tensor product of two corresponding RKHS's can be used to model the mean function of a spatio-temporal data (Gu and Wahba, 1993). Components in model (19) can be interpreted as spatio-temporal main effects and interactions. An autoregressive structure may be used to model possible temporal correlation, and exponential structures may be used to model spatial correlation. Both correlations may appear in the covariance matrix $\mathbf{C}$.

A direct generalization of the penalized weighted least square (8) is

$$\min_{f \in \mathscr{H}} \left\{ \frac{1}{n} (\mathbf{Y} - \mathbf{f})^T \mathbf{C}^{-1} (\mathbf{Y} - \mathbf{f}) + \lambda \sum_{\beta=1}^{p} \theta_\beta^{-1} \| P_\beta f \|^2 \right\}, \tag{20}$$

where P_β is the orthogonal projection in $\mathscr{H}$ onto $\mathscr{H}_\beta$. Let $\phi_1, \ldots, \phi_M$ be a basis of $\mathscr{H}_0$ and $\mathbf{T}_{n \times M} = \{\phi_\nu(\mathbf{x}_i)\}_{i=1\nu=1}^{n\ M}$. Denote $\mathbf{T} = (\mathbf{Q}_1\ \mathbf{Q}_2)(\mathbf{R}^T\ 0^T)^T$ as the QR decomposition of $\mathbf{T}$. Let R_β be the reproducing kernel of $\mathscr{H}_\beta$, $\boldsymbol{\Sigma}_\beta = \{R_\beta(\mathbf{x}_i, \mathbf{x}_j)\}_{i=1\ j=1}^{n\ n}$, and $\boldsymbol{\Sigma} = \Sigma_{\beta=1}^{p} \theta_\beta \boldsymbol{\Sigma}_\beta$.

The solution to (20) is

$$\hat{f}(\mathbf{x}) = \sum_{\nu=1}^{M} d_\nu \phi_\nu(\mathbf{x}) + \sum_{i=1}^{n} c_i \sum_{\beta=1}^{p} \theta_\beta R_\beta(\mathbf{x}_i, \mathbf{x}), \tag{21}$$

where $\mathbf{c} = (c_1, \ldots, c_n)^T$ and $\mathbf{d} = (d_1, \ldots, d_M)^T$ are solutions to (9), but with $\mathbf{T}$ and $\mathbf{\Sigma}$ defined in this section.

Denote $\boldsymbol{\tau} = (\lambda/\theta_1, \ldots, \lambda/\theta_p, \boldsymbol{\alpha})$. We propose methods to estimate all parameters simultaneously. The GML method is derived from the following Bayes model:

$$Y_i = F(\mathbf{x}_i) + \varepsilon_i, \qquad i = 1, \ldots, n, \ \mathbf{x}_i \in \chi,$$

with prior defined by assuming that

$$F(\mathbf{x}) = \sum_{\nu=1}^{M} \eta_\nu \phi_\nu(\mathbf{x}) + (n\lambda)^{-1/2} \sigma \sum_{\beta=1}^{p} \theta_\beta^{1/2} Z_\beta(\mathbf{x}), \qquad \mathbf{x} \in \chi,$$

where $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_M)^T \sim N(0, a\mathbf{I})$, and $Z_\beta(\mathbf{x})$ are independent, mean zero Gaussian stochastic processes, independent of $\boldsymbol{\eta}$, with covariance $\mathrm{E} Z_\beta(\mathbf{x}) Z_\beta(\mathbf{u}) = R_\beta(\mathbf{x}, \mathbf{u})$. We assume that $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T \sim N(0, \sigma^2 \mathbf{C})$ and is independent of F.

