A "New" Approach to the Smoothing Problem

by

W.E. Diewert and T.J. Wales*

March 1998

I. Introduction

In this paper, we consider a "new" method for nonparametrically smoothing a function of one variable. Specifically, we consider the following model:

$$(1) \qquad y_t = s_t + e_t \quad , \quad t = 1, 2, \ldots, N$$

where y_t is the observed value (the rough), s_t is the "smoothed" value and e_t is an unobserved erratic component at period or observation t.

There are a substantial number of smoothing methods (i.e., methods for choosing the s_t) that have been suggested in the literature.[1] Most of these methods involve a tradeoff between goodness of fit (closeness of the s_t to the y_t) and the "smoothness" of the smoothed values s_t.[2] Most authors define the smoothness of the s_t in terms of the magnitude of the sum of the second or third differences of the s_t. However, there are other possible definitions of smoothness. Our preferred definition of a smooth series is one where its second differences do not change sign too often.[3] This definition of smoothness is essentially due to the British actuary, T.B. Sprague (1887; 96).

In section 2, we discuss Sprague's definition of smoothness in more detail. In section 3, we discuss desirable properties for smoothing methods. In sections 4 and 5, we show how a Sprague smooth can be constructed by solving various quadratic programming problems. Our analysis here builds on the pioneering contributions of Hildreth (1954) on the nonparametric fitting of a concave function.

In section 6, we explain how our method of smoothing can be modified to deal with the problems that most nonparametric smoothing methods encounter in constructing the smooth at the end points of the data.

Two applications of our smoothing method are presented. In section 7, we apply our method to the smoothing of mortality data. In section 8, we apply our method to the fitting of a sawtooth function in a Monte Carlo study.[4] In this second application, we use the end point adjusted method explained in section 6.

Since our method of smoothing depends on the rapid solution of numerous quadratic programming problems, it is necessary to have an effective algorithm for solving these problems.[5] After various transformations, each of our problems has the following structure: minimize a strictly convex quadratic function subject to some nonnegativity constraints. We found that standard quadratic programming algorithms were not well suited to solving this class of problems, since the algorithms assumed that there were linear constraints in addition to nonnegativity constraints. Thus we found it useful to develop our own algorithm that made use of the special structure of our problem. This algorithm is explained in Appendix 1. Our algorithm draws on some of the ideas used by Beale (1955) (1959) (1967) in his algorithm and by Wilde and Beightler (1967; 58-81) in their differential algorithm. Our new algorithm was used to solve the problems that occurred in sections 7 and 8 and we found it to be very effective.

In our statement of the smoothing problem which follows (1), we have assumed that our independent variable t is evenly spaced. Our smoothing method can readily be extended to unevenly spaced data and this extension is presented in Appendix 2.

Section 9 concludes.

2. Alternative Definition of Smoothness

In this section, we shall review some of the definitions of smoothness that have been proposed by researchers over the years.

Define the first, second and third differences of the series $\{s_t\}$ by (2), (3) and (4) respectively:

$$(2) \qquad \Delta s_t \equiv s_t - s_{t-1} \quad , \qquad t = 2, 3, \ldots, N;$$

$$(3) \qquad \Delta^2 s_t \equiv \Delta s_t - \Delta s_{t-1} = s_t - 2s_{t-1} + s_{t-2} \quad , \qquad t = 3, 4, \ldots, N;$$

$$(4) \qquad \Delta^3 s_t \equiv \Delta^2 s_t - \Delta^2 s_{t-1} = s_t - 3s_{t-1} + 3s_{t-2} - s_{t-3} \quad , \qquad t = 4, 5, \ldots, N.$$

Whittaker (1923; 64) defined a series $\{s_t\}$ to be smooth if its squared third differences defined by (4) were "small" while Henderson (1924; 29-30) suggested that smallness of the squared first or second differences defined by (2) or (3) might suffice to define smoothness.[6] Whittaker and Henderson also suggested that smoothness and fit could be traded off by choosing a smoothing parameter λ. Given a choice for λ, Henderson's (1924;30) method for choosing the smoothed values in the model (1) reduced to solving the following unconstrained minimization problem in the s_t:

$$(5) \qquad \min_{s_1, s_2, \ldots, s_N} \Sigma_{t=1}^{N} (y_t - s_t)^2 + \lambda\, \Sigma_{t=3}^{N} (\Delta^2 s_t)^2.$$

