NBER TECHNICAL PAPER SERIES

SMOOTHNESS PRIORS AND NONLINEAR REGRESSION

Robert J. Shiller

Technical Paper No. 25

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge MA 02138

August 1982

Any opinions expressed are those of the author and not those of the National Bureau of Economic Research.

NBER Technical Working Paper #25
August 1982

Smoothness Priors and Nonlinear Regression

Abstract

In applications, the linear multiple regression model is often modified to allow for nonlinearity in an independent variable. It is argued here that in practice it may often be desirable to specify a Bayesian prior that the unknown functional form is "simple" or "uncomplicated" rather than to parametize the nonlinearity. "Discrete smoothness priors" and "continuous smoothness priors" are defined and it is shown how posterior mean estimates can easily be derived using ordinary multiple linear regression modified with dummy variables and dummy observations. Relationships with spline and polynomial interpolation are pointed out. Illustrative examples of cost function estimation are provided.

Professor Robert J. Shiller
Cowles Foundation
Yale University
Box 2125 Yale Station
New Haven, CT 06520

Smoothness Priors and Nonlinear Regression

Robert J. Shiller*

I. Introduction

Most applications of the multiple regression model are to relatively unstructured problems in which there is no hard knowledge that the relation is linear or has some other given functional form. Linearity is thus assumed for lack of reason to move to a more complicated structure. Frequently, however, there is reason to believe that some aspect of the relation is distinctly nonlinear, and thus certain simple modifications of the linear regression model are widely used. To take a concrete example, which serves to remind us of the unstructured sort of problems typically encountered, suppose one wishes to use the regression model to assess the impact of a number of variables: (percent industrialized, industry dummies, and per capita income) on the percent of the labor force unionized by county in a state. One may think that per capita income might have a positive association with unionization for low levels of income, and no impact or even a negative impact for high levels of income. Within the framework of the standard linear regression model one may allow for a nonlinear effect of income by including income squared, say, as well as income and the other variables in a multiple regression with percent unionized as the dependent variable. Such widely used simple expedients (in this example a multivariate polynomial regression which allows for a second degree or parabolic function for income) may seem adequate given the ambiguity regarding the theoretical framework. There is always a cost, however, if our assumptions are arbitrarily restrictive or misrepresent our prior information, however intangible this prior information may be. In this example, one probably did not want to rule out a function relating unionization to income which, say, approaches a horizontal asymptote, which was bell-shaped, etc., but low-order polynomials do rule out such shapes. Of course, certain alternative parametrizations besides the polynomial parametrization may seem satisfactory for a given application.

One might prefer instead as a general method to begin with a Bayesian prior on the functional form which merely represents the notion that the function ought to be "smooth" or "uncomplicated." Such a general method might then take the place of choosing among parametrizations. Methods are available which are as simple conceptually as including a squared term in a regression, but these methods are not widely used. I will develop here, and illustrate with an example, the use of "smoothness priors" in practical regression modelling. The estimators described here are explicitly Bayesian and in this respect are quite different from other estimators used for unknown functions such as the nearest neighbor methods (Cover [1968], Stone [1977]), recursive partitioning method (Breiman and Meisel [1976]) or the projection pursuit method (Friedman and Stuetzle [1982]).

What we will call "continuous smoothness priors" have been suggested by Wahba [1978] (drawing on earlier work of Kimeldorf and Wahba [1970]). She described her work as offering a Bayesian justification of spline smoothing for the estimation of an unknown function f(•) in the simple nonlinear regression model $y_i = f(x_i) + \varepsilon_i$ $i = 1, \ldots, n$ where ε_i is iid $N(0,\sigma^2)$. The "spline smoothing" that she justifies does not reduce to "spline regression" as it is commonly practiced (as surveyed, for example, in Poirier [1976]). Spline regression is a parametrization that involves restricting the function to be a spline with a specified number of knots at specified points, the restrictions achieved by making the number of knots less than n. Wahba's prior produces a posterior mean estimate which is a spline with n knots, one at each observation point. The priors restrict the function at the observation points instead by urging them to describe a "smooth" or "uncomplicated" curve. The priors involve a more natural and simple notion of smoothness than is implicit in spline regression with its arbitrary designation of knot points.

Wahba's prior effectively assumes that f(t) is a realization of a stochastic process in "time" t which is the d^{th} integral of a Wiener process which was started at zero in the infinite past. This prior produces a posterior mean for f which is a d+2 order spline. The prior is uninformative on a point of the function by itself. Instead, the priors in effect assert that the d+1 order derivatives with respect to t of the function are white noise. In the case d=1, if the prior variance of the white noise is "small" the priors assert that large changes in slope are unlikely, i.e., that the function is "smooth" or "uncomplicated." The priors assert nothing else, and their natural simplicity is appealing for relatively unstructured problems that are often encountered.

What we will call "discrete smoothness priors" have been used for the estimation of the parameters β_k $k=0,\dots,\lambda$ in the distributed lag model $y_t = \sum_{k=0}^{\lambda} \beta_k x_{t-k} + \varepsilon_t$ $t=1,\dots, n$, where ε_t is iid $N(0,\sigma^2)$, Shiller [1973]. These priors asserted that the d+1 order differences of the coefficients ordered by k have a given "small" variance, and are independently normally distributed. Equivalently, the function β_t is a realization of a stochastic process in "time" t which is the d^{th} sum of a discrete random walk started at zero in the infinite past. In the usual case d=1 these priors also assert, if the prior variance is small, that large changes in slope are unlikely, i.e., that the distributed lag coefficients β_k k=0,..., λ lie on a "smooth" curve; but the priors are uninformative on individual coefficients. Such a prior also had earlier antecedents in the literature on the graduation problem (Whittaker and Robinson [1967]). More recently, it has also been used in a seasonal regression model (Gersovitz and MacKinnon [1978]).

In this paper, both discrete and continuous smoothness priors will be developed and an application made to the estimation of simple cost functions. The application should serve to illustrate the appeal of the estimators based on such priors. The estimators will be developed in a simple way and without reference to the reproducing kernel Hilbert space framework of Wahba. The spline connection for continuous smoothness priors pointed out by Wahba will be established in a simple way and an analogous connection to interpolating polynominals for discrete smoothness priors will be shown. The estimators will be presented in terms of "dummy observations" and "dummy variables" as a natural and simple modification of linear regression. Attention will be confined to the case d=1 since in this case the priors have a simple interpretation in terms of change in slope and since the estimator then approaches linear ordinary least squares as the prior variance goes to zero.