It can be shown that when a approaches to infinity, the posterior mean of the Bayes model equals the smoothing spline estimate. That is, $\lim_{a\to\infty} \mathrm{E}(F(\mathbf{x})|\mathbf{Y}) = \hat{f}(\mathbf{x})$, where $\hat{f}(\mathbf{x})$ is given in (21). Let $B(\boldsymbol{\tau}) = \mathbf{\Sigma} + n\lambda\mathbf{C}$, where the dependencies on parameters are expressed explicitly. As argued in Wahba (1985), the maximum likelihood estimates of $\boldsymbol{\tau}$ should be based only on the marginal distribution of $\mathbf{z} = \mathbf{Q}_2^T \mathbf{Y}$. Accordingly, the generalized maximum likelihood (GML) estimates of $\boldsymbol{\tau}$ are maximizers of the log likelihood based on $\mathbf{z}$,

$$l_1(\boldsymbol{\tau}, \sigma^2|z) = -\frac{1}{2} \log \left| \frac{\sigma^2}{n\lambda} \mathbf{Q}_2^T \mathbf{B}(\boldsymbol{\tau}) \mathbf{Q}_2 \right| - \frac{n\lambda}{2\sigma^2} \mathbf{z}^T \left(\mathbf{Q}_2^T \mathbf{B}(\boldsymbol{\tau}) \mathbf{Q}_2 \right)^{-1} \mathbf{z} + \text{ constant}.$$

Maximizing l_1 with respect to σ^2, we have

$$\hat{\sigma}^2 = n\lambda \mathbf{z}^T \left(\mathbf{Q}_2^T \mathbf{B}(\boldsymbol{\tau}) \mathbf{Q}_2 \right)^{-1} \mathbf{z}/(n-m). \tag{22}$$

Then the GML estimates of $\boldsymbol{\tau}$ are maximizers of

$$l_2(\boldsymbol{\tau}|\hat{\sigma}^2) = -\frac{n-m}{2} \log \frac{\mathbf{z}^T \left(\mathbf{Q}_2^T \mathbf{B}(\boldsymbol{\tau}) \mathbf{Q}_2 \right)^{-1} \mathbf{z}}{\left[\det \left(\mathbf{Q}_2^T \mathbf{B}(\boldsymbol{\tau}) \mathbf{Q}_2 \right)^{-1} \right]^{\frac{1}{n-m}}}.$$

Equivalently, the GML estimates are the minimizers of

$$M(\boldsymbol{\tau}) = \frac{\mathbf{z}^T \left(\mathbf{Q}_2^T \mathbf{B}(\boldsymbol{\tau}) \mathbf{Q}_2 \right)^{-1} \mathbf{z}}{\left[\det \left(\mathbf{Q}_2^T \mathbf{B}(\boldsymbol{\tau}) \mathbf{Q}_2 \right)^{-1} \right]^{\frac{1}{n-m}}} = \frac{\mathbf{Y}^T \mathbf{C}^{-1} (\mathbf{I} - \mathbf{A}) \mathbf{Y}}{\left[\det{}^+ (\mathbf{C}^{-1}(\mathbf{I} - \mathbf{A})) \right]^{\frac{1}{n-m}}},$$

where $\det^+$ is the product of the nonzero eigenvalues.

Comparing the GML and GCV functions for independent observations, it is easy to see that a direct extension of the GCV function is the following:

$$V(\boldsymbol{\tau}) = \frac{\frac{1}{n} \| \mathbf{C}^{-1} (\mathbf{I} - \mathbf{A}) \mathbf{Y} \|^2}{\left[\frac{1}{n} \mathrm{Tr}(\mathbf{C}^{-1}(\mathbf{I} - \mathbf{A})) \right]^2}.$$

The GCV estimates of $\boldsymbol{\tau}$ are $\hat{\boldsymbol{\tau}} = \arg\min_{\boldsymbol{\tau}} V(\boldsymbol{\tau})$.

To introduce an extension of the UBR method, we define the weighted average squared errors (WASE) as $\mathrm{WASE} = \| \mathbf{C}^{-1} (\hat{\mathbf{f}} - \mathbf{f}) \|^2 / n$. Then,

$$\mathrm{E}(\mathrm{WASE}) = \frac{1}{n} \| \mathbf{C}^{-1} (\mathbf{I} - \mathbf{A}) \mathbf{f} \|^2 + \frac{\sigma^2}{n} \mathrm{Tr}(\mathbf{A}^T \mathbf{C}^{-2} \mathbf{A} \mathbf{C}).$$

An unbiased estimate of E(WASE) is

$$U(\boldsymbol{\tau}) = \frac{1}{n} \| \mathbf{C}^{-1} (\mathbf{I} - \mathbf{A}) \mathbf{Y} \|^2 - \frac{\sigma^2}{n} \mathrm{Tr}\ \mathbf{C}^{-1} + 2 \frac{\sigma^2}{n} \mathrm{Tr}\ (\mathbf{C}^{-1} \mathbf{A}). \tag{23}$$

The UBR estimates of $\boldsymbol{\tau}$ are $\hat{\boldsymbol{\tau}} = \arg\min_{\boldsymbol{\tau}} U(\boldsymbol{\tau})$. The UBR method needs an estimate of σ^2; one possible estimator is given in (22).

