Cowles Foundation Paper 593

Smoothness Priors and Nonlinear Regression

ROBERT J. SHILLER*

Smoothness priors represent prior information that an unknown function does not change slope quickly and hence that the function describes a simple curve (e.g., Wahba 1978). In this article such priors for the multiple nonlinear regression model are developed in such a way that estimates and "standard errors" can be obtained as a natural and conceptually straightforward extension of linear multiple-regression estimation with the addition of dummy variables and dummy observations. Relations to spline and polynomial interpolation are described. An illustrative example of cost-function estimation is provided.

KEY WORDS: Nonparametric regression; Spline smoothing; Polynomial regression.

1. INTRODUCTION

Those who use the multiple-regression model in applied work often find themselves seeking simple expedients to allow for nonlinearity in an independent variable. This article offers another such simple expedient based on recent Bayesian literature on spline smoothing.

In most applications of the linear regression model, there is apparently no theory that the relation should be linear. Linearity is assumed for simplicity. Similarly when, in such applications, concern is felt that a particular independent variable might have a distinctly nonlinear effect, there is probably little real knowledge about the nature of the nonlinear function that describes this effect. Thus some simple way of dealing with the nonlinearity is sought. In these situations the same sense of parsimony that inclines so many researchers to use the simple linear models often inclines those concerned with nonlinearity to modify the linear model minimally as by allowing only one (or a couple) of the independent variables to have a nonlinear effect.

Thus one often sees reported regression results in which, for example, the squared value of an independent variable is added as an extra independent variable. Such an application of polynomial regression may be called a parametric method because the nonlinear function is restricted to a certain class of functions. In this case the function must be parabolic. There is a question whether one's intuitive modeling sense is generally well served by such simple parametric methods—that is, whether one's actual intangible prior knowledge of the function is well represented by the prior restrictions implicit in the model. Did one really want to restrict the function to the U shape of the parabola when this was done?

The regression methods proposed here may be thought of as rubber-ruler methods. The straight hard ruler of strictly linear multiple regression is softened somewhat for an independent variable. The resulting estimators are restrictive only in the sense that a rubber ruler is restrictive. These estimators allow the estimated function to describe, for example, a parabola or a function with an asymptote as well. The multiple-regression estimation methods described here cannot be reduced in any useful way to univariate curve fitting with transformed data. In experiments with a variety of known smooth functions, the estimators produced nearly the true functions, even when their shape could not be spotted visually in any scatter diagram.

The model that will be considered here allows for nonlinearity in one independent variable:

$$y_t = f(x_{t1}) + x_{t2}\gamma + \epsilon_t, \tag{1}$$

where y_t is the tth observation of the dependent variable (a scalar), $f(\cdot)$ is an unknown function, x_{t1} is the tth observation of the first independent variable (a scalar), x_{t2} is the tth observation of the g element row vector of other independent variables, γ is a g element column vector of unknown regression coefficients, and ϵ_t is the unknown error term. (The natural generalization of (1) suggested by the approach here is $y_i = \sum_j f_j(x_{ij}) + \epsilon_i$—i.e., an additively separable model.) The vector ϵ, whose ith element is ϵ_i, is assumed independent of all observations of the independent variables and spherically normal with zero mean and variance σ^2. There are n observations of the vector $[x_{t1}, x_{t2}, y_t]$ $(t = 1, \ldots, n)$, ordered in terms of increasing x_{t1}.

The prior notion of smoothness used here is that the slope of the unknown function $f(\cdot)$ probably does not change too fast. Thus, loosely speaking, the priors will give high probability to any function that can easily be drawn with a rubber ruler—that is, for which one does not have to bend the ruler too hard. The rubber-ruler analogy is apt because the prior density used here has an analogy to the potential energy in an elastic beam subjected to loads at discrete points (see Sokolnikoff 1956).

* Robert J. Shiller is Professor, Department of Economics and School of Organization and Management, Yale University, and Research Associate, National Bureau of Economic Research; Cowles Foundation for Research in Economics, Yale University, Box 2125, Yale Station, New Haven, CT 06520. The author thanks Robert Engle, Jerry Hausman, Peter Phillips, Dale Poirer, and anonymous referees for helpful comments. Danny Quah and Nigel Wilson provided research assistance. Estimation was performed with the TROLL System, M.I.T. This research was supported by National Science Foundation Grant SES-8105837.

Reprinted from: © Journal of the American Statistical Association
September 1984, Volume 79, Number 387
Theory and Methods Section

Here the likelihood function will supply the loads at observation points. The priors are partially uninformative. They carry information about changes in slope but not about slope or individual values of the function. The primary estimation method developed here depends on a special case of a class of priors due to Wahba (1978) based on earlier work by Kimeldorf and Wahba (1970). I call this case *continuous smoothness priors*. They are continuous because the priors effectively concern derivatives of the unknown continuous function. In Section 5 slightly simpler priors (Whittaker and Robinson 1967, Shiller 1973, Gersovitz and MacKinnon 1978), which I call *discrete smoothness priors*, will be discussed. They are discrete because the priors concern differences (measured across discrete points) of the unknown function.