II The Regression Model and Preliminary Indication of the Estimator

The model to be considered here is nonlinear in the first variable:

(1) $y_i = f(x_{i1}) + x_{i2}\gamma + \varepsilon_i$

where y_i is the i^{th} observation of the dependent variable (a scalar), $f(\bullet)$ is the unknown function, x_{i1} is the i^{th} ovservation of the first independent variable (a scalar), x_{i2} is the i^{th} observation of the g element row vector of other independent variables, γ is a g element column vector of regression coefficients and ε_i is the error term. We assume that the vector ε whose i^{th} element is ε_i is independent of all observations of the independent variables and is spherically normal, i.e., the density of ε is:

(2) $f(\varepsilon) \propto h^{n/2}\exp(-h\varepsilon'\varepsilon/2)$

where $h \equiv$ precision $\equiv 1/\sigma^2$ where σ^2 is the variance of ε_i. We are given n observations of the vector $[x_{i1}, x_{i2}, y_i]$ $i = 1,\ldots, n$ ordered in terms of increasing x_{i1}.

The analysis here can be extended in obvious ways to the case where other variables are additively nonlinear. It would be more elegant to extend the model to allow for a general multivariate function $Y_i = f(x_{i1}, x_{i2}) + \varepsilon_i$, however for most actual applications this simpler model (1) is probably preferable, in that the model departs only minimally from the standard multiple linear regression model.

Because the simplicity of the estimators is a prime consideration here and because the notation below is rather complicated, for the sake of motivation let us look briefly at a numerical example of the estimator which will be derived below. The example will also yield a mixed regression interpretation of the estimator (Theil [1971]).

Let us suppose that x_2 is a single (scalar) variable, and that we have four observations on the vector $[x_{i1}\ x_{i2}\ y_i]$ which are, ordered in terms of increasing x_1, [4,1,5], [4,2,6], [5,8,7] and [7,8,6]. The number of observations in this example is very small since the purpose is illustrative only. Suppose we wish to evaluate the function at points $\overline{x}_1 = 3,4,5,6,7$. At two of these points we have one observation, at one point we have two observations and at two points we have no observations. Defining the vector $\beta = [f(3), f(4), f(5)\ f(6), f(7), \gamma]'$ we can estimate the posterior mean of β by first setting up the regression:

$$(3)\quad \tilde{Y} = \tilde{X}\beta + \tilde{u}$$

where: $\tilde{X} = \begin{bmatrix} X \\ k\bar{R} \end{bmatrix}, \tilde{Y} = \begin{bmatrix} Y \\ 0 \end{bmatrix}, \tilde{u} = \begin{bmatrix} \varepsilon \\ k\eta \end{bmatrix}$

and

$$X = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 2 \\ 0 & 0 & 1 & 0 & 0 & 8 \\ 0 & 0 & 0 & 0 & 1 & 8 \end{bmatrix} \qquad Y = \begin{bmatrix} 5 \\ 6 \\ 7 \\ 6 \end{bmatrix}$$

$$\bar{R} = \begin{bmatrix} 1 & -2 & 1 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 & 0 \\ 0 & 0 & 1 & -2 & 1 & 0 \end{bmatrix} = [R \vdots 0]$$

Here, the first four rows of $\tilde{X}$ and $\tilde{Y}$, represent the data on the independent variables and the dependent variable respectively. The first four observations on $\tilde{u}$ are the realizations of the regression error term ε. The next three rows of $\tilde{X}$ and $\tilde{Y}$ are the dummy observations, so that $\bar{R}\beta$ is the vector of second differences of the function and η is the error of the priors, i.e., the second differences of the actual function. The scalar k is the ratio of the standard deviation σ of the regression error ε to the prior standard deviation ξ. If k is large, i.e., the priors on the second differences are tight, then a lot of weight is given to the dummy observations which represent that the second differences of the function are probably small, i.e., that the function does not have sudden changes in slope.

The posterior mean estimate of β using discrete smoothness priors is just the ordinary least squares estimate $\hat{\beta} \equiv (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{Y} = (X'X + k^2\bar{R}'\bar{R})^{-1}X'Y$. It may be thought of as a sort of ridge estimator except that the parameters representing points along the curve are not shrunk towards zero but towards each other. With continuous smoothness priors the estimator is slightly different, reflecting the fact that second differences of the integral of a Wiener process are not serially uncorrelated but have MA(1) serial correlation with correlation coefficients at one lag of 0.25.[1] The variance matrix of the error term u is thus equal to $\sigma^2\Omega$ where Ω is block diagonal with the identity matrix in the upper block and the tridiagonal matrix H in the lower block:

$$(4)\quad H = \frac{2}{3}\begin{bmatrix} 1 & .25 & 0 \\ .25 & 1 & .25 \\ 0 & .25 & 1 \end{bmatrix}$$

The posterior mean of β is then the generalized least squares estimate:

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

One might argue that smoothness priors are a more natural extension to the regression model of the prior notions of smoothness which gave rise to the spline or polynomial regression models. For example, if one looks at the functional minimized by a cubic spline in Schoenberg [1964], one sees a resemblance to the posterior distribution for continuous smoothness priors developed below. The original cubic natural spline was developed as an approximation to the curve an elastic beam (or spline of wood) assumes if it is subjected to loads at certain points called knots. The curve assumes a shape which minimizes potential energy. The potential energy in the curve is approximately proportional to the integral of squared second derivatives (Sokolnikoff [1956]). Here, we may envision the likelihood function as pushing the curve at discrete points representing observations and the prior as representing the elasticity of the beam. It is thus natural to have a knot at each point where there is an observation.

The principal reason for considering the slightly more complicated continuous smoothness priors is that such priors produce a sort of time consistency for the estimate. We shall see that if we used the same data and procedure to estimate the function at points 3, 3.5, 4, 4.5, 5, 5.5, 6, 6.5, and 7 then, for an appropriately scaled and expanded $\bar{R}$ matrix of the same form, the estimates and standard errors at the points 3, 4, 5, 6, and 7 would be the same as in the above regression. We could expand the set of points more and estimate the curve at intervals of 0.1 from 2 to 10, say, and then we would find the estimates unchanged at the points used before. Thus, in effect, the procedure provides an estimate of a continuous function. In contrast, discrete smoothness priors are most appropriately applied when such expansion is not envisioned, as when x_1 takes on integer values only.

