

Spline smoothing with model-based penalties

J. O. RAMSAY

McGill University, Montreal, Quebec, Canada

N. HECKMAN

University of British Columbia, Vancouver, British Columbia, Canada

and

B. W. SILVERMAN

University of Bristol, Bristol, England

Nonparametric regression techniques, which estimate functions directly from noisy data rather than relying on specific parametric models, now play a central role in statistical analysis. We can improve the efficiency and other aspects of a nonparametric curve estimate by using prior knowledge about general features of the curve in the smoothing process. Spline smoothing is extended in this paper to express this prior knowledge in the form of a linear differential operator that annihilates a specified parametric model for the data. Roughness in the fitted function is defined in terms of the integrated square of this operator applied to the fitted function. A fast $O(n)$ algorithm is outlined for this smart smoothing process. Illustrations are provided of where this technique proves useful.

Function or curve estimation is among the oldest problems in experimental psychology, and remains one of its central statistical objectives. Indeed, every paper presented at the conference to which this was a contribution mentioned an explicit function-estimation problem or presented results based implicitly on this technology.

Traditionally, curve estimation proceeded by the scientist proposing a parametric function family such as $f(x) = \alpha x^\beta$ that seemed to both capture the observed features of the relationship as mirrored in the data and to be consistent with a priori intuitions and theoretical considerations. The statistical problem then reduced to finding appropriate estimates of these parameters, in this example α and β . But one often discovered that some minor modifications were necessary, and so, for example, might be led to consider $f(x) = \alpha(x - \gamma)^\beta$. But these seemingly innocent adjustments of a model often turn out to make the estimation of the principal or “structural” parameters much more difficult, and in any case there was often some ambiguity about when to stop adding “nuisance” parameters such as the threshold γ .

The last few decades have seen an explosion of methods that seek to estimate a function directly, without the intermediate device of parametric models. That is, in effect the function itself becomes the parameter to be estimated. These nonparametric regression methods have also been extended to multidimensional arguments. Use-

ful references are Eubank (1988), Härtle (1990), and Simonoff (1996). The three main classes of methods are kernel smoothing, local polynomial smoothing, and spline smoothing.

But there is clearly a need to retain some of the flavor of earlier parametric investigations. The scientist often has good reason to propose that a major part of a function has a linear, power-law, exponential, or sinusoidal character, and therefore is quite justified on theoretical grounds for asking that the functional estimation problem use this information. Moreover, it can be shown theoretically that a nonparametric regression technique will perform better in various ways if some large part of the actual relationship can be correctly specified in advance of collecting the data.

This paper presents an extension of spline smoothing technology that allows the investigator to retain some aspects of a parametric model in a nonparametric regression situation.

INTRODUCTION TO SPLINE SMOOTHING

Cubic Spline Smoothing

The classic spline smoothing method estimates a curve $x(s)$ from observations,

$$y_j = x(t_j) + \varepsilon_j, j = 1, \dots, n, \quad (1)$$

by making explicit two possible aims in curve estimation. On the one hand, we wish to ensure that the estimated curve gives a good fit to the data; for example, in terms of residual sum of squares:

$$SSE(x|y) = \sum_j [y_j - x(t_j)]^2. \quad (2)$$

J.O.R. and N.H. wish to acknowledge the support of the Natural Sciences and Engineering Research Council of Canada through Grants A320 and A7969, respectively. Correspondence should be addressed to J. O. Ramsay, Department of Psychology, 1205 Dr. Penfield Ave., Montreal, PQ, Canada H3A 1B1 (e-mail: ramsay@psych.mcgill.ca).

On the other hand, we do not wish the fit to be too good if this results in a curve x that is excessively “wiggly” or locally variable.

A very common way of measuring the roughness of a function on an interval \mathcal{T} is its integrated squared second derivative,

$$PEN_2(x) = \int_{\mathcal{T}} [D^2x(s)]^2 ds = \|D^2x\|^2, \quad (3)$$

using the notation $D^m x$ to indicate the m th derivative of x , $d^m x/dt^m$. We can then define the *penalized* residual sum of squares as

$$SSE_{\lambda}(x|y) = SSE(x|y) + \lambda PEN_2(x). \quad (4)$$

The parameter λ is a smoothing parameter that measures the “rate of exchange” between fit to the data, as measured by residual sum of squares, and variability of the function x , as quantified by $PEN_2(x)$. Our estimate of the function is obtained by finding the function x that minimizes $SSE_{\lambda}(x)$ over the space of functions x for which $PEN_2(x)$ is defined.

For this particular roughness penalty, the resulting curve $x(s)$ can be shown to be a cubic spline with knots at the data points t_j . Many details of the method are discussed by Green and Silverman (1994) and elsewhere, and we shall discuss it only fairly superficially here. For any fixed λ , the function $x(s)$ can be found in $O(n)$ operations, for example, by using the S-PLUS function `smooth.spline` (Statistical Sciences, 1995), which also contains options for choosing the smoothing parameter automatically from the data.