The methods developed here are explicitly Bayesian and in this respect are quite different from other estimation methods used for unknown functions, such as the nearest-neighbor methods (Cover 1968 and Stone 1977), the recursive partitioning method (Breiman and Meisel 1976), or the projection pursuit method (Friedman and Stuetzle 1981). The methods discussed are closer to those of Blight and Ott (1975) or Oman (1982). The estimation method produced here is nonparametric in the sense that the prior does not absolutely rule out any value for the estimate of the vector $f(x_{i1})$, $(i = 1, \ldots, n)$.

2. THE PRIOR AND POSTERIOR DISTRIBUTION FOR f and γ

The estimation approach here is designed to facilitate estimation of the values and corresponding "standard errors" of the function $f(\cdot)$ at an arbitrarily specified list of values of the argument of the function, not just those values for which there are observations. Thus the researcher may choose to estimate the function at more or less equally spaced values of the argument even if the observations are unequally spaced, or the researcher may choose to view the function more finely in an interval of the argument in which there is particular interest. For this, one must write the likelihood function for the model (1) in an unusual form that involves redundant parameters.

The $f(\cdot)$ in (1) will be estimated at $N \geq 3$ distinct points $\bar{x}_{i1}$ $(i = 1, \ldots, N)$ ordered in terms of increasing $\bar{x}_{i1}$, where all observed values of x_{i1} are included among $\bar{x}_{i1}$, but $\bar{x}_{i1} \neq \bar{x}_{j1}$ unless $i = j$. Write the $n \times (N + g)$ element matrix $X = [X_1 : X_2]$. The $n \times N$ element matrix X_1 has elements X_{1ij} equal to zero, except in $x_{i1} = \bar{x}_{j1}$, where the element is 1. One may think of the columns of X_1 as *dummy variables* representing points along the function. The $n \times g$ element matrix X_2 has x_{i2} as its ith row. With the n element vector, Y—whose ith element is y_i—and the $N + g$ element vector $\beta' = [f' : \gamma']'$, where the ith element of the N element column vector f is $f(\bar{x}_{i1})$, one can write the likelihood function

$$L(Y \mid \beta, h, X) \propto h^{n/2}\exp[-h(Y - X\beta)'(Y - X\beta)/2], \quad (2)$$

where h is the *precision* equal to $1/\sigma^2$. If there is only one observation at each point $\bar{x}_{i1}$ $(i = 1, \ldots, N)$, then the likelihood function will be saturated so that maximum likelihood estimation is not feasible.

The intuitive notion behind continuous smoothness priors is that the derivatives of the unknown function do not change too fast as the argument changes, and thus the function is smooth and not choppy or irregular. This suggests that the second derivatives of the function at all points on the real line are probably small. One may wish to make the prior on the second derivatives spherically normal. This intuitive notion will be formalized later, by forming a prior representing that the unknown function is a realization of the integral of a Wiener process for which the standard deviation of an increment is small. Wahba's (1978) priors were more general: she allowed the $(m - 1)$th integral $m > 0$ of a Wiener process rather than merely the first integral. For m other than 2, her priors do not have the same interpretation as a softening of the linearity restriction in ordinary regression.

To derive continuous smoothness priors then, which will be *cylindrically uniform* priors (as in Leamer 1978), let $Z(t)$ be a stochastic process such that $Z(0) = 0$, $dZ(0) = 0$, where $Z(t)$ is the integral of a unit Wiener process and $dZ(t)$ is the stochastic differential of $Z(t)$. Then the autocovariance function for Z is the integral of the autocovariance function of a Wiener process:

$$q(s, t) \equiv E(Z(s) \cdot Z(t)) = \int_0^s \int_0^t \min(s', t')ds'dt'.$$

Then

$$\begin{aligned} q(s, t) &= s^2t/2 - s^3/6, \qquad s \leq t \\ &= st^2/2 - t^3/6, \qquad s \geq t. \end{aligned} \quad (3)$$

Let Q_t be the $N \times N$ matrix whose ijth element is $q((\bar{x}_{i1} + t), (\bar{x}_{j1} + t))$. Clearly, as t approaches infinity, all elements of Q_t approach infinity. As one moves away from initial conditions at $t = 0$, the variance of this nonstationary process approaches infinity. However, Q_t^{-1} approaches a finite nonzero limit. The prior distribution for f will be multivariate normal with zero mean and precision matrix (the inverse of the prior variance matrix) equal to $\xi^{-2} \lim_{t\to\infty} Q_t^{-1}$, where ξ is the standard deviation of the change in slope (derivative) of the function over a one-unit interval in x_1. Because the prior uses $\lim_{t\to\infty} Q_t^{-1}$ rather than Q_t^{-1} for finite t, the prior is partially uninformative as noted in the introduction.

Define the $(N - 2) \times N$ matrix R such that Rf is a vector whose elements correspond to the slope changes in the function, so our notion of smoothness of the function can be described as implying that the elements of Rf are probably small.