III <u>The Prior and Posterior Distribution</u>

We wish to estimate $f(\cdot)$ in (1) at $N \geqslant 3$ 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 $i=j$. N may be either greater than, equal to, or less than n. N could be greater than n if, as in the example in the previous section, we wish to estimate the function at points where there are no observations. Below, we shall say that such a point $\bar{x}_{i1}$ (where $\bar{x}_{i1} \neq x_{j1}$ for any j) does not correspond to an observation. N could be less than n if there are repeated observations of x_1, i.e., $x_{i1} = x_{j1}$ for some $i \neq j$. Such repeated observations were also represented in the example in the preceding section. Write the $n \times (N+g)$ element matrix $X = [X_1 \vdots X_2]$.

The n × N element matrix X_1 has elements X_{1ij} equal to zero except where $x_{i1} = \overline{x}_{j1}$ where the element is 1. The n × g element matrix X_2 has x_{i2} as its i^{th} row. With the n element vector, Y--whose i^{th} element is y_i--and the N+g element vector $\beta' = [f' \vdots \gamma']'$ where the i^{th} element of the N element column vector f is $f(\overline{x}_i)$, we can write the likelihood function:

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

We wish to express improper prior distributions for the vector f which are uninformative on the function at any point or on its slope between any two points, but which represent the notion that the function is probably fairly smooth. These will by "cylindrically uniform" priors as in Leamer [1978]. To derive continuous smoothness priors, 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:

(6) $$q(s,t) = \begin{cases} s^2t/2 - s^3/6 & s \leqslant t \\ st^2/2 - t^3/6 & s \geqslant t \end{cases}$$

Let Q_t be the N × N matrix whose ij^{th} element is $q((\overline{x}_{i1}+t), (\overline{x}_{j1}+t))$. Clearly, as t approaches infinity all elements of Q_t approach infinity. However, $V_t \equiv Q_t^{-1}$ approaches a finite nonzero limit as t approaches infinity. We will make our prior precision matrix (inverse of prior variance matrix) for f equal to $\xi^{-2}\lim_{t\to\infty} V_t$, where ξ is the standard deviation of the change in slope (derivative) of the function over a one unit interval in x_1.

We define the (N-2) × N matrix R by:

$$(7)\quad R_{ij} = \begin{cases} \Delta_i^{-1} & i=j \\ -(\Delta_i^{-1} + \Delta_{i+1}^{-1}) & i = j-1 \\ \Delta_{i+1}^{-1} & i = j-2 \\ 0 & \text{otherwise} \end{cases}$$

where $\Delta_i \equiv (\overline{x}_{i+1,1} - \overline{x}_{i1})$. 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 $(\overline{x}_{i1}, f_i)$, $(\overline{x}_{i+1,1}, f_{i+1})$ and a line connecting $(\overline{x}_{i+1,1}, f_{i+1})$, $(\overline{x}_{i+2,1}, f_{i+2})$. In the case of the chosen values of $\overline{x}_1$ spaced at unit intervals, and N = 5, this R matrix reduces to the R matrix shown in the example in the introduction above. Thus in general R is a matrix of second differences for (possibly) unequally spaced observations. We define the N × 2 matrix A by $A_{i1} = 1$ and $A_{i2} = \overline{x}_{i1}$, i=1,..., N. Then 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, we find that:

$$(8)\quad H_{ij} = \begin{cases} (\Delta_i + \Delta_{i+1})/3 & i=j \\ \Delta_{i+1}/6 & i=j-1 \\ \Delta_i/6 & i=j+1 \\ 0 & \text{otherwise} \end{cases}$$

where the tridiagonal (N-2) × (N-2) matrix H is of full rank. Again, if the points estimated $\overline{x}_{i1}$ i=1,...N are spaced at unit intervals and N=5, then the matrix H is as shown in the introduction above.

Proposition 1:

The limit as $t\to\infty$ of Q_t^{-1} is $R'H^{-1}R$. Proof: Define the matrix Φ as $[R' \vdots A]$. Φ is non-singular since we have chosen the $\bar{x}_{i1}$ i=1,...,N so that no two values are the same, i.e., so that Δ_i is nonzero for i=1,...,N-1. The columns of R' are independent and so R' has rank N-2, and then also A has rank 2. Since $\Phi'\Phi$ is block diagonal with nonsingular blocks RR' and A'A, $\Phi'\Phi$ is nonsingular and hence Φ is nonsingular. 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[2] for the inverse of a partitioned matrix and taking limits as $t\to\infty$ we find that $\lim_{t\to\infty}(\Phi'Q_t\Phi)^{-1}$ is block diagonal with upper block equal to H and lower block equal to zero. Premultiplying by Φ and postmultiplying by Φ' yields the proposition.

We now specify the prior distribution based on continuous smoothness priors of all parameters of the model: the N+g element vector $\beta = [f' \vdots \gamma']'$ and the precision (h) of the regression error term. For simplicity, we adopt the partially uniformative conjugate prior of the kind discussed in Raiffa and Schlaifer [1961], which conveniently produces a multivatiate student posterior. Such a prior provides a justification for running a regression with dummy observations representing the priors and allows Bayesian interpretations of the estimated coefficients and standard errors. We thus choose the case of this prior which makes the prior of h (independent of β) $f_p(h) \propto 1/h$, i.e., the uninformative prior for a nonnegative variable proposed by Jeffreys [1961]. The prior of γ (independent of f and h) will be flat: $f_p(\gamma) \propto$ constant. The prior on the (N-2) element vector Rf will be multivariate normal with zero mean and precision k^2hH^{-1}. Finally, using flat priors on any two coefficients on f, forming the product of these independent priors, we find the prior density

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

where $\overline{R}$ is the $(N-2) \times (N+g)$ matrix $[R\vdots 0]$.

The posterior distribution for β and h is by Bayes law proportional to the product of the prior (9) and the likelihood (5). The marginal posterior for β is then the multivariate student distribution:

(10) $(\beta) \propto [n-g-2 + \grave{h}(\beta-\grave{\beta})'(X'X+k^2\overline{R}'H^{-1}\overline{R})(\beta-\grave{\beta})]^{-(n+N-2)/2}$

where $\grave{\beta} = (\grave{f}',\grave{\gamma}')'$, and where $\grave{h}^{-1} \equiv \grave{\sigma}^2 = (\tilde{Y}-\tilde{X}\grave{\beta})'\Omega^{-1}(\tilde{Y}-\tilde{X}\grave{\beta})/(n-g-2)$ where $\tilde{Y}$ and $\tilde{X}$ are as defined in (3). The posterior mean is $\grave{\beta} = (X'X + k^2\overline{R}'H^{-1}\overline{R})^{-1}X'Y$ which can then be found by regressing $\tilde{Y}$ on $\tilde{X}$ using generalized least squares with block diagonal variance matrix Ω with I in the upper block and H in the lower block.[3] The estimate of the variance-covariance matrix of estimated coefficients that would then be printed out by a standard generalized least squares regression program is $\grave{\sigma}^2(X'X + k^2\overline{R}'H^{-1}\overline{R})^{-1}$. The marginal distribution of the i^{th} 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.[4] Thus, the standard t statistics have a Bayesian interpretation.