Wang (1998b) conducted simulations to compare the extended GML, GCV, UBR methods and Diggle and Hutchinson's (1989) method based on $\phi(n, d) = d \ln(n)$. It was found that the GCV and Diggle and Hutchinson methods are not stable when the sample size is small and/or the correlation is large. That is, there is a certain probability that the GCV and Diggle and Hutchinson methods select the smoothing parameter as zero, resulting in interpolation. This problem diminishes quickly when the sample size increases. Beside these obvious undersmoothed cases, the GCV method worked as well as the GML method for small to moderate sample sizes. The WASE of a spline estimate with the GCV choice of the smoothing parameter converges faster than the WASE of a spline estimate with the GML choice of the smoothing parameter. The Diggle and Hutchinson method works as well as the GML method for moderate to large sample sizes, but it fails badly for small sample sizes. The UBR method estimates the smoothing parameter very well, but estimates the correlation parameters poorly. Furthermore it needs an estimate of the variance. The GML method is stable and works very well for all situations. Therefore the GML method is recommended when the sample size is small to moderate. The GCV method is recommended when the sample size is large. Similar

results have been found for independent observations (Wahba, 1985, Kohn, Ansley and Tharm, 1991 and Wahba and Wang, 1993).

SS ANOVA models have connections to linear mixed-effects models. Consider the following linear mixed-effects model:

$$\mathbf{Y} = \mathbf{Td} + \sum_{\beta=1}^{p} \mathbf{u}_\beta + \boldsymbol{\varepsilon} = \mathbf{Td} + \mathbf{Zu} + \boldsymbol{\varepsilon},$$

where $\mathbf{d}$ is fixed, $\mathbf{u}_\beta$ is random and distributed as $\mathbf{u}_\beta \sim N(0, \sigma^2\theta_\beta\boldsymbol{\Sigma}_\beta/n\lambda)$, $\boldsymbol{\varepsilon} \sim N(0, \sigma^2\mathbf{C})$, $\mathbf{u}_\beta$'s and $\boldsymbol{\varepsilon}$ are mutually independent, $\mathbf{Z}_{n\times np} = (\mathbf{I}_n, \ldots, \mathbf{I}_n)$ and $\mathbf{u} = (\mathbf{u}_1^T, \ldots, \mathbf{u}_p^T)^T$. $\mathrm{Var}(\mathbf{u}) = \sigma^2\mathbf{D}/n\lambda$ where $\mathbf{D} = \mathrm{diag}(\theta_1\boldsymbol{\Sigma}_1, \ldots, \theta_p\boldsymbol{\Sigma}_p)$.

Writing $\mathbf{D}/n\lambda = (\mathbf{I})(\mathbf{D})(\mathbf{I}/n\lambda)$, the equation (3.3) in Harville (1976) is

$$\begin{pmatrix} \mathbf{T}^T\mathbf{C}^{-1}\mathbf{T} & \mathbf{T}^T\mathbf{C}^{-1}\mathbf{ZD} \\ \mathbf{DZ}^T\mathbf{C}^{-1}\mathbf{T} & n\lambda\mathbf{D} + \mathbf{DZ}^T\mathbf{C}^{-1}\mathbf{ZD} \end{pmatrix} \begin{pmatrix} \mathbf{d} \\ \boldsymbol{\phi} \end{pmatrix} = \begin{pmatrix} \mathbf{T}^T\mathbf{C}^{-1}\mathbf{Y} \\ \mathbf{DZ}^T\mathbf{C}^{-1}\mathbf{Y} \end{pmatrix}. \tag{24}$$