The first sum of squares in (5) serves to measure the fit of the "smooth" $s \equiv (s_1, s_2, \ldots, s_N)$ to the "rough" $y \equiv (y_1, y_2, \ldots, y_N)$ while the second sum of squares measures the smoothness of the s_t. The parameter λ trades off fit and smoothness. Whittaker's (1923) smoothing model was similar except that third differences $\Delta^3 s_t$ were used in place of second differences $\Delta^2 s_t$ in (5). In the actuarial literature, the smoothing method that chooses the s_t by solving (5) for an exogenous λ is known as the Whittaker-Henderson method of graduation[7] and in the economics literature as the Hodrick-Prescott filter.[8] In the statistical literature, the smooth generated by solving (5) for an exogenous λ is called a smoothing spline.[9]

Bizley (1958; 126) criticized the above definitions of smoothness by noting that smooth looking series generated by an exponential function (so that $s_t = e^t$) have differences which never become small, no matter how high a difference is taken. This seems to be a very valid criticism which has been somewhat ignored in the current smoothing literature.

Bizley (1958; 133) proposed his own definition of smoothness: namely that the absolute value of the rate of change of curvature with respect to distance measured along the curve should be "small". However, Bizley was unable to work out any satisfactory way of determining what "small" might be in actual applications, and so his new definition of smoothness did not lead anywhere.[10]

The final definition of smoothness that we consider here was proposed many years ago by the British actuary, Thomas Bond Sprague (1887; 96): a series $\{s_t\}$ is smooth if the number of changes in sign of its second differences defined above by (3) is "small".[11] Sprague (1887; 93-95) observed that when the second differences of the data series changed sign, the curve spanned by the discrete series changed from being convex to concave or vice versa. Thus each time the second difference changed sign, the curve spanned by the data would exhibit a half wave or wiggle. Thus minimizing the number of sign changes would minimize the number of waves or wiggles in the curve. Sprague originally proposed his smoothness criterion in order to criticize the moving average method of smoothing proposed by Woolhouse (1870). Sprague (1887; 82) observed that Woolhouse's method applied to actuarial data led to an excessive number of wiggles in the smooth. This criticism of moving average methods applies equally well to virtually all nonparametric smoothing methods, with the exception of some regression based methods such as polynomial regression or regression splines.

Sprague's solution to the smoothing problem was to use graphical methods. Unfortunately, the graphical method is not sufficiently precise and reproducible for many scientific and business applications.[12] We shall attempt to put Sprague's graphical method of smoothing on a more scientific basis in sections 4 and 5 below.

Having considered alternative definitions of smoothness, we briefly turn our attention to some axiomatic properties for smoothing methods that have been suggested over the years. These properties will be helpful in evaluating the usefulness of our "new" smoothing method.

3. <u>Properties of Smoothing Methods</u>

What are desirable properties for smoothing methods? Before we attempt to answer this question, we first formally define a smoothing method.

A smoothing method could be defined as a function f from R^N into R^N; i.e., the <u>function</u> f maps an arbitrary "rough" vector y into the smoothed vector s = f(y). However, to allow for more complex smoothing procedures, we will allow the function f to be set valued; i.e., f maps a point y R^N into a subset f(y) of R^N. Thus in general, the smoothing method represented by f can be a <u>correspondence</u>.[13]

We define a point y R^N to be a <u>smooth point</u> for the smoothing method f if

(6) y f(y);

i.e., the initial unsmoothed point y belongs to the set of smoothed points f(y) which the smoothing method f generates. If the smoothing correspondence f is actually a function, then (6) becomes

(7) $y = f(y)$.

Sprague (1887; 79) seems to have been one of the first researchers to evaluate smoothing methods on the basis of their axiomatic properties. The first desirable property he suggested that a smoothing method should have is that the sum of the values of the smoothed series should equal the sum of the values of the original series; i.e., he suggested the following <u>sum preserving test</u>:[14]

(8) if $s \quad f(y)$, then $1_N \cdot s = 1_N \cdot y$

where 1_N is a vector of N ones and $x \cdot y$ signifies the inner product of the vectors x and y.

Our next test was first suggested by Whittaker (1923; 68); we shall call it the <u>first moment preserving test</u>:[15]

(9) if $(s_1, \ldots, s_N) \quad f(y_1, \ldots, y_N)$, then $\Sigma_{t=1}^{N} ts_t = \Sigma_{t=1}^{N} ty_t$.

If a discrete probability distribution is being smoothed, then a reasonable smoothing method should satisfy tests (8) and (9).

Our next four tests are implicitly due to Sprague (1887; 108-109) and may be found explicitly in Greville (1944; 204-211). Greville used these tests (and some additional ones) to evaluate alternative smoothing methods based on regression splines.[16]

Consider the following <u>identity test</u>:[17]

(10) if $s \quad f(k1_N)$, then $s = k1_N$ for every scalar k;

i.e., if the rough y is a constant vector, then the corresponding unique smooth s is the same constant vector, $k1_N$.