$$\begin{aligned} R_{ij} &= \Delta_i^{-1}, && i = j \\ &= -(\Delta_i^{-1} + \Delta_{i+1}^{-1}), && i = j - 1 \\ &= \Delta_{i+1}^{-1}, && i = j - 2 \\ &= 0, && \text{otherwise}, \end{aligned} \quad (4)$$

where $\Delta_i \equiv (\bar{x}_{i+1,1} - \bar{x}_{i1})$. Thus

$$R_i f = f_i \Delta_i^{-1} - f_{i+1}(\Delta_i^{-1} + \Delta_{i+1}^{-1}) + f_{i+2}\Delta_{i+1}^{-1}$$
$$= (f_i - f_{i+1})/\Delta_i - (f_{i+1} - f_{i+2})/\Delta_{i+1}$$

is the difference in the slope between a line connecting $(\bar{x}_{i1}, f_i), (\bar{x}_{i+1,1}, f_{i+1})$ and a line connecting $(\bar{x}_{i+1,1}, f_{i+1})$, $(\bar{x}_{i+2,1}, f_{i+2})$. Thus if the chosen values of $\bar{x}_1$ are spaced at unit intervals, then Rf is the vector of second differences of the function values at these points. Define the $N \times 2$ matrix A by $A_{i1} = 1$ and $A_{i2} = \bar{x}_{i1}$ $(i = 1, \ldots, N)$. The $RA = 0$. The variance matrix B_t of $[Z(t), dZ(t)]'$ is given by $B_{t11} = t^3/3$, $B_{t12} = B_{t21} = t^2/2$, and $B_{t22} = t$. Hence $Q_t = AB_tA' + Q_0$. Thus H_t defined as RQ_tR' equals $R(AB_tA' + Q_0)R'$, which equals RQ_0R', and hence H does not depend on t. Multiplying, one finds that

$$\begin{aligned} H_{ij} &= (\Delta_i + \Delta_{i+1})/3, && i = j \\ &= \Delta_{i+1}/6, && i = j - 1 \\ &= \Delta_i/6, && i = j + 1 \\ &= 0, && \text{otherwise}, \end{aligned} \tag{5}$$

where the tridiagonal $(N - 2) \times (N - 2)$ matrix H is of full rank.

Proposition 1. The limit as $t \to \infty$ of Q_t^{-1} is $R'H^{-1}R$.

Proof. Define the nonsingular matrix Φ as $[R' \vdots A]$. Since $RQ_tR' = H$ and $RQ_tA = RQ_0A$, only the lower right 2×2 block of $\Phi' Q_t \Phi$ depends on t. Using the rule for the inverse of a partitioned matrix and taking limits as $t \to \infty$, one finds that $\lim_{t\to\infty}(\Phi' Q_t \Phi)^{-1}$ is block diagonal with the upper block equal to H^{-1} and the lower block equal to zero. Premultiplying by Φ and postmultiplying by Φ' yields the proposition.

The prior distribution based on continuous smoothness priors of the parameters of the model, the $N + g$ element vector $\beta = [f' \vdots \gamma']'$ and the precision (h) of the regression error term, will be a partially uninformative conjugate prior of the kind discussed in Raiffa and Schlaifer (1961). This prior conveniently produces a multivariate Student posterior. Such a prior provides an estimator that can be implemented by running a regression with dummy observations representing the priors, and it allows Bayesian interpretations of the estimated coefficients and standard errors. The case of this prior used here for h (independent of β) is $p(h) \propto 1/h$. This is the uninformative prior for a scale parameter proposed by Jeffreys (1961). The prior of γ (independent of f and h) will be flat: $p(\gamma) \propto$ constant. Forming the product of these independent priors, one finds the prior density

$$p(\beta, h) \propto h^{(N-4)/2}\exp(-k^2 h \beta' \bar{R}' H^{-1} \bar{R}\beta/2), \tag{6}$$

where $\bar{R}$ is the $(N - 2) \times (N + g)$ matrix $[R \vdots 0]$ and $k = \sigma/\xi$.

The posterior distribution for β and h is by Bayes law, proportional to the product of the prior (6) and the likelihood (2). To write the posterior, we first define the matrices $\tilde{X}$ $(n + N - 2 \times n + g)$, $\tilde{Y}$ $(n + N - 2 \times 1)$, $\tilde{\epsilon}$ $(n + N - 2 \times 1)$, and Ω $(n + N - 2 \times n + N - 2)$:

$$\tilde{X} = \begin{bmatrix} X \\ \cdots \\ k\bar{R} \end{bmatrix}, \tilde{Y} = \begin{bmatrix} Y \\ \cdots \\ 0 \end{bmatrix}, \tilde{\epsilon} = \begin{bmatrix} \epsilon \\ \cdots \\ k\eta \end{bmatrix}, \Omega = \begin{bmatrix} I_n & \vdots & 0 \\ \cdots & \vdots & \cdots \\ 0 & \vdots & H \end{bmatrix} \tag{7}$$

so that $\tilde{Y} = \tilde{X}\beta + \tilde{\epsilon}$ and $\eta = Rf$. This setup is similar to that in Shiller (1973). The marginal posterior for β is then the multivariate Student distribution

$$p(\beta) \propto [n - g - 2 + \hat{h}(\beta - \hat{\beta})'(\tilde{X}'\Omega^{-1}\tilde{X}) \times (\beta - \hat{\beta})]^{-(n+N-2)/2}, \tag{8}$$

where $\hat{\beta} = (\hat{f}', \hat{\gamma}')'$ is the posterior mean of β, given by

$$\hat{\beta} = (\tilde{X}'\Omega^{-1}\tilde{X})^{-1}\tilde{X}'\Omega^{-1}\tilde{Y}$$
$$= (X'X + k^2\bar{R}'H^{-1}\bar{R})^{-1}X'Y. \tag{9}$$

If $g = 0$ and $X = I$, expression (9) is implicit in expressions 4.2–4.4 in Wahba (1978), where m in her expressions is set to 2. The posterior mean of h is given by

$$\hat{h}^{-1} \equiv \hat{\sigma}^2 = (\tilde{Y} - \tilde{X}\hat{\beta})'\Omega^{-1}(\tilde{Y} - \tilde{X}\hat{\beta})/(n - g - 2) \tag{10}$$

(see Raiffa and Schlaifer 1961 for theorems that establish this).