It is now possible to show that any point of the function $\grave{f}_i = \grave{f}(\overline{x}_{i1})$ where there is no observation (so that the i^{th} row of X'X and X'Y = 0) lies on a cubic natural spline which interpolates the other coefficients. A cubic spline v(x) is the third integral of a step function, where the values of x, $\overline{x}_{i1}$ i=1,...,N at which the steps occur are called "knots." It is a natural spline if the function is linear in x for $x < \overline{x}_{1,1}$ and for $x > \overline{x}_{N1}$. There is a unique cubic natural spline which interpolate any set of N 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 [1971]:

$$(11)\ v(x) = \theta_0 + \theta_1 x + \sum_{i=1}^{N} q(x,\bar{x}_{i1})C_i \text{ where } \sum_{i=1}^{N} C_i = 0 \text{ and } \sum_{i=1}^{N} C_i\bar{x}_{i1} = 0.$$

That such a function is a cubic natural spline with knots at $\bar{x}_{i1},\ldots,\bar{x}_{N1}$, is easily seen. $q(x,\bar{x}_{i1})$ is a linear function of x for $x > \bar{x}_{i1}$ and a cubic function of x for $x < \bar{x}_{i1}$. Moreover, the zeroth, first and second derivatives are continuous at $\bar{x}_{i1}$. Clearly, any linear combination of $q(x,\bar{x}_{i1})$ $i=1,\ldots,N$ is a cubic spline. Moreover, the function $v(x)$ is linear for $x > \bar{x}_{N1}$, and the two restrictions on the C_i, $i=1,\ldots,N$ assure that it is linear for $x < \bar{x}_{1,1}$ and so the spline is a natural spline. We may write the vector f of estimated function values in terms of the parameters of the cubic natural spline that interpolates them, $f = A\theta + Q_0C$, $\theta = [\theta_0,\theta_1]'$ and $C = [C_1,C_2,\ldots,C_N]'$. We then have:

<u>Proposition 2</u>: With continuous smoothness priors, if, for a given i, $\bar{x}_{i1}$ does not correspond to an observation 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 $RA=0$), $R'H^{-1}R\grave{f} = R'(RQ_0R')^{-1}R\grave{f} = R'(RQ_0R')^{-1}R(A\theta+Q_0R'\tilde{C}) = R'(RQ_0R')^{-1}RQ_0R'\tilde{C} = R'\tilde{C} = C$. From the definition of $\grave{\beta}$, $(X'X + k^2\bar{R}'H^{-1}\bar{R})\dot{\beta} = X'Y$ and hence $X'X\grave{\beta} + k^2\bar{C} = X'Y$, where $\bar{C}' = [C' \vdots 0]$. Because $\bar{x}_{i1}$ does not correspond to an observation, the i^{th} rows of both $X'X$

and X'Y are zero and then $C_i = 0$. Q.E.D. It follows that the knot at $\bar{x}_{i1}$ is superfluous, i.e., $\hat{f}_i$ lies on a natural spline which interpolates the other estimated points along the function.

We define discrete smoothness priors simply by replacing the matrix H in the prior density (9) with the identity matrix. Clearly, the posterior mean $\hat{\beta}$ is the ordinary least squares estimate $\beta = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{Y} = (X'X + k^2\bar{R}'\bar{R})^{-1}X'Y$. We then have a proposition analogous to proposition 2 above:

Proposition 3: With discrete smoothness priors, and equally spaced estimated points $\bar{x}_{i1}$ $i=1,\dots,N$, if for a given i, $2 < i < N-1$, $\bar{x}_{i1}$ does not correspond to an observation then $\hat{f}_i = f(\bar{x}_{i1})$ lies along a cubic polynomial in $\bar{x}_{i1}$ which interpolates the adjacent two points on either side.

Proof: Because the estimated points are equally spaced, $\Delta_i = \Delta_j \; \forall \; i,j$ and hence R has rows which produce second differences (i.e. $R_i\hat{f} = \Delta_i(\hat{f}_i - 2\hat{f}_{i+1} + \hat{f}_{i+2})$), where R_i is the i^{th} row of R. Therefore the i^{th} row of R'R, $2 < i < N-1$, produces second differences of second differences or fourth differences, i.e. the i^{th} row or $R'R\hat{\beta}$ is $\Delta_i^2(\hat{f}_{i-2} - 4\hat{f}_{i-1} + 6\hat{f}_i - 4\hat{f}_{i+1} + \hat{f}_{i+2})$, which is proportional to the fourth difference in the estimated function. As noted above, the fact that $\bar{x}_{i1}$ does not correspond to an observation x_{j1} implies that the i^{th} rows of both X'X and X'Y are zero. Hence, from the i^{th} row of the normal equations $(X'X + k^2\bar{R}'\bar{R})\hat{\beta} = X'Y$ it follows that the fourth difference of the estimated function at i is zero, which implies that if the five points $\hat{f}_{i-2}$, $\hat{f}_{i-1}$, $\hat{f}_i$, $\hat{f}_{i+1}$, $\hat{f}_{i+2}$ lie on a cubic polynomial. Q.E.D.

We might elaborate on the differences between these propositions. When we use continuous smoothness priors and $k^2h\bar{R}'H^{-1}\bar{R}$ as the prior precision matrix for β and $\bar{x}_{i1}$ does not correspond to an observation, then $\hat{f}_i$ cannot be determined by the adjacent estimated coefficients alone as in the case with discrete smoothness priors. The cubic natural spline interpolator at any point depends on all points interpolated. For discrete smoothness priors, the first element of $R'R\hat{f}$ is $\hat{f}_1 - 2\hat{f}_2 + \hat{f}_3$ (the last element is $\hat{f}_{N-2} - 2\hat{f}_{N-1} + \hat{f}_N$), which means that if $\bar{x}_{1,1}$ does not correspond to an observation, then f_1 with discrete smoothness priors will lie on a straight line through $\hat{f}_2$ and $\hat{f}_3$. The second element of $R'R\hat{f}$ is $-2\hat{f}_1 + 5\hat{f}_2 - 4\hat{f}_3 + \hat{f}_4$ (the last but one element is $\hat{f}_{N-3} - 4\hat{f}_{N-2} + 5\hat{f}_{N-1} - 2\hat{f}_N$) which means that if the likelihood function carries no information about $f(\bar{x}_{2,1})$ then f_2 will lie on a cubic polynomial through $\hat{f}_1$, $\hat{f}_3$ and $\hat{f}_4$, which makes "f_0," $\hat{f}_1$, and $\hat{f}_2$ collinear. Thus, the estimated function lies along a unique interpolating function for the points where there are observations. Like the cubic natural spline, segments of the function are cubic polynomials and the function is linear beyond the range of the observations (where the likelihood function offers no "reason" for curvature). However, for the interior points, continuity of first and second derivatives is replaced by the requirement that the polynomial interpolate one point further on either side.