We term the next test, the <u>linear trend test</u>: if the rough y exhibits a linear trend, then the corresponding unique smooth s exhibits the same linear trend; i.e.,

(11) $y_t = \alpha + \beta t$ for $t = 1, \ldots, N$; $y \equiv (y_1, \ldots, y_N)$ implies $f(y) = \{y\}$.

The analogous *quadratic trend* and *cubic trend* tests are (12) and (13) respectively:[18]

(12) $y_t = \alpha + \beta t + \gamma t^2$ for $t = 1, \ldots, N$ implies $f(y) = \{y\}$;

(13) $y_t = \alpha + \beta t + \gamma t^2 + \delta t^3$ for $t = 1, 2, \ldots, N$ implies $f(y) = \{y\}$.

Roughs y that are constant or are generated by low order polynomials are evidently very smooth. Hence a reasonable smoothing method should simply reproduce these smooth roughs.

Schoenberg (1946; 52) proposed the following test which we term the *diminishing variation test*:[19]

(14) $s \quad f(y)$ implies $s \cdot s \leq y \cdot y$.

Thus if the smoother f satisfies (8) and (14), then the variance of the resulting smooth cannot exceed the variance of the rough. This too is a very reasonable test: a smoothing method should not generate a smooth that has a higher variance than the rough.

It can be shown that the Henderson (1924) smoothing spline method based on solving (5) satisfies all of the above tests with the exception of (12) and (13) while the Whittaker (1923) method satisfies all of the tests except (13).[20] Thus the above tests possess some discriminating power.

Our final test is due to Sprague. Before we formally define his test, it is worth quoting him in some detail to indicate his pioneering contributions:

> "I now proceed to a different part of my subject, and will prove that it is undesirable to employ such formulas as Mr. Woolhouse's, Mr. Higham's, or Mr. Ansell's, not only because, as already mentioned, they will never entirely get rid of the irregularities in our observations, but also because they all have a tendency to introduce an error even into a regular series of numbers. I start with the proposition which I think will command universal assent, that, if we attempt to graduate a perfectly regular series of numbers, the result should be to leave it unaltered; and that, if our method of procedure alters the law of the series, and

> substitutes for the original series one following a different law, this proves that our method of procedure is faulty".
>
> [Sprague (1887; 107-108)]

As the above quotation indicates, Sprague made two fundamental points: (i) the smooths that are generated by moving average smoothers do not appear to be visually smooth (too many minor wiggles appear in the smooth)[21] and (ii) the smooth that results from smoothing a perfectly smooth rough should be identical to the rough. We formalize Sprague's second point as the following smoothing invariance test:

(15) $s \quad f(y)$ implies $s \quad f(s)$.

If the smoothing correspondence f is actually a function, then (15) becomes $f[f(y)] = f(y)$.

Test (15) is very stringent; most smoothing methods do not pass this test. For example, of the seven main types of nonparametric smoothing models listed by Buja, Hastie and Tibshirani (1989): (i) running mean smoothers; (ii) bin smoothers; (iii) running line smoothers; (iv) polynomial regression; (v) cubic smoothing splines (the Henderson (1924) model); (vi) regression splines with fixed knots or break points; (vii) kernel smoothers, only methods (iv) and (vi) pass test (15).

Another way of evaluating smoothing methods is to look at the set of smooth points (recall definition (6)) that each method admits. For method (i) listed in the previous paragraph, only constant points of the form $y = k1_N$ will be recognized as smooth points. For methods (iii) and (v), only "linear" points $y = (y_1, \ldots, y_N)$ of the form $y_t = \alpha + \beta t$ will be recognized as being smooth points and only "quadratic" points of the form $y_t = \alpha + \beta t + \gamma t^2$ will be smooth points for Whittaker's (1923) smoothing spline method.

Buja, Hastie and Tibshirani (1989) consider linear smoothing models where the smoothing correspondence f(y) reduces to the linear function Ay of the rough y where A is a square matrix whose elements do not depend on y. For these models, they show that the set of smooth points is simply the set of eigenvectors of A which have a unit eigenvalue; i.e., it is the set of points y such that

(16) $y = Ay$.

Consider a linear regression smoother where the N by K matrix of independent variables is X. The smooth in this case is

(17) $s = X(X^TX)^{-1}X^Ty \equiv Ay.$

Thus in the case of a linear regression smoothing model, the A matrix is simply the projection matrix $X(X^TX)^{-1}X^T$ spanned by the columns of X and the set of smooth points is the K dimensional subspace of points spanned by the columns of X.