To comment briefly on the role of the smoothing parameter λ , note that if λ is very large, all functions that are not linear will incur a substantial roughness penalty in $SSE_{\lambda}(x)$. For this reason, as $\lambda \rightarrow \infty$, the fitted curve x approaches the standard linear regression to the observed data. On the other hand, for small λ , the curve will tend to become more and more variable, since there is less and less penalty placed on its roughness, and as $\lambda \rightarrow 0$, the curve x will approach an interpolant to the data, satisfying $x(t_j) = y_j$ for all j s. However, even in this limiting case, the interpolating curve will not be arbitrarily variable; instead, it will be the smoothest twice-differentiable curve that can be found that exactly fits the data.

The Bias–Variance Tradeoff

In fact, the spline smoothing criterion (Equation 3) simply states one of the basic principles of statistics:

$$\text{Mean Squared Error} = \text{Bias}^2 + \text{Sampling Variance}$$

That is, according to the error model (Equation 1) that is used to motivate most smoothing methods, a completely unbiased estimate of function value $x(t_j)$ can always be produced by fitting y_j exactly, since this observed value is itself an unbiased estimate of $x(t_j)$. But, of course, we recognize that lack of bias is not everything in statistics, and in fact mean squared error,

$$MSE = E[\hat{x}(t_j) - x(t_j)]^2, \quad (5)$$

comes closer to capturing what we usually mean by badness of estimate (as opposed to fit). *MSE* can very often be dramatically reduced by sacrificing some bias in order to reduce sampling variance, $\text{Var}[\hat{x}(t_j)]$, and this is the main reason why we want to impose smoothness on \hat{x} . By requiring that \hat{x} vary only gently from one value to another, we are in fact borrowing information from neighboring data values and thereby expressing our faith in the regularity of the underlying function \hat{x} that we are trying to estimate.

The L-spline Criterion

The cubic spline smoothing can be extended by using measures of roughness other than PEN_2 . There are really two different (but related) reasons for doing this. On the one hand, we may wish the class of functions that have zero roughness to be wider than, or otherwise different from, those that are of the form $a + bs$. For example, if we desire a smooth estimate of acceleration D^2x , we may well want to penalize the size of D^4x , thereby directly controlling the curvature of the acceleration function and at the same time imposing zero penalty on any cubic polynomial.

On the other hand, we may have in mind that, locally at least, curves x should ideally satisfy a particular differential equation, and we may wish to penalize departure from this. For instance, if we were observing periodic data on an interval $[0, T]$ and there were some reason to suppose that simple harmonic motion with period ω was a natural approximate model for the data, we know that $\omega^2x + D^2x = 0$ is the linear differential equation satisfied by this type of variation. If the actual x deviates from purely harmonic behavior, we can expect that $\omega^2x + D^2x$ will be nonzero.

We can achieve both of these aims by replacing the second derivative operator D^2 by a more general linear differential operator L , defined as

$$Lx = w_0x + w_1Dx + \dots + w_{m-1}D^{m-1}x + D^m x, \quad (6)$$

where w_j , $j = 0, \dots, m-1$ are also functions and where m is the order of the differential operator L . We can then define

$$PEN_L(x) = \int_{\mathcal{T}} [Lx(s)]^2 ds = \|Lx\|^2, \quad (7)$$

the integral of the square of $Lx(s)$. For instance, the simple harmonic motion example above would lead us to use $Lx(s) = \omega^2x(s) + D^2x(s)$. Other examples will be given below.

Now it can be shown (Wahba, 1990) that the integrated squared bias in a spline smoothing estimate \hat{x} of a smooth function x assumed to underly the noisy discrete observations has the following bound:

$$\text{Bias}^2(\hat{x}) = \int \{x(t) - E[\hat{x}(t)]\}^2 dt \leq \int (Lx)^2 dt,$$

or more compactly, $\text{Bias}^2(x) \leq \|Lx\|^2$. This is an important result because it tells us that bias, which we would naturally like to be small, is limited by how closely our choice of differential operator L comes to annihilating the “true” function x . Thus, if x is predominantly of some known form and one for which there is a linear differential operator L that annihilates it, then it makes sense to choose this

same L so as to at least annihilate this component, thereby ensuring that the Lx is probably small, and consequently that the estimate \hat{x} is relatively unbiased when compared with, say, using $L = D^2$.

TWO EXAMPLES

Gross Domestic Product Data

The gross domestic product of a country shares with many economic indicators an overall tendency for exponential growth. Moreover, when the data are available on a quarterly basis, there is also obvious seasonal variation. Figure 1 displays the gross domestic product for Sweden over a 15-year period (OECD, 1995).

This suggests the use of the order 4 composite operator

$$L = (-\gamma D + D^2)(\omega^2 I + D^2) = -\gamma\omega^2 D + \omega^2 D^2 - \gamma D^3 + D^4 \quad (8)$$

to annihilate

$$\mathbf{u}(t) = (1, \exp \gamma t, \sin \omega t, \cos \omega t)'$$

In this application we know that $\omega = 2\pi$ for time measured in years, and a nonlinear least squares estimate for γ yields 0.078.