Whether we use continuous or discrete smoothness priors, i.e., whether Ω in $\beta = (\tilde{X}'\Omega\tilde{X})^{-1}\tilde{X}'\Omega^{-1}\tilde{Y}$ is the identity matrix, it can be shown that if $(\tilde{X}'\tilde{X})$ is nonsingular, $\lim_{k\to\infty} \hat{\beta} = \bar{A}(\bar{A}'X'X\bar{A})^{-1}\bar{A}'X'Y$ and $\lim_{k\to\infty} \hat{\sigma}^2(\tilde{X}'\Omega^{-1}\tilde{X})^{-1} = s^2\bar{A}(\bar{A}'X'X\bar{A})^{-1}\bar{A}'$ 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^2 is the estimated standard error of regression of Y on $X\bar{A}$. If $\tilde{X}'\tilde{X}$ or $\tilde{X}'\Omega^{-1}\tilde{X}$ is nonsingular for any k then $\tilde{X}'\tilde{X}$ and $\tilde{X}'\Omega^{-1}\tilde{X}$ are nonsingular for all nonzero k. Moreover, $\tilde{X}'\tilde{X}$ is nonsingular if and only if $A'X'XA$ is nonsingular. The n x (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, $\lim_{k\to\infty} \hat{\beta} = A\hat{\delta}$ 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.

Proposition 4: With continuous smoothness priors, if we drop some points from the list of points $\bar{x}_{i1}$ i=1,...,N where the function is to be evaluated and then use the same procedure with the same k to estimate, then, as long as these points dropped do not correspond to observations, the estimates of the remaining function values and coefficients (and their estimated standard errors) will be unaffected. Proof: We will show this for the case g = 0 (only one independent variable in the regression). The more general result is then easily established, though with rather messy partitioning of matrices. Call the subset of points (in increasing order) to be included $\bar{x}_{si}$ i=1,...,N_s $3 \leq N_s < N$. Call the N_s element vector of points to be estimated f_s. We can choose an N × N element matrix $J = (J')^{-1}$ such that $\beta'J = [f_s' \vdots f_e']$ where the $N-N_s$ element vector f_e' contains the values of the function that we will drop from estimation. We define X_s, R_s $(=\bar{R}_s)$, H_s, and Q_{st} as before but with $\bar{x}_{si}$ i=1,..., N_s in place of $\bar{x}_{i1}$ i=1,..., N. It follows that $J'X'XJ$ has $X_s'X_s$ in its upper diagonal N_s x N_s block and is zero elsewhere while $J'X'Y$ has $X_s'Y$ as its upper N_s element partition and zero in its lower $(N-N_s)$ element partition. To establish the proposition we need only show that $J'(X'X + k^2\bar{R}'H^{-1}\bar{R})^{-1}J$ has $(X_s'X_s$ +

$k^2\bar{R}_s'H_s^{-1}\bar{R}_s)^{-1}$ in its upper diagonal $N_s \times N_s$ block. To show this, partition $U \equiv k^2J'\bar{R}'H^{-1}\bar{R}J$ into $U = \begin{bmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{bmatrix}$ where U_{11} is $N_s \times N_s$ and partition $V_t \equiv k^2(J'Q_tJ)^{-1}$ conformably. The matrix U_{22}, whose order is at most N-3, is nonsingular. By the rule for inverting partitioned matrices, and because of the special structure of J'X'XJ, the upper N_s X N_s block of $J'(X'X + k^2\bar{R}'H^{-1}\bar{R})^{-1}J$ is $(X_s'X_s + U_{11} - U_{12}U_{22}^{-1}U_{21})^{-1}$. By proposition 1, $\lim_{t\to\infty} V_t = U$. Since V_{22_t} is nonsingular, the rule for inverting partition matrices implies that $k^2Q_{st}^{-1} = V_{11_t} - V_{12_t}V_{22_t}^{-1}V_{21_t}$. Thus, $k^2 \lim_{t\to\infty} Q_{st}^{-1} = U_{11} - U_{12}U_{22}^{-1}U_{21}$. But by proposition 1 with Q_{st}, R_s, and H_s in place of Q_t, R, and H, $\lim_{t\to\infty} Q_{st}^{-1} = R_s'H_s^{-1}R_s$. Substituting, we find that the upper N_s x N_s block of $J'(X'X + k^2\bar{R}'H^{-1}\bar{R})^{-1}J$ has the form indicated.

III Illustrative Examples and Discussion

One observation that comes from experimenting with the estimators is that it often makes very little difference for the general appearance of the estimated function what k one chooses (over a fairly wide range) or whether one uses discrete or continuous smoothness priors. The main effect of the smoothness priors is to entrain the various points along the estimated function, regardless of the exact variance matrix chosen for the prior. If there are, say, 25 points spaced at unit intervals estimated along the function then the prior standard deviation ξ of the change in slope of the function over its whole range is 5 times the standard deviation 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 curves should be smooth. By analogy, it may not matter much exactly what the stiffness is of the flexible curve a draftsman uses to fit a curve through a scatter of points. A draftsman may own only one such flexible curve!

The illustrative examples presented here use data published in Nerlove's well-known study of cost curves in the electric utility industry [1963]. 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 output. The variable x_1 is the log of output measured 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. Further description of the data, as well as a discussion of potential problems in estimation of cost functions from such data may be found in Nerlove [1963].