It is clear that $\hat{\beta}$ may be produced by running a conventional generalized least-squares regression of $\tilde{Y}$ on $\tilde{X}$ using variance matrix Ω. The final $N - 2$ rows of $\tilde{X}$ and $\tilde{Y}$ may be viewed as representing dummy observations that the changes in slope of the function (the elements of Rf) equal zero plus a noise term η. The larger k is the more weight will be given to these dummy observations and hence the more the regression estimator will smooth the estimated function. (There is a substantial literature regarding choice of nuisance parameters like k here. One might look at Craven and Wahba 1979 or Ullah and Raj 1979.) The estimate of the variance–covariance matrix of estimated coefficients that would be printed out by a standard generalized least squares regression program is $\hat{\sigma}^2(X'X + k^2\bar{R}'H^{-1}\bar{R})^{-1}$. The marginal distribution of the ith coefficient β_i is Student with $n - g - 2$ degrees of freedom (as would be computed by the program) and scale parameter given by the standard error of the coefficient printed by the program. Thus the standard t statistics have a Bayesian interpretation.

3. ESTIMATING POINTS OF THE FUNCTION WHERE THERE ARE NO OBSERVATIONS

Proposition 2. Adding to or deleting from the list $\bar{x}_{i1}$ $(i = 1, \ldots, N)$ some elements that do not correspond to observations has no effect on estimates or standard errors of parameters. That is, suppose that one increases or decreases the number of extra points at which the function is to be estimated. If the same procedure with the same value of k is used, one will get the same estimates and standard errors of the function (at points corresponding to the values of $\bar{x}$ that are included in both estimates) and of the coefficient vector γ.

Proof. It will suffice to show this for the case in which one such element of f is deleted. The partially improper prior $p(\beta)$ on β was $p(\beta) = \lim_{t\to\infty} p_t(\beta)$. If one uses the same procedure to construct a prior on the $(N + g - 1)$ element vector $\beta^{(i)}$ of elements of β other than f_i, then the prior

$$p^{(i)}(\beta^{(i)}) = \lim_{t\to\infty} p_t^{(i)}(\beta^{(i)}) = \lim_{t\to\infty} \int p_t(\beta) df_i.$$

Interchanging the limit and the integral shows that $p^{(i)}(\beta^{(i)}) = \int p(\beta) df_i$, that is, $p^{(i)}(\beta^{(i)})$ is the marginal prior distribution of $\beta^{(i)}$ from $p(\beta)$. By construction, if $\bar{x}_{i1}$ does not correspond to an observation, then the ith column of X_1 is zero and hence f_i drops out of the likelihood function. The marginal posterior distribution $\int p(\beta) df_i$ can therefore be found as the product of the same likelihood function (2) with $\int p(\beta) df_i = p^{(i)}(\beta^{(i)})$. The marginal posterior $\int p(\beta) df_i$ is thus the same as the posterior that the same procedure would give for $\beta^{(i)}$, and thus the estimates and standard errors as defined here are the same.

4. RELATIONS TO SPLINE INTERPOLATION

A cubic spline $v(x)$ is the third integral of a step function in which the values of x, $\bar{x}_{i1}$ $(i = 1, \ldots, N)$ at which the steps occur are called *knots*. It is a natural spline if the function is linear in x for $x < \bar{x}_{1,1}$ and for $x > \bar{x}_{N1}$. There is a unique cubic natural spline that interpolates any set of $N \geq 2$ points $(\bar{x}_{i1}, y_i)$, $i = 1, \ldots, N$, for which $\bar{x}_{i1} \neq \bar{x}_{j1}$ unless $i = j$ (Greville 1969).

It is convenient to write the general cubic natural spline with knots $\bar{x}_{i1}$ $(i = 1, N)$ in terms of N parameters C_i $(i = 1, \ldots, N)$ in a form used by Kimeldorf and Wahba (1970):

$$v(x) = \theta_0 + \theta_1 x + \sum_{i=1}^{N} q(x, \bar{x}_{i1}) C_i,$$

$$\text{where } \sum_{i=1}^{N} C_i = 0 \quad \text{and} \quad \sum_{i=1}^{N} C_i \bar{x}_{i1} = 0. \quad (11)$$

One may write the vector $\hat{f}$ of estimated function values in terms of the parameters of the cubic natural spline that interpolates them, $\hat{f} = A\theta + Q_0 C$, $\theta = [\theta_0, \theta_1]'$, and $C = [C_1, C_2, \ldots, C_N]'$. Then one has Proposition 3, which is a restatement in this context and a generalization of a result from Wahba (1978).

Proposition 3. With continuous smoothness priors, if the ith column of X_1 is spanned by X_2, then $C_i = 0$.