The Melanoma Data

These data, the incidence of malignant melanomas in the Connecticut Tumor Registry, taken from Andrews and Herzberg (1985) and displayed in Figure 2, represent another complex relationship, with a cyclic effect superimposed on a linear development. The operator that would be interesting would be

$$L = \omega^2 D^2 + D^4 \quad (9)$$

for some appropriate constant ω , since this would annihilate the four functions

$$\mathbf{u}(t) = (1, t, \sin \omega t, \cos \omega t)'$$

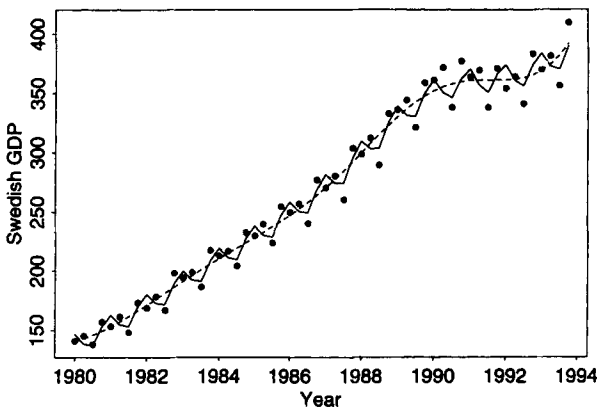


Figure 1. The gross domestic product for Sweden with seasonal variation. The solid line is the smooth using operator $L = (-\gamma D + D^2)(\omega^2 I + D^2)$, and the dashed line is the smooth for $L = D^4$, the smoothing parameter being determined by minimizing the *GCV* criterion in both cases.

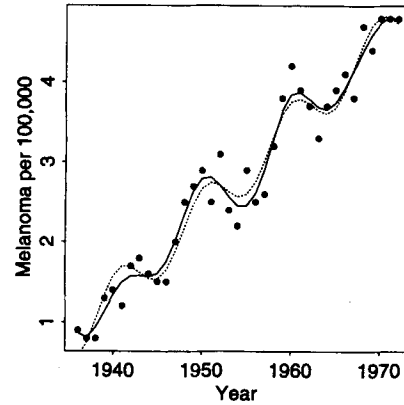


Figure 2. Age-adjusted incidences of melanoma for the years 1936 to 1972. The solid curve is the polynomial smoothing spline fit to the data penalizing the norm of the fourth derivative, with the smoothing parameter chosen by minimizing the *GCV* criterion. The dotted line is the fit of $x(t) = c_1 + c_2 t + c_3 \sin(\omega t) + c_4 \cos(\omega t)$ with period $\omega = 0.65$ and coefficients estimated by least squares.

SOME LINEAR DIFFERENTIAL EQUATION FACTS

In order to present some of the details about how to smooth data using this more general class of penalties, we need to first mention some basic facts from the theory of linear ordinary differential equations. It is, of course, far beyond the scope of this article to offer even a cursory treatment of a topic as rich as the theory of differential equations, and there would be little point since there are many fine texts on the topic. One of our favorites is Coddington (1989), and for advice on a wide range of practical matters, Press, Teukolsky, Vetterling, and Flannery (1992) is recommended. Instead what we offer here are a few facts about linear differential equations that are useful in the context of *L*-spline smoothing.

Identifying Linear Differential Operator *L*

For a linear differential operator L of order m coupled with appropriate initial value constraints, the homogeneous differential equation $Lu = 0$ has exactly m solutions u_j that are linearly independent. Such a set is not uniquely defined, and in fact for any such set an arbitrary linear combination $\sum_j \beta_j u_j$ is also a solution. Thus the homogeneous equation in fact defines an m -dimensional space of functions, called the *kernel* of L and indicated by $\ker L$.

We have already cited a number of examples where we had a set of known functions $\mathbf{u} = (u_1, \dots, u_m)'$ and where at the same time we were aware of the linear differential operator L that solved the homogeneous linear differential equations $Lu_j = 0, j = 1, \dots, m$. Suppose, however, that we have the u_j s in mind but that the linear differential operator that annihilates them is not obvious, and we want to find it.

For example, suppose that two functions u_1 and u_2 are to be annihilated. Then $m = 2$, and L is of the form

$$Lx = w_0u + w_1Du + D^2u.$$

The relation $Lu_1 = Lu_2 = 0$ defines the differential operator, and it can be expressed as follows by taking the second derivatives over to the other side of the equation:

$$\begin{aligned} w_0u_1 &= w_1Du_1 = -D^2u_1 \\ w_0u_2 &= w_1Du_2 = -D^2u_2, \end{aligned} \tag{10}$$

or, in matrix notation,

$$[u \quad Du] w = -D^2u. \tag{11}$$

This is a linear matrix equation for the unknown weight functions w_0 and w_1 , and its solution is simple provided that the matrix

$$\mathbf{W} = [u \quad Du] \tag{12}$$

is nowhere singular, or in other words, that its determinant $|\mathbf{W}|$ does not vanish for any value of the argument t . This coefficient matrix is called the *Wronskian matrix*, and its determinant $|\mathbf{W}|$ is called the *Wronskian* for the system. Clearly w can be calculated simply as $-\mathbf{W}^{-1} D^2u$ if $|\mathbf{W}| \neq 0$.