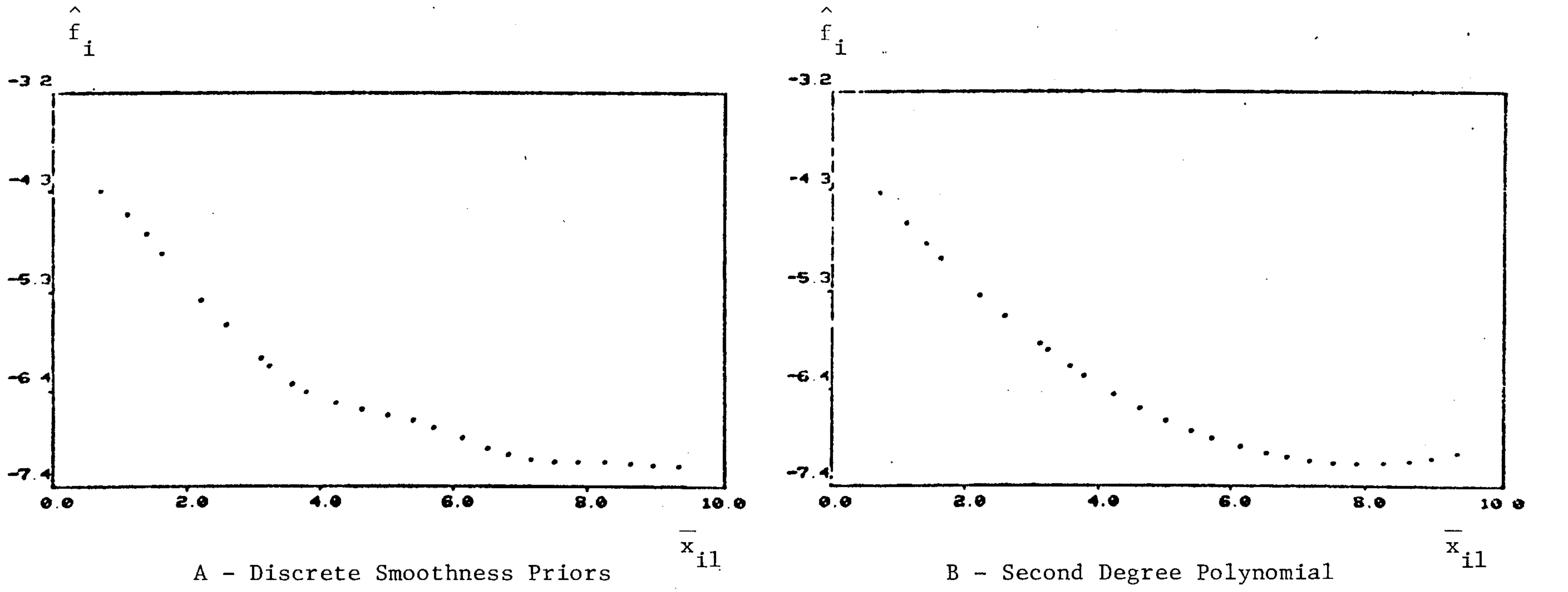
Nerlove used a piecewise linear function for $f(x_1)$, and concluded that there was no evidence for the U-shaped cost curve hypothesized by theorists. He concluded that costs per unit of output were declining with output for small firms and were essentially constant for large firms. When the discrete smoothness priors with $k = 1.5$ were applied to these data, the estimated function is as shown in figure 1-A.[5] The estimated curve indeed shows declining costs at first, then a region of essentially constant costs.

Figure 1-B shows the estimated function using second order polynomial regression. In this case , polynomial regression does fairly well in capturing the shape seen in figure 1-A. If one is particular, however, one notes that the estimated curve does have a U shape and suggests a cost minimizing level of output. Obviously, the U shape is due to the fact that parabolas are U-shaped! Of course, the estimation procedures might have given a parabola which does not turn up over the relevant range, but such a parabola apparently could not fit the other points of the function well. The appearance of the estimated function could be improved in this instance by moving to a cubic polynomial. In the cubic and quartic cases the estimated curve is not quite flat either in the high output range, but "oscillates" in the region. This is of course due to the fact that such polynomials cannot represent asymptotes. Also, the extra parameters may be "used up" to some extent in improving the fit of the estimated function in other regions. A fifth degree polynomial estimate showed a sharp upturn in cost in the highest region of x_1 reflecting the indifference of high degree polynomials to sudden changes in slope.

In this example the observations were chosen so as to be more or less equally spaced in terms of x_1. It often happens that observations of x_1 are 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. He therefore used spline regression and chose more knot positions at the low end of the maturity scale. Smoothness priors, it has already been noted, automatically place knots at each observation point.

The advantages of the smoothness priors might be made more obvious if we show an example where the true curve is known and where the average cost curve has a more interesting shape. For this purpose, artificial data y' were created equal to a function of x_1 (shown in figure 2-A), plus the vector x_2 times a vector of constants, plus an error term. The cost function created for this purpose has a region of constant costs for small firms, a region of declining then rising costs, then a lower region of constant costs for large firms. The normal error term was given a standard deviation of .05, much smaller than the estimated standard error of regression of .52 in the regression whose coefficients are shown in figure 1-A. The smaller error was chosen so that the data would contain a lot of information about the true curve.