Proof. The two restrictions on C can be written $C = R'\tilde{C}$, where $\tilde{C}$ is an $N - 2$ element vector of parameters. Now note that (using $RA = 0$),

$$R'H^{-1}R\hat{f} = R'(RQ_0R')^{-1}Rf$$
$$= R'(RQ_0R')^{-1}R(A\theta + Q_0R'\tilde{C})$$
$$= R'(RQ_0R')^{-1}RQ_0R'\tilde{C} = R'\tilde{C} = C.$$

From the definition of $\hat{\beta}$,

$$(X'X + k^2\bar{R}'H^{-1}\bar{R})\hat{\beta} = X'Y,$$

and hence

$$X'X\hat{\beta} + k^2\bar{C} = X'Y, \quad (12)$$

where $\bar{C}' = [C', 0]$. Since the ith row of X_1 is spanned by X_2, one can write $Xr = 0$ where the first N elements of the vector r are zero except for the ith. Premultiplying Equation (12) by r' then gives $C_i = 0$.

Proposition 2 showed that estimates depend only on those x_i sites in which there are observations. The most important consequence of Proposition 3 is that estimates for the other points fall on the cubic natural spline that interpolates the estimates for the actual observations. Specifically, if one includes in the $\bar{x}$ a value of the function at a point $\bar{x}_{i1}$ where there is no observation, then the estimate at this point $\hat{f}(\bar{x}_{i1})$ is a cubic natural spline interpolation of the other estimated points $\hat{f}(\bar{x}_{11})$, $\hat{f}(\bar{x}_{12})$, $\ldots$, $\hat{f}(\bar{x}_{i-1,1})$, $\hat{f}(\bar{x}_{i+1,1})$, $\ldots$, $\hat{f}(\bar{x}_{N1})$. In this case the ith column of X_1 is zero and thus spanned by X_2, even if X_2 has zero columns $(g = 0)$.

Proposition 3 also implies that the estimate $\hat{f}(\bar{x}_{i1})$ is the cubic natural spline interpolation of the other function values if one dummies all observations corresponding to $\bar{x}_{i1}$ (i.e., if one includes among the columns of X_2 a dummy variable that is 1 in all rows where x_1 attains this value and zero otherwise) or if one includes in X_2 a column whose elements are collinear with x_1, except in a region where $x_1 = \bar{x}_{i1}$ (and there has a different value).

Although the estimated function is a cubic spline, the procedure giving rise to $\hat{\beta}$, as noted, should not be confused with the fitting of regression splines as it is commonly practiced (for a discussion see Wegman and Wright 1983). Fitting regression splines achieves restrictions on the function at the points x_{i1} $(i = 1, \ldots, n)$ by assuming the function lies on a spline in x_1 with *fewer than n* knots, and it makes no prior assumption about the probability of large changes in slope.

5. DISCRETE SMOOTHNESS PRIORS

What is the role of the H^{-1} matrix in the expression (Equation (9)) for $\hat{\beta}$? One could have begun with what might be called *discrete* smoothness priors by substituting the identity matrix for H in the prior (6). If the values of $\bar{x}_{i1}$ are chosen to be equally spaced (i.e., $\bar{x}_{i1} - \bar{x}_{i-1,1}$ is independent of i), then this prior would make f_i the ith partial sum of a random walk in i (rather than the integral to i of a Wiener process in i) and the estimate of β would be the ordinary least squares estimate of $\tilde{Y}$ on $\tilde{X}$:

$$\hat{\beta} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{Y} = (X'X + k^2\bar{R}'\bar{R})^{-1}X'Y.$$

In the case in which elements $\bar{x}_{i1}$ are equally spaced, $R'R$ has the property that $R'R\hat{f}$ (except for the first two elements and last two elements) is a constant times the vector of fourth differences of $\hat{f}$. Thus the same circumstances that with continuous smoothness priors caused $\hat{f}(\bar{x}_{i1})$ to lie on a cubic spline interpolating the other es-

timated function values would, with discrete smoothness priors, cause $\hat{f}(\bar{x}_{i1})$ to lie on a cubic polynomial that interpolates the adjacent two estimated values on either side—that is, interpolates $\hat{f}(\bar{x}_{i-2,1})$, $\hat{f}(\bar{x}_{i-1,1})$, $\hat{f}(\bar{x}_{i+1,1})$, $\hat{f}(\bar{x}_{i+2,1})$. In contrast, the spline interpolator places $\hat{f}(\bar{x}_{i,1})$ on a cubic polynomial that passes through $\hat{f}(\bar{x}_{i-1,1})$ and $\hat{f}(\bar{x}_{i+1,1})$ but does not generally pass through $\hat{f}(\bar{x}_{i-2,1})$ or $\hat{f}(\bar{x}_{i+2,1})$, and the polynomial is determined with the additional restriction of continuity of first and second derivatives. The estimate $\hat{f}(\bar{x}_{i1})$ with discrete smoothness priors is determined locally by adjacent estimated points, but the estimate $\hat{f}(\bar{x}_{i1})$ with continuous smoothness priors is determined globally by all other estimated points.

Proposition 2 does not hold for discrete smoothness priors (unless the element dropped is $\bar{x}_{11}$ or $\bar{x}_{N1}$). Discrete smoothness priors are thus most appropriately applied where $\bar{x}_{i1}$ ($i = 1, \ldots, N$) is the complete list of points at which the function might be estimated, as when x_{i1} take on integer values only.