Note that the functions to be annihilated need not be known analytically; in many problems, such as the growth curve example described at the end of this paper, they are known only as a result of numerical calculations. But numerical techniques can be employed to estimate the weight functions w_j in this case. Moreover, Ramsay (1996) has developed a method for estimating these weight functions, and hence the operator L , directly from replicated curves. This approach, called *principal differential analysis*, has a strong conceptual connection to principal components analysis.

The Green's Function for L

Suppose now that we want to reverse the effect of applying an m th order linear differential operator L . That is, we have a *forcing function* f satisfying

$$Lx = f \tag{13}$$

and we want to find x . Well, we recognize of course that the solution will not be unique; if we add to any solution x some linear combination of m functions u_j that span $\ker L$, this function also satisfies the equation. But there is a unique solution to the equation given the associated equation

$$B_t x = \mathbf{x}_0, \tag{14}$$

where $B_t x$ is the set of m values

$$\mathbf{x}_0 = [x(0), Dx(0), \dots, D^{m-1}x(0)]'.$$

The operator B_t is called the *initial value operator*, and when applied to a function x returns the values of the function and its first $m-1$ derivatives at time $t = 0$. It is assumed here that the only function x simultaneously satisfying $B_t x = 0$ and $Lx = 0$ is zero itself.

It can be shown that there exists a bivariate function $G(t;w)$ called the *Green's function*, associated with the pair (B_t, L) that satisfies

$$\begin{aligned} x(t) &= \int_{\mathcal{T}} G(t;w) Lx(w), dw \\ &= \int_{\mathcal{T}} G(t;w) f(w) dw, x \in \ker B. \end{aligned} \tag{15}$$

Thus the Green's function defines an integral transform that inverts the linear differential operator L . Applying $\int G(t; \cdot)$ to Lx gets us back to x itself. The Green's function G is called the *kernel* of the integral transform.

There is a simple recipe for constructing the Green's function corresponding to the initial value constraint B_t and any linear differential operator L . First, compute the Wronskian matrix \mathbf{W} defined in Equation 12. Second, define the vector-valued function \mathbf{v} as the vector containing the elements of the last row of \mathbf{W}^{-1} . Then, it turns out that

$$\begin{aligned} G(t;w) &= \sum_{j=1}^m u_j(t)v_j(w) \\ &= \mathbf{u}'(t)\mathbf{v}(w), w \leq t, \text{ and } 0 \text{ otherwise.} \end{aligned} \tag{16}$$

The Reproducing Kernel Associated with L

The concept of a reproducing kernel plays a central role in the theory of spline functions. This is a bivariate function $k(s,t)$ defined by the interesting property that

$$\int_{\mathcal{T}} L_s k(t,s) Lx(s) ds = x(t),$$

provided that $B_t x = 0$. The notation L_s means that the differential operator is applied to the second argument. The reproducing kernel k can be easily calculated as follows

$$k(s,t) = \int_{\mathcal{T}} G(s,w) G(t,w) dw. \tag{17}$$

Now let us assume that we have in hand m linearly independent functions u_j , each satisfying $Lu_j = 0$. Then the importance of the reproducing kernel derives from the following theorem, a proof of which can be found in references on splines such as Wahba (1990):

Optimal basis theorem: The function x minimizing the spline smoothing criterion (Equation 7) defined by a linear differential operator L of order m has the expansion

$$x(t) = \sum_{j=1}^m d_j u_j(t) + \sum_{i=1}^n c_i k(t_i, t). \tag{18}$$

AN EFFICIENT ALGORITHM FOR L-SPLINE SMOOTHING

In smoothing long sequences of observations, it is critical to devise a smoothing procedure that requires a number of arithmetic operations that is proportional to n , the length of the sequence. Such an algorithm is referred to as being $O(n)$, or of order n . An algorithm that was of $O(n^2)$ or $O(n^3)$ would be, for example, quite impractical for thousands of sampling points t_j .

The following algorithm is based on the theoretical paper of Anselone and Laurent (1967), but is also known

as the Reinsch algorithm because of the application to the cubic polynomial smoothing case ($L = D^2$) by Reinsch (1967, 1970). It was subsequently extended by Hutchison and de Hoog (1985). A technical account and rationale are available in Heckman and Ramsay (1996).

The algorithm requires the computation of values of two types of functions:

1. $u_j, j = 1, \dots, m$: a set of m linearly independent functions satisfying $Lu_j = 0$, that is, spanning $\ker L$. We shall refer to these collectively as the vector-valued function \mathbf{u} .

2. k : the reproducing kernel function for the subspace of functions e satisfying $B_l e = 0$, where B_l is the initial value constraint operator.

These two sets of functions are the user-supplied components of the algorithm and are, of course, defined by the particular choice of operator L used in the smoothing application.

The algorithm naturally splits into three phases: (1) an initial setup phase that does not depend on the smoothing parameter λ , (2) a smoothing phase in which the data are smoothed, and (3) a summary phase in which performance measures for the smooth are computed. This division of the task is of practical importance because we may want to try smoothing with many values of λ , and will naturally not want to needlessly repeat either the initial setup phase or the final descriptive phase.