Let $\mathbf{c}$ and $\mathbf{d}$ be a solution to (9). Because $\mathbf{ZDZ}^T = \boldsymbol{\Sigma}$, it can be shown that $\mathbf{d}$ and $\boldsymbol{\phi} = \mathbf{Z}^T\mathbf{c}$ is a solution to (24) if $\boldsymbol{\Sigma}$ is invertible. The estimate of $\mathbf{u}$ is $\hat{\mathbf{u}} = \mathbf{D}\boldsymbol{\phi} = \mathbf{DZ}^T\mathbf{c}$. Thus, $\theta_\beta\boldsymbol{\Sigma}_\beta\mathbf{c} = \hat{\mathbf{u}}_\beta$, the smoothing spline ANOVA estimate of the component in the subspace $\mathscr{H}_\beta$, is a BLUP. Therefore the smoothing spline ANOVA estimates of the main effects, the interactions and the overall function are BLUP's. If $\boldsymbol{\Sigma}$ is not invertible, the smoothing spline estimate of the overall function $\hat{f} = T\mathbf{d} + \boldsymbol{\Sigma}\mathbf{c}$ is still a BLUP since it is unique. Furthermore, the GML estimates of parameters are also REML estimates because $\mathbf{z}$ represents $n - M$ linearly independent contrasts of $\mathbf{Y}$.

Because of the above connection between a SS ANOVA model and a linear mixed-effects model, SAS procedure `proc mixed` can be used to calculate coefficients $\mathbf{c}$ and $\mathbf{d}$ in (21). Note that the spline estimate $\hat{f}(\mathbf{x})$ is defined on the whole domain χ, whereas an estimate of a linear mixed-effects model is only defined on the design points. Our ultimate goal is to find a spline estimate; the relationships between smoothing spline models and mixed-effects models are used to achieve this goal. Several examples using `proc mixed` are available from the second author's homepage: `http://www.pstat.ucsb.edu/~yuedong`.

In the following, we apply the GML method to fit two datasets. We use the drum roller data introduced in Section 4.1 as our first example. The series consisting of the odd-numbered observations ($n = 575$) are fitted by a cubic spline for the deterministic mean function ($m = 2$) and an AR(1) model for errors. The left plot in Figure 5 shows the data (points), the estimate of f under the AR(1) model for the errors (solid line) and its 95% Bayesian confidence intervals (Wahba, 1983). We also plot the estimate of f under the independence assumption (dotted line) with the smoothing parameter selected by the GCV method. The estimate of f under the independence assumption is very variable. The estimates of the first-order autoregressive parameter and the residual variance are 0.3 and 0.35, respectively. The fitted mean function is comparable to that found by local linear regression using CDPI in Figure 4.

In the second example, we fit spatial data which consist of water acidity measurements (surface pH), the calcium concentration and geographic information (latitude and longitude) of 284 lakes in Wisconsin. This is a subset of the 1984 *Survey on Lakes in the USA* by the Environmental Protection Agency. Of interest is the dependence of the water acidity on the calcium concentration:

$$\mathrm{pH}_i = f(\mathrm{calcium}_i) + \varepsilon_i, \qquad i = 1, \ldots, 284. \tag{25}$$

The estimate of f under the independence assumption is variable (dotted line in the right plot of Figure 5), which indicates that the estimate of the smoothing parameter is too small. Measurements taken at lakes in relatively close proximity may be more highly correlated than those taken at more distant lakes. Therefore, an adjustment is needed to remove the effect of spatial correlation. We model the spatial correlation by an exponential structure with a nugget effect: $\mathrm{Var}(\varepsilon_i) = \sigma^2 + \sigma_1^2$ and $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = \sigma^2 \exp(-d_{ij}/\rho)$, where d_{ij} is the Euclidean distance between the geographic locations i and j. The estimates of the parameters σ^2, ρ and σ_1^2 are 0.13, 0.05 and 0.1, respectively. The estimate of f under this covariance structure is shown in the right plot of Figure 5 (solid line). We also fitted the data with a spherical spatial correlation structure and obtained similar estimates.

4.3 Rates of Convergence for Multidimensional Regression

Earlier in this section, local polynomial models and smoothing spline ANOVA models are studied for multidimensional regression. When the regression function is high dimensional, one often faces the so-called curse of dimensionality; that is, the traditional methods (e.g., series expansion up to a certain order, multidimensional kernel) often cannot provide a reasonably accurate estimate based on a moderately sized sample. In this situation, more parsimonious multivariate techniques have the potential to overcome the curse of dimensionality.