6. RELATIONS TO LINEAR ORDINARY LEAST SQUARES

Whether one uses continuous or discrete smoothness priors—that is, whether Ω in $\hat{\beta} = (\bar{X}'\Omega\bar{X})^{-1}\bar{X}'\Omega^{-1}\bar{Y}$ is as shown in (7) or is the identity matrix—strict multicollinearity is a problem if and only if it is a problem with ordinary linear regression. Moreover, the estimates approach the ordinary linear regression estimates as k approaches infinity. It can be shown that if $(\bar{X}'\bar{X})$ is nonsingular,

$$\lim_{k\to\infty} \hat{\beta} = \bar{A}(\bar{A}'X'X\bar{A})^{-1}\bar{A}'X'Y$$

and

$$\lim_{k\to\infty} \sigma^2(\bar{X}'\Omega^{-1}\bar{X})^{-1} = s^2\bar{A}(\bar{A}'X'X\bar{A})^{-1}\bar{A}',$$

where $\bar{A}$ is the $(N + g) \times (g + 2)$ block diagonal matrix with A in the upper block and I_g in the lower block and where s is the estimated standard error of regression of Y on $X\bar{A}$. If $\bar{X}'\bar{X}$ or $\bar{X}'\Omega^{-1}\bar{X}$ is nonsingular for any k, then $\bar{X}'\bar{X}$ and $\bar{X}'\Omega^{-1}\bar{X}$ are nonsingular for all nonzero k. Moreover, $\bar{X}'\bar{X}$ is nonsingular if and only if $A'X'XA$ is nonsingular. The $n \times (g + 2)$ matrix $X\bar{A}$ has 1's in its first column, the x_1 observations in its second column, and the observations of the remaining variables in succeeding columns. Thus $\lim_{k\to\infty} \hat{\beta} = \bar{A}\delta$, where the $g + 2$ element vector δ is the vector of linear ordinary least squares regression coefficients of y on a constant, x_1 and x_2.

Smoothness priors are uninformative on the same space where the Bayesian priors that produce linear ordinary least squares are uninformative but are uninformative on a space of dimension smaller than that of the Bayesian priors that produce classical polynomial regression (i.e., a regression of y on x_1, x_2 and higher powers of x_1) as an estimator. Thus polynomial regression may fail because of multicollinearity even when linear ordinary least squares regression (and hence the estimate with smoothness priors) does not.

Either too small or too large a value of k may result in an $\bar{X}'\bar{X}$ matrix, which, although technically nonsingular, may be close enough to singularity that a computer program will not invert it. All ill-conditioned $\bar{X}'\bar{X}$ matrix may also occur if two observations of x_1 are very close together but not equal. One may wish to deal with this problem by rounding the data so that the two observations become identical.

7. ATTEMPTS TO REDUCE THE ESTIMATION PROBLEM

Does the estimator of the unknown function in this multiple regression context reduce in some sense to simple univariate curve fitting on transformed data? One might expect that one could at least get a good idea of the nature of the curve in a scatter plot. Call e_j the residuals of a vector j in a regression on X_2 (so that $e_j = (I - X_2(X_2'X_2)^{-1}X_2')j$). One might then consider plotting e_y versus x_1, e_y versus e_{x_1}, or just y against x_1. No such scatter plot, however, can properly take into account the covariances of all of the dummy variables representing points along the function with X_2. In experiments with a known function $f(x_1)$ one finds that such scatter plots may give no indication of the true curve even when the estimator (9) produces very nearly the true curve. Thus the multiple regression context is a good showplace for smoothness priors.

Even when the vector $x_1 = [x_{11}, x_{21}, \ldots, x_{n1}]'$ is orthogonal with all columns of X_2, the estimate of the function $\hat{f}$ is not the same as the estimate one would obtain by using the smoothness priors estimator (9) with e_y in place of y. Rather, the estimate is the same when the vector $[\hat{f}(x_{11}), \hat{f}(x_{21}), \ldots, \hat{f}(x_{n1})]$ is orthogonal with all columns of X_2, as can be seen directly from the normal equations implied by (9). This is not helpful, since one must know $\hat{f}$ to reduce the problem to univariate curve fitting in this way.

8. ILLUSTRATIVE EXAMPLE

In an oft-cited study, Nerlove (1963) sought evidence in the electric utility industry for the U-shaped cost curve for individual firms hypothesized by microeconomic theorists. Both very small firms and very large firms were thought to be inefficient and hence to face higher costs of producing output. Such cost curves have been the cornerstone of the classical theory of the size distribution of firms: Competition should tend to drive the size of firms to that corresponding to the minimum cost level of output.

The data used here are from Nerlove's (1963) Appendix C. Each of the 25 observations used here represents a firm and applies to the year 1955. The dependent variable y is the log of cost in millions of dollars per billion kilowatt-hours of output. The variable x_1 is the log of output measures in billions of kilowatt-hours. The vector x_2 consists of two variables: the log of the wage rate and the log of the fuel cost.

In estimating the model (1), Nerlove used a piecewise linear function for $f(\cdot)$ and concluded that there was no evidence for the U-shaped cost curve hypothesized by

theorists. He concluded that costs per unit of output were declining with output for small firms but essentially constant for large firms.

When the continuous smoothness priors with $k = 1$ were applied to these data, the estimated function was as shown in Figure 1. The estimated curve indeed shows declining costs at first, followed by essentially constant costs.