Setting Up the Smoothing Procedure

In the initial phase, we define two symmetric band-structured matrices \mathbf{H} and $\mathbf{C}'\mathbf{C}$, both of order $n - m$, where m is the order of operator L . A matrix is said to be band structured if all entries except those no more than a fixed number of positions away from the diagonal are zero. These band-structured matrices require only $(n - m)(m + 1)$ storage locations, and can be processed in various ways in a number of operations that are proportional to their order, whereas otherwise most matrix operations such as matrix multiplies take either $O(n^2)$ or $O(n^3)$ operations.

The band-structured character of these two matrices depends on computing for each $i = 1, \dots, n - m$ a set of $m + 1$ coefficients

$$\mathbf{c}_i = (c_{i1}, \dots, c_{i,m+1})'$$

with a special property. Let $m + 1$ by m matrix \mathbf{U}_i have rows $\mathbf{u}'(t_{i+\ell}), \ell = 0, \dots, m$. The property that \mathbf{c}_i must have is

$$\mathbf{U}_i' \mathbf{c}_i = \mathbf{c}_i' \mathbf{U}_i = 0. \quad (19)$$

That is, coefficients \mathbf{c}_i must be orthogonal to all columns of \mathbf{U}_i .

There are various ways to compute such a coefficient vector, but probably the most efficient general method is to use the QR decomposition:

$$\mathbf{U}_i = \mathbf{Q}\mathbf{R}$$

where matrix \mathbf{Q} is square, of order $m + 1$, and orthonormal, and where matrix \mathbf{R} is $m + 1$ by m and upper triangular. The QR decomposition is a standard tool in matrix

computations and can be found in any matrix-oriented programming language such as MATLAB (MathWorks, 1993) or S-PLUS (Statistical Sciences, 1995), and in all comprehensive libraries of subroutines or procedures in lower level languages. The required coefficient \mathbf{c}_i is simply the last column of matrix \mathbf{Q} . In special cases, however, there are also other computational alternatives, and in the famous polynomial spline-smoothing case, coefficients defining divided differences are employed.

With these coefficients in hand, we can now define the n by $n - m$ matrix \mathbf{C} as follows: In column i are to be found the $m + 1$ coefficients \mathbf{c}_i starting in row i ; elsewhere the matrix contains zeros. This defines one of the setup-phase matrices $\mathbf{C}'\mathbf{C}$, which is symmetric, of order $n - m$, and band structured.

Now let the symmetric order n matrix \mathbf{K} contain the values $k(t_i, t_j), i, j = 1, \dots, n$. Then the other setup-phase matrix is

$$\mathbf{H} = \mathbf{C}'\mathbf{K}\mathbf{C}. \quad (20)$$

This, too, is symmetric, of order $n - m$, and band structured.

The Smoothing Phase

The actual smoothing consists of two steps:

1. The solution to the order $n - m$ linear equation system

$$(\mathbf{H} + \lambda\mathbf{C}'\mathbf{C})\mathbf{d} = \mathbf{C}'\mathbf{y} \quad (21)$$

for the vector \mathbf{d} of length $n - m$, where vector \mathbf{y} contains the values to be smoothed.

2. The vector of values \mathbf{x} of the smoothing function x at the n argument values are then computed by

$$\mathbf{x} = \mathbf{y} - \lambda\mathbf{C}\mathbf{d}. \quad (22)$$

Both of these steps can be computed in $O(n)$ operators, and references on efficient matrix computation such as Golub and van Loan (1989) can be consulted for details.

The Performance Assessment Phase

The vector of smoothed values \mathbf{x} and the values that were smoothed are related as follows:

$$\mathbf{x} = [\mathbf{I} - \lambda\mathbf{C}(\mathbf{H} + \lambda\mathbf{C}'\mathbf{C})^{-1}\mathbf{C}']\mathbf{y} = \mathbf{S}\mathbf{y}. \quad (23)$$

The matrix \mathbf{S} defined by $\mathbf{S} = \mathbf{I} - \lambda\mathbf{C}(\mathbf{H} + \lambda\mathbf{C}'\mathbf{C})^{-1}\mathbf{C}'$ is often called the *hat matrix* and in effect defines a linear transformation that maps the unsmoothed data into its smooth image.

Various measures of performance depend on the diagonal values in \mathbf{S} . Of these the most important are

$$GCV = SSE / (1 - n^{-1} \text{trace } \mathbf{S})^2, \quad (24)$$

where

$$SSE = \sum_{i=1}^n [y_i - x(t_i)]^2 = \|\mathbf{y} - \mathbf{x}\|^2,$$

and

$$CV = \sum_{i=1}^n \{[y_i - x(t_i)]/[1 - s_{ii}]\}^2, \quad (25)$$

where s_{ii} is the i th diagonal entry of \mathbf{S} . Both measures can be computed in $O(n)$ operations given the band-structured nature of the matrices defining \mathbf{S} .