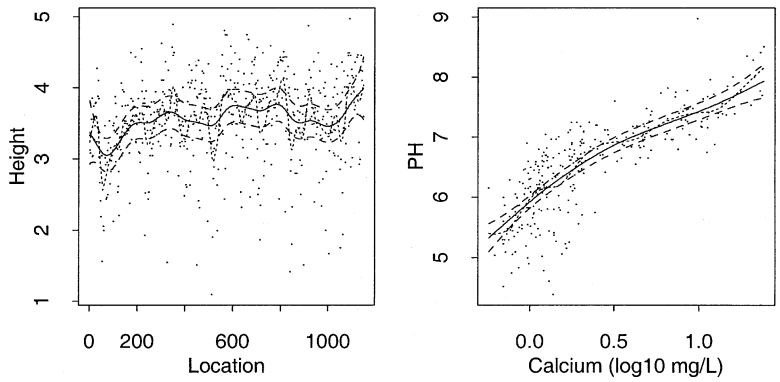

FIG. 5. *Left: drum Roller. Right: water acidity. Dots: observations. Solid lines: estimates with smoothing parameters estimated by GML. Dashed lines:* 95% *Bayesian confidence intervals. Dotted lines: estimates assuming independence and smoothing parameters are estimated by GML.*

Examples of such techniques are additive models as considered in Section 4.1, neural nets (Barron and Barron, 1988), CART (Breiman, Friedman, Olshen and Stone, 1984), projection pursuit (Friedman and Stuetzle, 1981), low-order tensor product splines (Stone, Hansen, Kooperberg and Truong, 1997) or smoothing spline ANOVA (as in Section 4.2). The challenge there lies in both the choice of a good parsimonious type of modeling among many candidates and also the choice of a good–sized model within the same type of modeling. Some adaptation results in that direction with independent errors are in Yang (2000). We here illustrate the advantages of additive or low-order interaction models for the case of dependent data by studying the rates of convergence of additive and low-order interaction classes.

Consider Sobolev classes with various interaction order and smoothness [the one-dimensional Sobolev space is defined in (7)]. For $r \geq 1$, let $\mathbf{z}_r = (z_1, \ldots, z_r) \in [0,1]^r$. For $\mathbf{k} = (k_1, \ldots, k_r)$ with nonnegative integer components k_i, define $|\mathbf{k}| = \sum_{i=1}^r k_i$. Let $D^{\mathbf{k}}$ denote the differentiation operator $D^{\mathbf{k}} = \partial^{|\mathbf{k}|}/\partial z_1^{k_1} \cdots \partial z_r^{k_r}$. For an integer α, define the Sobolev norm $\|g\|_{W_2^{\alpha,r}} = \|g\|_2 + \sum_{|\mathbf{k}|=\alpha} \int_{[0,1]^r} |D^{\mathbf{k}} g|^2 \, d\mathbf{z}_r$. Let $W_2^{\alpha,r}(C)$ denote the set of all functions g on $[0,1]^r$ with $\|g\|_{W_2^{\alpha,r}} \leq C$. Consider the following function classes of different interaction orders and smoothness:

$$S_1(\alpha; C) = \left\{ \sum_{i=1}^d g_i(x_i) \colon g_i \in W_2^{\alpha,1}(C), 1 \leq i \leq d \right\},$$

$$S_2(\alpha; C) = \left\{ \sum_{1 \leq i < j \leq d} g_{i,j}(x_i, x_j) \colon g_{i,j} \in W_2^{\alpha,1}(C), \right.$$
$$\left. 1 \leq i < j \leq d \right\},$$

$$S_d(\alpha; C) = W_2^{\alpha,d}(C),$$

with $\alpha \geq 1$ and $C > 0$. The simplest class $S_1(\alpha; C)$ contains additive functions (no interaction), and with larger r, functions in $S_r(\alpha; C)$ have higher order interactions. The unknown regression function f is assumed to be in (at least) one of the Sobolev classes but with r and α unknown.

Assume that the density of the covariates (with respect to Lebesgue measure) is supported in $[0,1]^d$ and is bounded above and away from zero. The errors are assumed to have a long-range dependence with known dependence parameter $0 < \gamma < 1$.