Consideration of this example may help to elicit one's priors. In such a study it would not seem appropriate to use second-order polynomial regression for x_1. To do that would be to restrict the function either to the U shape of a parabola or to a linear shape. Higher order polynomials are also incapable of representing an asymptote, but they would allow shapes much closer to that seen in Figure 1, as well as U shapes. A fifth-order polynomial-regression estimate (which retained linearity in X_2), Figure 2, looked more or less like the estimate in Figure 1. Although one should not make too much of the difference between Figures 1 and 2 (the standard errors of the coefficient estimates in both figures around $\bar{x} = 9$ are a little more than one), it is interesting to note that Figure 2 *does* suggest a minimum cost level of output—that is, the function suddenly curves up at the end. What might account for this difference between 1 and 2?

The first question to ask is how well a fifth-degree polynomial can approximate the curve in Figure 1. To answer this, $\hat{f}$ from Figure 1 was fitted to x_1 with a fifth-order polynomial. The fitted curve came much closer to the $\hat{f}$ in Figure 1 than the curve shown in Figure 2 did, but it also showed an upturn at the end, establishing that the inability of polynomials to represent asymptotes is a problem here. The upturn, however, was only one-fourth as big as in Figure 2. Furthermore, note that the curve in Figure 2 lies everywhere above that of Figure 1, so one should not regard the curve in Figure 2 as a polynomial approximation to the curve in Figure 1.

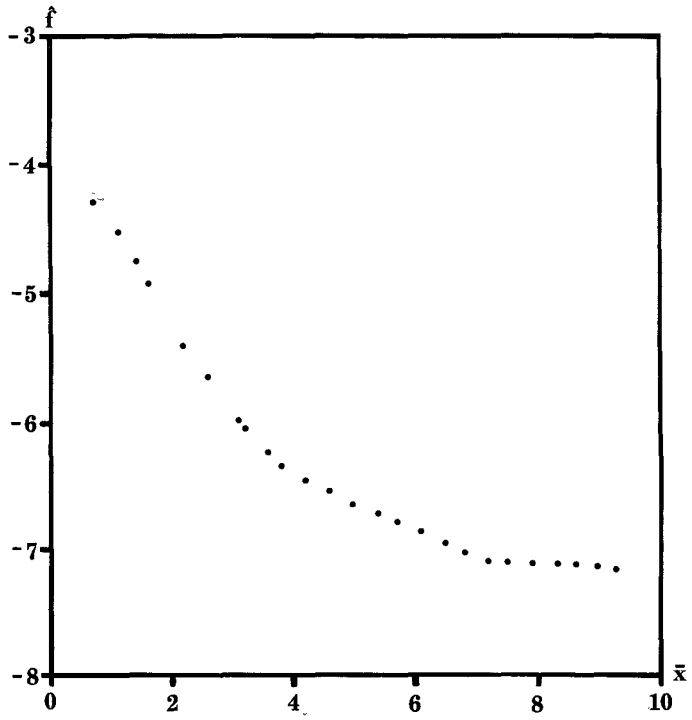

Figure 1. Estimated Function Based on Continuous Smoothness Priors ($\hat{f}(\bar{x}_{i1})$ v $\bar{x}_{i1}$, $i = 1, \ldots, N$) Using Estimator (9) With $k = 1$ Data are described in Section 8

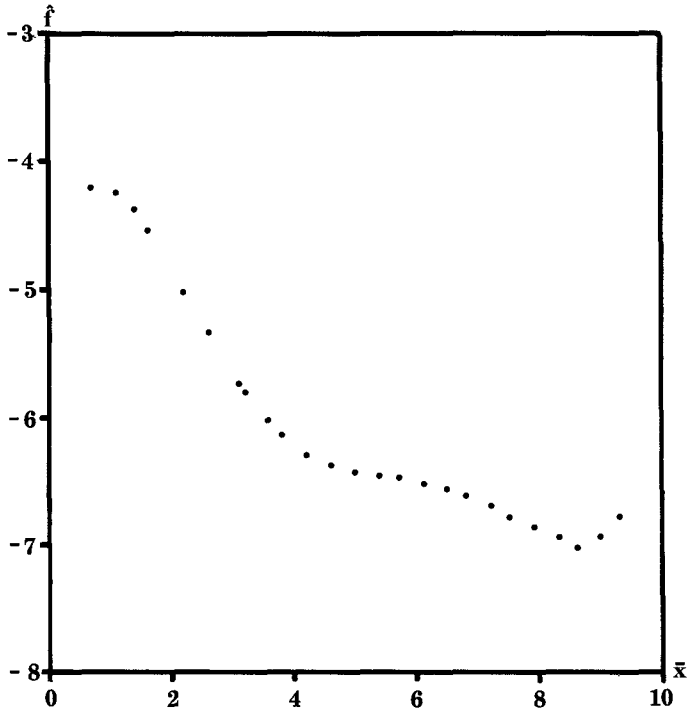

Figure 2. Estimated Function Using the Same Data That Produced the Estimates Shown in Figure 1 but Using Fifth-Degree Polynomial Regression Instead of the Estimator (9).

Polynomial regression has no objection, so to speak, to sudden changes in slope. Thus, for example, if x_1^5 is nearly collinear with variables in x_2, then its coefficient and hence the behavior of the estimated function at the end may be highly erratic. For Nerlove's purposes, rubber-ruler priors may be more appropriate than a prior in which the function is a polynomial.