One of the main applications of these two criteria, both of which are types of “discounted” error sums of squares, is as a guide for choosing the value of the smoothing parameter λ . It is relatively standard practice to look for the value that minimizes one of these two criteria, just as in standard regression analysis various variable selection procedures attempt to minimize discounted error sums of squares. Interestingly, the GCV measure was originally introduced by Craven and Wahba (1979) as an approximation to the CV criterion that could be computed in $O(n)$ operations, but in fact now tends to be preferred in practice, even though CV is also available in $O(n)$ operations because various simulated studies have indicated that GCV tends to be a better basis for choosing the smoothing parameter λ . Hutchison and de Hoog (1985) developed the $O(n)$ techniques for computing these values.

Also of great value is a measure of the effective number of degrees of freedom of the smoothing operation. Two measures are

$$DF_1 = \text{trace } \mathbf{S} \text{ and } DF_2 = \text{trace } \mathbf{S}'\mathbf{S}. \quad (26)$$

It can be shown that in the limit as $\lambda \rightarrow \infty$, both measures become simply m , and similarly as $\lambda \rightarrow 0$, both measures converge to n . In between, they give slightly different impressions of how much of the variation in the original unsmoothed data remains in the smoothed version, or, alternatively, how big the “dimensionality” is of the smoothing function.

As a rough indication of the time taken for a typical smooth, our implementation in S-PLUS of this algorithm running on a Sparcstation 2 was able to smooth 10,000 observations using an order 4 penalty in about 35 sec. Of course, timings will vary enormously even for the same machine, depending on factors such as memory available, programming environment, and disk hardware.

A RETURN TO THE EXAMPLES

The Melanoma Data

The parameter ω was estimated to be 0.650 by the minimization of the nonlinear least squares criterion

$$SSE = \sum_j^n [y_j - \beta_1 - \beta_2 t_j - \beta_3 \sin(\omega t_j) - \beta_4 \sin(\omega t_j)]^2$$

with respect to the four linear parameters β_j and the phase parameter ω . This yielded to a period of 9.66 years, roughly the period of the sunspot cycle affecting solar radiation and consequently melanoma. When we smooth the data with the spline defined by the operator (Equation 9) and select λ so as to minimize GCV , it turns out that λ becomes arbitrarily large, corresponding to a smooth using only the basis functions \mathbf{u} , consuming four degrees of freedom, and yielding $GCV = 0.076$. However, the polynomial smoothing spline with order $m = 4$ —that is,

using the differential operator $L = D^4$ —displayed in Figure 2 produced a minimum GCV estimate corresponding to $DF_1 = 12.0$ and $GCV = 0.095$. That is, it required three times the degrees of freedom to produce a fit that was still worse in GCV terms than the L-spline smooth. Clearly of the two order-4 methods, the operator (Equation 9) is much to be preferred to $L = D^4$.

GDP Data with Seasonal Effects

The minimum GCV L-spline for these data is the solid line in Figure 1, and was fit using $\omega = 2\pi$ and $\gamma = 0.078$, the latter value being estimated by nonlinear least squares as in the melanoma example. The fit by this spline yielded $GCV = 142.9$, $SSE = 5,298$, and $DF_1 = 10.4$. This fairly low-dimensional spline is able to track both the seasonal and long-term variation rather well.

By contrast, the minimum GCV polynomial spline corresponding to $L = D^4$ is shown by the dashed line, and corresponds to $GCV = 193.8$, $SSE = 8,169$, and $DF_1 = 7.4$. As both the curve itself and the GCV value indicate, the polynomial spline was completely unable to model the seasonal variation, and treated it as noise. On the other hand, reducing the smoothing parameter λ to the point where SSE was reduced to the same value as was attained for the L-spline required $DF_1 = 28.2$, or nearly three times the degrees of freedom. Again we see that building into the operator L the capacity to model important sources of variation pays off handsomely.

SIMULATED HUMAN GROWTH DATA

One of the triumphs of nonparametric regression techniques has been their capacity to uncover previously unsuspected aspects of growth in skeletal height (Gasser, Müller, Köhler, Molinari, & Prader, 1984; Ramsay, Bock, & Gasser, 1996). In this illustration, spline smoothing using an estimated differential operator is applied to simulated smoothing data. The objective was to see whether estimating the smoothing operator improves the estimation of the height and height acceleration growth functions over an a priori “off-the-rack” smoother.

To investigate how the performance of the L-spline would compare with a polynomial spline in practice, we simulated data to resemble as much as possible actual human growth curve records. Two samples were generated: a training sample of 100 records that was analyzed in a manner representative of actual practice; and a validation sample of 1,000 records to see how these analyses would perform on data for which the analyses were not “tuned.”

The simulated data for both the training and validation samples consisted of growth records generated by using the triple logistic parametric nine-parameter growth model proposed by Bock and Thissen (1980). According to this model, height $h_i(t)$ at age t for individual i is

$$h_i(t) = \sum_{j=1}^3 c_{ij} / \{1 + \exp[-a_{ij}(t - b_{ij})]\}. \quad (27)$$

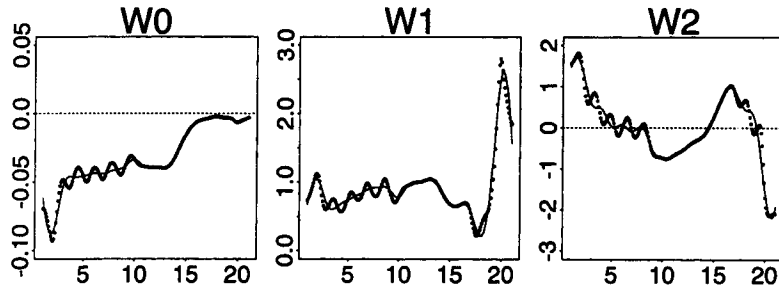


Figure 3. The three weight functions ω_0 , ω_1 , and ω_2 for the operator $L = \omega_0 I + \omega_1 D + \omega_2 D^2 + D^3$: The points indicate the pointwise-approximation, and the solid line indicates the basis function expansion.