THEOREM 4.3. *Under the above conditions:*

(i) *The minimax rate of convergence for estimating a regression function in $S_r(\alpha; C)$ is $n^{-\min(2\alpha/(2\alpha+r), \gamma)}$, that is,*

$$\min_{\hat{f}} \max_{f \in S_r(\alpha, C)} E\|f - \hat{f}\|^2 = O\big(n^{-\min(2\alpha/(2\alpha+r), \gamma)}\big)$$

for $1 \leq r \leq d$ and $\alpha \geq 1$.

(ii) *Without knowing the hyperparameters α and the interaction order r, one can construct a minimax rate adaptive estimator. That is, a single estimator $\hat{f}^*$ can be constructed such that*

$$\max_{f \in S_r(\alpha, C)} E\|f - \hat{f}^*\|^2 = O\big(n^{-\min(2\alpha/(2\alpha+r), \gamma)}\big)$$

automatically for all $1 \leq r \leq d$ and $\alpha \geq 1$.

The first result of the theorem suggests the advantage of additive or low interaction order models. For a fixed hyperparameter α for the Sobolev classes, if one knew the true interaction order r, then a faster convergence rate could be achieved compared to an estimator obtained assuming the highest interaction order (i.e., $r = d$). The improvement in rate of convergence is substantial when r is small compared to d. For similar results on effects of dimension reduction with independent errors,

see Stone (1994), Nicoleris and Yatracos (1997) and Yang (1999).

Because the smoothness parameter α and the interaction order are unknown in most applications, adaptive estimators of f relative to these parameters are desired. The second result of the above theorem guarantees the existence of such adaptive estimators. General adaptive estimation with unknown dependence γ is currently being investigated by one of the authors of this paper. For the Sobolev (and more generally Besov) classes, as studied in Wang (1996) and Johnstone and Silverman (1997) for the one-dimensional case, wavelet estimators are a natural consideration. It seems reasonable to expect that, with proper thresholding and a method to select the interaction order r, tensor-product wavelet estimators will have the desired adaptation capability.

5. CONCLUSION

Correlation is a common occurrence in practical applications. In nonparametric regression, this correlation can have important consequences on the statistical properties of the estimator and on the selection of the smoothing parameter using data-driven methods, as shown in this article. We have reviewed the existing literature on the effect of correlation for several important types of nonparametric regression methods, including kernel, spline and wavelet methods and discussed some recent work in these areas. Clearly, the list of nonparametric regression techniques discussed in this article is not exhaustive. Other approaches could include a fully Bayesian approach in which formal priors are defined for the correlation function and on the structure of the mean function.

The techniques reviewed in this article were shown to be able to handle the correlation, if its shape can either be parametrically specified or if the errors are assumed to be short-range dependent. For such correlation structures, data-driven smoothing parameter selection methods are available, as discussed in Sections 3 and 4.

Currently, the theoretical results for kernel-based and wavelet expansion methods under correlation are more extensive than those for smoothing splines. In the former two types of method, it was shown that the convergence rates of the estimators are unaffected by short-range dependence or the errors. Under long-rang dependence, however, the rate of convergence can be severely damaged, but level-dependent thresholding for wavelet estimation still leads to optimal rates of convergence. Theoretical results for adaptive estimation, a useful approach for modelling for high-dimensional datasets under the assumption of independent errors, are shown to continue to hold under several correlation scenarios.

There is still ample room for future methodological research in nonparametric regression with correlated errors. For kernel-based methods, there is a need for data-driven bandwidth selection methods for higher dimensional datasets, especially spatial data where correlation is induced by the location of the observations. For smoothing splines, research on their large-sample properties under short-range and long-range dependence is still lacking, as well as methods for estimating smoothing parameters when the correlation is not parametrically specified. The results on adaptation also need to be extended to more general correlation structures for greater flexibility and applicability.

Perhaps the most pressing area for future research is the implementation of many of the existing theoretical and methodological results into practical algorithms and software. This would provide users of nonparametric regression with tools to analyze datasets with correlated observations with as much ease as those with assumed independence.

ACKNOWLEDGMENTS

Yuedong Wang is supported in part by NIH Grant R01 GM58533. Yuhong Yang was supported in part by NSA Grant MDA 9049910060.