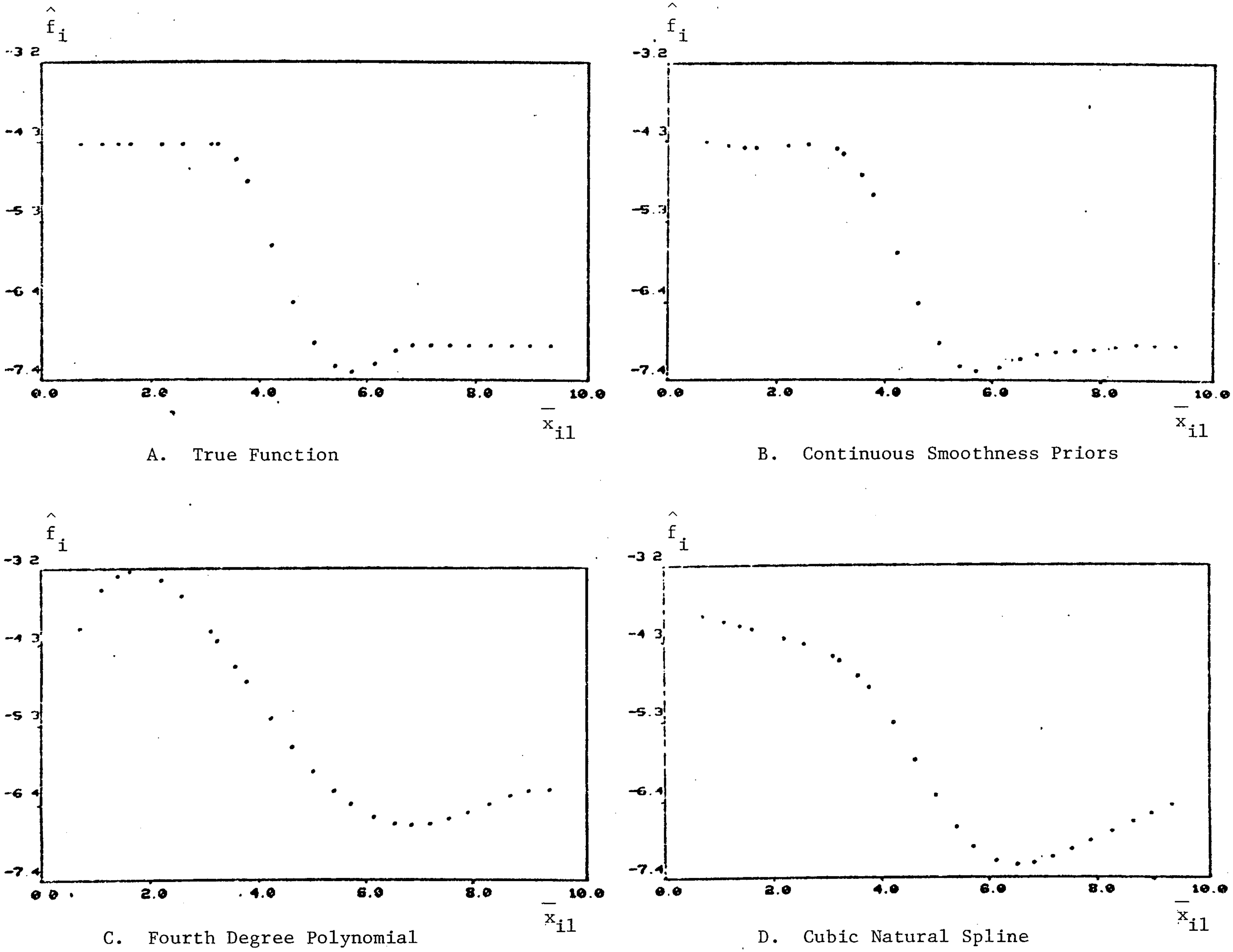
An estimate of the function using continuous smoothness priors and k=.1 is shown in figure 2-B. The actual shape is captured quite well. The flat region at the beginning and at the end are captured fairly well, and the point of minimized cost is right on the mark. In constrast, a fourth degree polynomial regression (figure 2-C) gives no indication of either flat region, displaces the entire function upward, and substantially misses the point of minimized cost. A fifth degree polynomial looked only a little better, and a sixth degree could not be run due to an ill-conditioned X'X matrix. In general, very high degree polynomials will allow "unsmooth" erratic behavior of the estimated function. Figure 2-D shows the estimted function from a cubic natural spline regression with 4 equally spaced knots, at the 5th, 10th, 15th, and 20th observations. The appearance of the curve



A - Discrete Smoothness Priors

B - Second Degree Polynomial

Figure 1 - Estimates of Average Cost Functions with A. discrete smoothness priors k = 1.5 and B. second order polynomial regression



A. True Function

B. Continuous Smoothness Priors

C. Fourth Degree Polynomial

D. Cubic Natural Spline

Figure 2 - Estimates of known function with Artificial Data: A. Actual function, B. Estimate with continuous smoothness priors k = .1, C. Estimate from fourth degree polynomial regression, D. Estimate from spline regression with knots as 5th, 10th, 15th, and 20th observation

might be improved by a better choice of knot locations, but in general one cannot expect to place knots in the right place.

How does one arrive at a value of k? One should not use a default option k=1 since k is not unit free. According to the assumptions of the Bayesian estimation k $(=\sigma/\xi)$ is known in advance. One must judge how large σ (the standard error of the regression error term) is relative to ξ (with continuous smoothness priors, the standard deviation of the change in slope over a unit interval). In the example with continuous smoothness priors shown in figure 2, σ was known to equal .05. The assumed k of .1 therefore implies that ξ was .5. This means that a one prior standard deviation change in the slope (derivative) of the function (or one might say, in the "coefficient" of x_1) when x_1, the log of output, increases one unit, was .5. The derivative of the true curve in figure 1a ranges from -1.92 to +.49, or just under 5 prior standard deviations over or range where x_1 varies over a range of 9 units. These priors are therefore a little on the tight side to capture this curve, as evidenced by the rounding off in the estimate of the two "corners" of the true curve at the eight and eighteenth observations where slope changes rapidly. With discrete smoothness priors, the parameter ξ is the prior standard deviation of the change in slope (or "coefficient" of x_1) between successive observations of x_1. In the example shown in figure 1, k=1.5 and if $\sigma=1/2$ then $\xi=1/3$. Since the average space between observations is .36, we might say that the prior standard deviation of the change in slope over a unit interval is $(1/3)/\sqrt{.36} = 5/9$, or just a little more than with prior in figure 2. The estimated curve shows slope which ranges from $-.63$ to 0, substantially less variation than the prior would allow.

It may be helpful, if for no more reason than to check that one has the units straight, to compare one's k with a rule of thumb. For continuous smoothness priors a useful rule of thumb is $\hat{k} = \sqrt{r}s/|4b|$.[6] Here, s is the estimated standard error of the regression based on a preliminary linear ordinary least squares regression of y on a constant, x_1 and x_2, and b is the coefficient of x_1 in the same regression. The symbol r denotes the sample range of x_1, i.e. its maximum value minus its minimum value. The rule of thumb $\hat{k}$ is equal to an estimate s of σ divided by an "estimate" $\hat{\xi}$ $(=|4b|/\sqrt{r})$ of ξ. The "estimate" $\hat{\xi}$ is suggested by noting that, since the slope (derivative) of the function is a Wiener process or random walk, the standard deviation of the change in slope over the range r is $\sqrt{r}\,\xi$. If we want to be fairly uninformative about the shape of the function, we might want to allow this standard deviation to be $|4b|$, hence the "estimate" of ξ. One may wish to decrease the rule of thumb value $\hat{k}$ if one thinks that the function has such a shape that b will be small even though the function varies a lot. One might in that case use a preliminary polynomial regression and use for b the "representative" slope of the curve. For discrete smoothness priors with roughly equally spaced observations one will want to consider the above $\hat{k}$ divided by the square root of the usual space $\overline{\Delta}$ between two adjacent observations of x_1, i.e. $\hat{k}_d = \sqrt{r}s/|4b\sqrt{\overline{\Delta}}|$. In this example of figure 1 this rule of thumb gave k=2.4.

Too small or too large a value of k may result in an $\tilde{X}'\tilde{X}$ matrix which, while technically nonsingular, may be close enough to singularity that a computer program will not invert it. The rule of thumb does not guarantee that the $\tilde{X}'\tilde{X}$ matrix won't be ill-conditioned. Indeed, the k that would minimize the condition number of $\tilde{X}'\tilde{X}$ in a particular application does not depend on the standard error of the regression, while the rule of thumb does. An ill-conditioned $\tilde{X}'\tilde{X}$ 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.