This model, although not completely adequate to account for actual growth curves, does capture their salient features rather well. The actual number of parameters in the model turns out to be only eight, since it turns out that parameter $a_{i,1}$ can be expressed as a function of the other parameters.

Each record was generated by first sampling from a population of coefficient vectors having a random distribution estimated from actual data for males in the Fels growth study (Roche, 1992). The errorless growth curves (in centimeters) were computed for the 41 age values ranging from 1 to 21 in half-yearly steps, and the simulated data were generated by adding independent normal error with mean 0 and standard deviation 0.5 to these values. These simulated data had roughly the same variability as actual growth measurements.

Our two smoothing operators, one an L-spline and the other a polynomial spline, were both of order 3. The first step was to use the training sample to estimate the order 3 L-spline that comes as nearly as possible to annihilating the curves. To this end, the first analysis consisted of polynomial spline smoothing of the simulated data to get estimates of the first three derivatives. The smoothing operator used for this purpose was D^5 , implying that the smoothing splines were piece-wise polynomials of degree 9. This permits us to control the roughness of the third derivative in much the same way as a cubic smoothing spline controls the roughness of the smoothing function itself. The smoothing parameter was chosen to minimize the *GCV* criterion. With this amount of replicated data, this criterion yields a very stable value. Since our principal differential analysis estimate of the operator L required numerical integration, we also obtained function and derivative estimates at 201 equally spaced values $1(.1)21$.

A third-order differential operator L was estimated using both the point-wise technique and the basis function expansion approach outlined in Ramsay (1996). For the latter approach, we used the 23 order 4 B-splines defined by positioning knots at the integer values of age. The estimated weight functions ω_0 , ω_1 , and ω_2 for the operator $L = \omega_0 I + \omega_1 D + \omega_2 D^2 + D^3$ are displayed in Figure 3. Although these are difficult to interpret, we can see that ω_0 is close to 0, suggesting that the operator could be

simplified by dropping the first term. On the other hand, ω_1 is close to 1 until the age of 15, when the growth function has strong curvature as the pubertal growth spurt ends, and its strong variation after 15 is undoubtedly helping the operator to deal with this pronounced curvilinearity. The acceleration weight ω_2 varies substantially over the whole range of ages.

The three solutions u_j to $Lu = 0$, computed by the successive approximation method, are shown in Figure 4. Linear combinations of these three functions can produce good approximations to actual growth curves.

The next step is to use the estimated functions u_j to estimate the Green's function G and the reproducing kernel k associated with this operator. The integrals involved were approximated using the trapezoidal rule applied to the values at the 201 argument values.

Now we were ready to actually smooth the training sample data by the two techniques, L-spline and polynomial spline smoothing, both of order 3, very much as one would in practice. For both techniques we relied on the *GCV* criterion to choose the smoothing parameter. The polynomial smooth gave values of *GCV*, *DF*, and λ of 487.9, 9.0, and 4.4, respectively, and the L-spline smooth produced corresponding values of 348.2, 11.2, and 0.63.

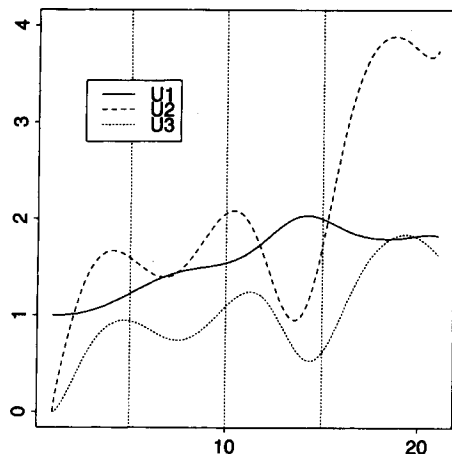


Figure 4. The lines show the three solutions to the homogeneous equation $Lu = 0$ corresponding to the linear differential operator L estimated for the simulated human growth data.