REFERENCES

ADENSTEDT, R. K. (1974). On large sample estimation for the mean of a stationary sequence. *Ann. Statist.* **2** 1095–1107.

ALTMAN, N. (1994). Krige, smooth, both or neither? Technical report, Biometrics Unit, Cornell Univ.

ALTMAN, N. S. (1990). Kernel smoothing of data with correlated errors. *J. Amer. Statist. Assoc.* **85** 749–759.

ARONSZAJN, N. (1950). Theory of reproducing kernels. *Trans. Amer. Math. Soc.* **68** 337–404.

BARRON, A. R. and BARRON, R. L. (1988). Statistical learning networks: a unifying view. In *Computer Science and Statistics: Proceedings of the 21st Interface*, 192–203.

BERAN, J. (1992). Statistical methods for data with long-range dependence. *Statist. Sci.* **7** 404–416.

BERAN, J. (1994). *Statistics for Long-Memory Processes*. Chapman and Hall, New York.

BIERENS, H. (1983). Uniform consistency of kernel estimators of a regression function under generalized conditions. *J. Amer. Statist. Assoc.* **78** 699–707.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. and STONE, C. J. (1984). *Classification and Regression Trees*. Wadsworth, Belmont, CA.

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

CHU, C.-K. and MARRON, J. S. (1991). Comparison of two bandwidth selectors with dependent errors. *Ann. Statist.* **19** 1906–1918.

COLLOMB, G. and HÄRDLE, W. (1986). Strong uniform convergence rates in robust nonparametric time series analysis. *Stochastic Process. Appl.* **23** 77–89.

COX, D. R. (1984). Long-range dependence: a review. In *Statistics: An Appraisal. Proceedings 50th Anniversary Conference*, (H. A. David and H. T. David, eds.) 55–74. Iowa State Univ. Press.

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

DONOHO, D. L. and JOHNSTONE, I. M. (1998). Minimax estimation via wavelet shrinkage. *Ann. Statist.* **26** 879–921.

EFROMOVICH, S. (1999). How to overcome curse of long-memory errors in nonparametric regression. *IEEE Trans. Inform. Theory* **45**, 1735–1741.

FRIEDMAN, J. and STUETZLE, W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.* **76** 817–823.

GU, C. and WAHBA, G. (1993). Semiparametric ANOVA with tensor product thin plate spline. *J. Roy. Statist. Assoc. Ser. B* **55** 353–368.

HALL, P. and HART, J. D. (1990). Nonparametric regression with long-range dependence. *Stochastic Process. Appl.* **36** 339–351.

HALL, P. Lahiri, S. N. and POLZEHL, J. (1995). On bandwidth choice in nonparametric regression with both short- and long-range dependent errors. *Ann. Statist.* **23** 1921–1936.

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** 86–95.

HÄRDLE, W., HALL, P. and MARRON, J. S. (1992). Regression smoothing parameters that are not far from their optimum. *J. Amer. Statist. Assoc.* **87** 227–233.

HART, D. (1997). *Nonparametric Smoothing and Lack-of-Fit Tests*. Springer, New York.

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

HART, J. D. (1994). Automated kernel smoothing of dependent data by using time series cross-validation. *J. Roy. Statist. Assoc. Ser. B* **56** 529–542.

HARVILLE, D. (1976). Extension of the Gauss–Markov theorem to include the estimation of random effects. *Ann. Statist.* **4** 384–395.

HERRMANN, E., GASSER, T. and KNEIP, A. (1992). Choice of bandwidth for kernel regression when residuals are correlated. *Biometrika* **79** 783–795.

JOHNSTONE, I. and SILVERMAN, B. (1997). Wavelet threshold estimators for data with correlated noise. *J. Roy. Statist. Assoc. Ser. B* **59** 319–351.

KIMELDORF, G. S. and WAHBA, G. (1971). Some results on Tchebycheffian spline functions. *J. Math. Anal. Appl.* **33** 82–94.

KOHN, R., ANSLEY, C. F. and THARM, D. (1991). The performance of cross-validation and maximum likelihood estimators of spline smoothing parameters. *J. Amer. Statist. Assoc.* **86** 1042–1050.

KOHN, R., ANSLEY, C. F. and WONG, C. (1992). Nonparametric spline regression with autoregressive moving average errors. *Biometrika* **79** 335–346.