9. DISCUSSION

The versatility of the estimators is apparent in experiments in which the likelihood is quite informative—that is, where σ is small. For example, the estimators do nearly as well as a polynomial regression in which the true curve is a polynomial, and they do much better if the true curve is bell-shaped. This, of course, is just a reflection of the aptness of the prior.

One observation that comes from experimenting with the estimators is that it often seems to make relatively little difference for the general appearance to the estimated function whether one uses discrete or continuous smoothness priors or even, over a substantial range, which k one uses. There is a sense in which the main effect of the smoothness priors is to entrain the various points along the estimated function, regardless of the exact variance matrix Ω chosen for the prior. If there are, say, 25 points spaced at unit intervals estimated along the

function, then the prior standard deviation of the change in slope of the function over its whole range is five times the standard deviation ξ of the change in slope between successive points. Thus over a wide range of chosen values for ξ, the prior is essentially uninformative about the overall shape of the function while being fairly informative that the curve should be smooth.

One might contrast the estimator with the generalization of polynomial regression of Blight and Ott (1975). They weakened the strict polynomial parameterization by representing the function $f(x_1)$ as a polynomial of known degree plus an error term $e(x_1)$, which was a first-order autoregressive process in x_1. Their prior allows rapid changes in slope if the standard of e is large, and rules out asymptotes if the standard deviation of e is small.

In the preceding illustrative example, observations were chosen to be more or less equally spaced in terms of x_1. It often happens that observations of x_1 are, instead, relatively clustered in certain regions. In that case, the estimators based on smoothness priors will tend to give a more detailed estimated curve in those regions. This ought to be considered an advantage of the estimators. For example, McCulloch (1975), who wished to estimate the yield curve (yield as a function of time to maturity) on U.S. Treasury securities, found that most observations were at the low end of the maturity scale. He observed that when polynomial regression was used to fit the curve, the detail apparent in the scatter diagram at low maturities was lost in the estimated polynomial. Therefore he used spline regression and chose more knot positions at the low end of the maturity scale. Smoothness priors automatically place knots at each observation point.

[*Received March 1983. Revised February 1984.*]

REFERENCES

BLIGHT, B.J.N., and OTT, L. (1975), "A Bayesian Approach to Polynomial Regression," *Biometrika*, 62, 79–88.

BREIMAN, L., and MEISEL, W.S. (1976), "General Estimates of the Intrinsic Variability of Data in Nonlinear Regression Models," *Journal of the American Statistical Association*, 71, 301–307.

COVER, T.M. (1968), "Estimation by the Nearest Neighbor Rule," *IEEE Transactions on Information Theory*, 14, 50–55.

CRAVEN, PETER, and WAHBA, GRACE (1979), "Smoothing Noisy Data With Spline Functions," *Numerische Mathematik*, 31, 370–403.

FRIEDMAN, JEROME H., and STUETZLE, WERNER (1981), "Projection Pursuit Regression," *Journal of the American Statistical Association*, 76, 817–823.

GERSOVITZ, MARK, and MACKINNON, JAMES G. (1978), "Seasonality in Regression: An Application of Smoothness Priors," *Journal of the American Statistical Association*, 73, 264–273.

GREVILLE, T.N.E. (1969), *Theory and Applications of Spline Functions*, New York: Academic Press.

JEFFREYS, H. (1961), *Theory of Probability*, London: Oxford University Press.

KIMELDORF, GEORGE S., and WAHBA, GRACE (1970), "A Correspondence Between Bayesian Estimation on Stochastic Processes and Smoothing by Splines," *Annals of Mathematical Statistics*, 41, 495–502.

LEAMER, EDWARD E. (1978), "Regression Selection Strategies and Revealed Priors," *Journal of the American Statistical Association*, 73, 580–587

McCULLOCH, J. HUSTON (1975), "The Tax Adjusted Yield Curve," *Journal of Finance*, 3, 811–830.

NERLOVE, MARC (1963), "Returns to Scale in Electricity Supply," in *Measurement in Economics; Studies in Mathematical Economics and Econometrics in Memory of Yehuda Grunfeld*, Stanford, Calif.: Stanford University Press, 167–198.

OMAN, SAMUEL D. (1982), "Shrinking Toward Subspaces," *Technometrics*, 24, 307–311.

RAIFFA, H., and SCHLAIFER, R. (1961), *Applied Statistical Decision Theory*, Cambridge, Mass.: Harvard University Press.

SHILLER, ROBERT J. (1973), "A Distributed Lag Estimator Derived From Smoothness Priors," *Econometrica*, 41, 775–788.

SOKOLNIKOFF, I.S. (1956), *Mathematical Theory of Elasticity*, New York: McGraw-Hill.

STONE, CHARLES J. (1977), "Consistent Nonparametric Regression," *Annals of Statistics*, 5, 595–645.

ULLAH, A., and RAJ, B. (1979), "A Distributed Lag Estimator Derived From Shiller's Smoothness Priors," *Economic Letters*, 2, 219–223.

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

WEGMAN, EDWARD J., and WRIGHT, IAN W. (1983), "Splines in Statistics," *Journal of the American Statistical Association*, 78, 351–366.

WHITTAKER, EDMUND, and ROBINSON, G. (1967), *The Calculus of Observations* (4th ed.), New York: Dover Publications.