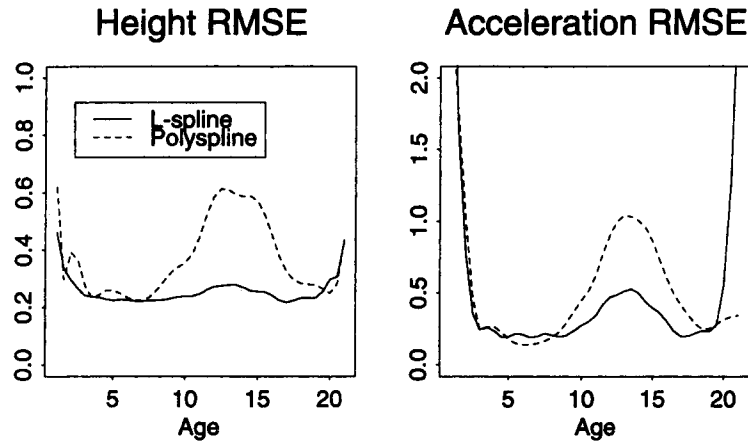


Figure 5. The left panel displays root-mean-squared error (RMSE) as a function of age for the simulated growth data. The solid line is for smoothing using the estimated differential operator L , and the dashed line is for polynomial smoothing using $L = D^4$. The right panel shows these results for the estimated height acceleration.

How well would these two smoothing techniques approximate the curves generating the data? To answer this question, we then generated 1,000 new simulated curves using the same generation process, and applied these two smoothers using the training sample values of λ . Since we knew the values of the true curves, we could compute the root-mean-squared-error (RMSE) criterion

$$RMSE(t) = \sqrt{E[\hat{x}(t) - x(t)]^2},$$

where the average was computed across the 1,000 curves for a given specific age t . This yielded the two RMSE curves displayed in Figure 5. We see that the estimate of both the growth curve itself and its acceleration by the L-spline procedure is much better for all but the final adult period, where the L-spline estimate of the acceleration curve becomes rather noisy and unstable. The improvement in the acceleration estimate prior to and during the pubertal growth spurt is especially impressive: The mean square error for the polynomial smooth is about four times that of the L-spline smooth. That is, using the L-spline is roughly equivalent to using the polynomial smooth with quadruple the sample size.

REFERENCES

- ANDREWS, D. F., & HERZBERG, A. M. (1985). *Data: A collection of problems from many fields for the student and research worker*. New York: Springer-Verlag.
- ANSELONE, P. M., & LAURENT, P. J. (1967). A general method for the construction of interpolating or smoothing spline-functions. *Numerische Mathematik*, **12**, 66-82.
- BOCK, R. D., & THISSEN, D. (1980). Statistical problems of fitting individual growth curves. In F. E. Johnston, A. F. Roche, & C. Susanne (Eds.), *Human physical growth and maturation: Methodologies and factors*. New York: Plenum.
- CODDINGTON, E. A. (1989). *An introduction to ordinary differential equations*. New York: Dover.
- CRAVEN, P., & WAHBA, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, **31**, 377-403.
- EUBANK, R. L. (1988). *Spline smoothing and nonparametric regression*. New York: Marcel Dekker.
- GASSER, T., MÜLLER, H.-G., KÖHLER, W., MOLINARI, L. & PRADER, A. (1984). Nonparametric regression analysis of growth curves. *Annals of Statistics*, **12**, 210-229.
- GOLUB, G., & VAN LOAN, C. F. (1989). *Matrix computations* (2nd ed.). Baltimore: Johns Hopkins University Press.
- GREEN, P. J., & SILVERMAN, B. W. (1994). *Nonparametric regression and generalized linear models: A roughness penalty approach*. London: Chapman & Hall.
- HÄRTLE, W. (1990). *Applied nonparametric regression*. Cambridge: Cambridge University Press.
- HECKMAN, N., & RAMSAY, J. O. (1996). *Some general theory for spline smoothing*. Unpublished manuscript. McGill University.
- HUTCHISON, M. F., & DE HOOG, F. R. (1985). Smoothing noisy data with spline functions. *Numerische Mathematik*, **47**, 99-106.
- MATHWORKS, INC. (1993). *MATLAB reference guide*. Natick, MA: Author.
- ORGANIZATION FOR ECONOMIC COOPERATION AND DEVELOPMENT. (1995). *Quarterly National Accounts*, **3**.
- PRESS, W. H., TEUKOLSKY, S. A., VETTERLING, W. T., & FLANNERY, B. P. (1992). *Numerical recipes in Fortran* (2nd ed.). Cambridge: Cambridge University Press.
- RAMSAY, J. O. (1996). Principal differential analysis: Data reduction by differential operators. *Journal of the Royal Statistical Society, Series B*, **58**, 495-508.
- RAMSAY, J. O., BOCK, R. D., & GASSER, T. (1996). Comparison of height acceleration curves in the Fels, Zurich, and Berkeley growth data. *Annals of Human Biology*, **22**, 413-426.
- REINSCH, C. (1967). Smoothing by spline functions. *Numerische Mathematik*, **10**, 177-183.
- REINSCH, C. (1970). Smoothing by spline functions II. *Numerische Mathematik*, **16**, 451-454.
- ROCHE, A. (1992). *Growth, maturation and body composition: The Fels longitudinal study 1929-1991*. Cambridge: Cambridge University Press.
- SIMONOFF, J. S. (1996). *Smoothing methods in statistics*. New York: Springer-Verlag.
- STATISTICAL SCIENCES. (1995). *S-PLUS guide to statistical and mathematical analysis, Version 3.3*. Seattle: StatSci, a division of MathSoft, Inc.
- WAHBA, G. (1990). *Spline models for observational data*. Philadelphia: Society for Industrial and Applied Mathematics.