KÜNSCH, H., BERAN, J. and HAMPEL, F. (1993). Contrasts under long-range correlations. *Ann. Statist.* **21** 943–964.

LASLETT, G. (1994). Kriging and splines: an empirical comparison of their predictive performance in some applications. *J. Amer. Statist. Assoc.* **89** 392–409.

NASON, G. P. (1996). Wavelet shrinkage using cross-validation. *J. Roy. Statist. Soc. Ser. B* **58** 463–479.

NICOLERIS, T. and YATRACOS, Y. G. (1997). Rate of convergence of estimates, Kolmogrov's entropy and the dimensionality reduction principle in regression. *Ann. Statist.* **25** 2493–2511.

OPSOMER, J.-D. (1995). Estimating a function by local linear regression when the errors are correlated. Preprint 95-42, Dept. Statistics, Iowa State Univ.

OPSOMER, J.-D. (1997). Nonparametric regression in the presence of correlated errors. In *Modelling Longitudinal and Spatially Correlated Data: Methods, Applications and Future Directions* (T. Gregoire, D. Brillinger, P. Diggle, E. Russek-Cohen, W. Warren and R. Wolfinger, eds.) 339–348. Springer, New York.

OPSOMER, J.-D. and RUPPERT, D. (1997). Fitting a bivariate additive model by local polynomial regression. *Ann. Statist.* **25** 186–211.

OPSOMER, J.-D. and RUPPERT, D. (1998). A fully automated bandwidth selection method for fitting additive models by local polynomial regression. *J. Amer. Statist. Assoc.* **93** 605–619.

PRIESTLEY, M. B. and CHAO, M. T. (1972). Nonparametric function fitting. *J. Roy. Statist. Assoc. Ser. B* **34** 385–392.

RUPPERT, D., SHEATHER, S. J. and WAND, M. P. (1995). An effective bandwidth selector for local least squares regression. *J. Amer. Statist. Assoc.* **90** 1257–1270.

RUPPERT, D. and WAND, M. P. (1994). Multivariate locally weighted least squares regression. *Ann. Statist.* **22** 1346–1370.

STONE, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation. *Ann. Statist.* **22** 118–184.

STONE, C. J. HANSEN, M. H. KOOPERBERG, C. and TRUONG, Y. (1997). Polynomial splines and their tensor products in extended linear modeling (with discussion). *Ann. Statist.* **25** 1371–1470.

WAHBA, G. (1978). Improper priors, spline smoothing, and the problem of guarding against model errors in regression. *J. Roy. Statist. Assoc. Ser. B* **40** 364–372.

WAHBA, G. (1983). Bayesian confidence intervals for the cross-validated smoothing spline. *J. Roy. Statist. Assoc. Ser. B* **45** 133–150.

WAHBA, G. (1985). A comparison of GCV and GML for choosing the smoothing parameters in the generalized spline smoothing problem. *Ann. Statist.* **4** 1378–1402.

WAHBA, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia.

WAHBA, G. and WANG, Y. (1993). Behavior near zero of the distribution of GCV smoothing parameter estimates for splines. *Statist. Probab. Lett.* **25** 105–111.

WAHBA, G. WANG, Y., GU, C., KLEIN, R. and KLEIN, B. (1995). Smoothing spline ANOVA for exponential families, with application to the Wisconsin Epidemiological Study of Diabetic Retinopathy. *Ann. Statist.* **23** 1865–1895.

WANG, Y. (1996). Function estimation via wavelet shrinkage for long-memory data. *Ann. Statist.* **24** 466–484.

WANG, Y. (1998a). Mixed-effects smoothing spline ANOVA. *J. Roy. Statist. Assoc. Ser. B* **60** 159–174.

WANG, Y. (1998b). Smoothing spline models with correlated random errors. *J. Amer. Statist. Assoc.* **93** 341–348.

YANG, Y. (1997). Nonparametric regression with dependent errors. Preprint 97-29, Dept. Statistics, Iowa State Univ. (A shorter version is accepted in *Bernoulli.*)

YANG, Y. (1999). Model selection for nonparametric regression. *Statist. Sinica* **9** 475–499.

YANG, Y. (2000). Combining different procedures for adaptive regression. *J. Multivariate Anal.* **74** 135–161.