In some applications, one must clearly correct for serial correlation of error terms. One might assume a first order autoregressive structure for the error term and put an uninformative prior on the autoregressive coefficient ρ. The posterior distribution can be described along lines shown in Tiao and Zellner [1964]. One can also justify a sort of Cochrane-Orcutt [1949] procedure (Shiller [1973]). Starting from an initial guess, $\hat{\rho}$, this procedure will entail transforming the X matrix (not the underlying x variables nor the $\tilde{X}$ matrix!) by subtracting from each observation $\hat{\rho}$ times the lagged observation (remembering that the x variables have been sorted in terms of x_1 so that the preceding observation may not be the lagged observation). One may then find the vector of residuals $e = y-x\hat{\beta}$ and regress residuals on their lagged values to get a new estimate ρ. One repeats the procedure, and when it converges one will have estimates $\hat{\beta}$ and $\hat{\rho}$ which simultaneously satisfy first order conditions for maximum of the posterior distribution.

Footnotes

*Yale University and National Bureau of Economic Research. I am grateful for discussions with Robert Engle, Jerry Hausman, and Dale Poirier. Danny Quah and Nigel Wilson provided research assistance and computing was done on the TROLL System. This research was supported by the National Science Foundation, under grant #SES-8105837.

[1] Recall the well-known fallacy of testing for the random walk character of stock prices by testing with annual data for serial correlation of changes in annual average stock market prices.

[2] For a nonsingular matrix Ψ partitioned as $\Psi = \begin{matrix} \Psi_{11} & \Psi_{12} \\ \Psi_{21} & \Psi_{22} \end{matrix}$ if Ψ_{22} is nonsingular and $\Gamma \equiv \Psi^{-1}$ is partitioned conformably then Γ_{11} is nonsingular, $\Gamma_{11}^{-1} = \Psi_{11} - \Psi_{12}\Psi_{22}^{-1}\Psi_{21}$ and $\Gamma_{21} = - \Psi_{22}^{-1}\Psi_{21}\Gamma_{11}$. The analogous rule holds for Γ_{22}^{-1} and Γ_{12} if Ψ_{11} is nonsingular.

[3] If $g = 0$ and if $X = I$, this expression for $\hat{\beta}$ is implicit in expressions (4.2) to (4.4) in Wahba [1978]

[4] These results regarding the student posterior can all be found in Raiffa and Schlaifer [1964].

[5] Estimation was performed using TROLL Macros &SMP, for discrete smoothness priors, and &GLSSMP, for continuous smoothness priors. Both were written by Nigel Wilson.

[6] The rule of thumb here has not been justified formally. The choice of k with smoothness priors in the context of the distributed lag estimation problem has been discussed by Fomby [1979], Maddala [1974], Ullah and Raj [1979], and Thurman and Swamy [1980].

REFERENCES

Breiman, L. and W.S. Meisel, "General Estimates of the Intrinsic Variability of Data in Nonlinear Regression Models," Journal of the American Statistical Association, 71:301-7, 1976.

Cochrane, D. and Orcutt, G.H., "Application of Least Squares Regressions to Relationships Containing Auto-Correlated Error Terms," Journal of the American Statistical Association, 44:32-61, 1949.

Cover, T.M., "Estimation by the Nearest Neighbor Rule," IEEE Transactions on Information Theory, 14:50-55, 1968.

Fomby, T.B., "MSE Evaluations of Shiller's Smoothness Priors," International Economic Review Vol. 20, 203-15, 1979.

Friedman, Jerome H. and Werner Stuetzle, "Projection Pursuit Regression," Journal of the American Statistical Association, Vol. 76 No. 376 pp 817-23, 1981.

Gersovitz, Mark and James G. MacKinnon, "Seasonality in Regression: An Application of Smoothness Priors," Journal of the American Statistical Association, 73:264-73, Vol. 73 No. 362 June 1978.

Greville, T.N.E., Theory and Applications of Spline Functions, Academic Press, New York, 1969.

Jeffreys, H., Theory of Probability, London, Oxford University Press, 1961.

Kimeldorf, George S. and Grace Wahba, "A Correspondence between Bayesian Estimation on Stochastic Processes and Smoothing by Splines," Annals of Mathematical Statistics, 41:495-502, No. 2, 1970.

Leamer, Edward, E., "Regression Selection Strategies and Revealed Priors," Journal of the American Statistical Association, 73:1978, pp. 580-7.

Maddala, G.S., "Ridge Estimators for Distributed Lag Models," NBER Computer Research Center Working Paper No. 69, Cambridge, MA, 1974.

McCulloch, J. Huston, "The Tax Adjusted Yield Curve," Journal of Finance, Vol. (No. 3), pp. 811-30, June 1975.

Poirier, Dale J., The Econometrics of Structural Change with Special Emphasis on Spline Functions, North-Holland Publishing Co., Amsterdam, 1976.

Raiffa, H. and R. Schlaifer, Applied Statistical Decision Theory, Cambridge, Harvard Univeristy Press, 1964.

Schoenberg, I.J., "Spline Functions and the Problem of Graduation," Proceedings of the National Academy of Science, U.S.A. 52 (1964), 947-950.

Sokolnikoff, I.S., Mathematical Theory of Elasticity, McGraw-Hill, New York, 1956.

Stone, Charles J., "Consistent Nonparametric Regression," The Annals of Statistics, Vol. 5 No. 4, 595-645, July 1977.

Theil, Henri, Principles of Econometrics, 1971.

Thurman, S.S. and P.A.V.B. Swamy, "An Operational Method for Estimating Distributed Lag Coefficient with Smoothness Priors," Board of Governors of the Federal Reserve System, 1980.

Tiao, George C. and Arnold Zellner, "Bayesian Analysis of the Regression Model with Autocorrelated Errors," Journal of the American Statistical Association, 59:763-78, 1964.

Ullah, A. and B. Raj, "A Distributed Lag Estimator Derived from Shiller's Smoothness Priors," Economic Letters Vol. 2, 219-23, 1979.

Wahba, Grace, "Improper Priors, Spline Smoothing and the Problem of Guarding Against Model Errors in Regression," Journal of the Royal Statistical Society B, Vol. 40, No. 3, pp. 364-72, 1978.

Whittaker, Edmund and G. Robinson, The Calculus of Observations, 4th Edition, Dover, New York, 1967.