Of the smoothing methods listed by Buja, Hastie and Tibshirani, only methods (iv) and (vi) will admit a reasonably large class of smooth points.

In the context of linear smoothing models, the smoothing invariance test (15) can be given a more statistical justification as follows. Let $f(y) = Ay = s$. Then if (15) holds $f(y - s) = f(y) - f(s) = s - s = 0_N$. Thus if our smoothing method has systematically removed the trend from the original rough vector y, then y - s should be equal to a vector of random variables with zero expectations and hence the smooth of y - s should be 0_N.

Having considered alternative definitions of smoothness and a few "reasonable" axiomatic properties which can be used to evaluate alternative smoothing methods, we now turn to a description of our proposed smoothing method.

4. A Single Turning Point Smoothing Procedure

Before we define our smoothing method, we require some preliminary definitions.

Define the linear extension of the series $(y_1, \ldots, y_N)$ as the function y(t) of the continuous variable t such that $y_t = y(t)$ for t = 1, 2, . . ., N and for t between the integers i and i + 1, $y(t) \equiv y_i + (y_{i+1} - y_i)(t - i)$ for i = 1, 2, . . ., N - 1. Thus for t between the integers i and i + 1, the linear extension function y(t) simply travels along the straight line joining the points (i, y_i) and $(i + 1, y_{i+1})$.

Suppose that the second differences $\Delta^2 y_t$ that correspond to a given series $(y_1, \ldots, y_N)$ are all nonnegative. Then it is easy to prove that the corresponding linear extension y(t) is a convex function over the interval [1, N].[22] Suppose now that

(18) $\Delta^2 y_t \geq 0$ for $t = 3, 4, \ldots, t_1$ and

(19) $\Delta^2 y_t \leq 0$ for $t = t_1 + 1, t_1 + 2, \ldots, N$.

Then the linear extension y(t) is a convex function over the interval $[1, t_1]$ and is a concave function over the interval $[t_1 - 1, N]$. Thus between $t_1 - 1$ and t_1, y(t) changes its curvature and thus we say that the original discrete series y^t has a point of inflection or turning point between t_1-1 and t_1. The above line of reasoning is explicitly laid out in Sprague (1887; 94-95).[23]

Sprague (1887; 96) also observed than an "irregular" curve would have a large number of points of inflection or equivalently, there would be a large number of changes in the signs of the second differences $\Delta^2 y_t$ as t traveled from 3 to N. This is equivalent to the existence of a large number of regions where the linear extension y(t) of $(y_1, \ldots, y_N)$ alternates between being a convex and concave function. Thus, as in section 2, we follow Sprague in defining $(y_1, \ldots, y_N)$ to be a smooth series if its second differences, $\Delta^2 y_t$ for $t = 3, 4, \ldots, N$, do not change sign "too often".

In order to be more precise, let $N \geq 4$ and let k be any integer such that $1 \leq k \leq N - 3$. Then define the series $(y_1, \ldots, y_N)$ to be k smooth[24] if and only if $\Delta^2 y_t$ changes sign at most k - 1 times. Thus if a series is k smooth, it can exhibit at most k distinct regions of convexity and concavity, or put in other words, it can have at most k bumps and dips[25] or k hills and valleys. If a series is 1 smooth, then its linear extension is either a convex or a concave function over [1, N] (or both in which case it is a linear function).

We now consider the problem of mathematically representing k smooth series or curves. Specifically, consider a situation where we want our smoothed curves $\equiv (s_1, s_2, \ldots, s_N)$ to be convex over observations 1, 2, . . ., i (where i satisfies $3 \leq i \leq N - 1$) and then concave over observations i - 1, i, i + 1, . . ., N. Necessary and sufficient conditions for the linear extension s(t) of the discrete series s_t to be convex over the interval [1, i] and concave over the interval [i - 1, N] are:

(20) $s_t - 2s_{t-1} + s_{t-2} = \alpha_t^2$ for $t = 3, 4, \ldots, i$;

(21) $s_t - 2s_{t-1} + s_{t-2} = -\alpha_t^2$ for for $t = i + 1, i + 2, \ldots, N$,

where $\alpha_3, \alpha_4, \ldots, \alpha_N$ are scalar parameters. Now use the N - 2 equations in (20) and (21) to solve for $s_3, s_4, \ldots, s_N$ in terms of $s_1, s_2, \alpha_3, \alpha_4, \ldots, \alpha_N$:

$$
\begin{aligned}
s_1 &= s_1 \\
s_2 &= s_2 \\
s_3 &= -s_1 + 2s_2 + \alpha_3^2 \\
s_4 &= -2s_1 + 3s_2 + 2\alpha_3^2 + \alpha_4^2 \\
s_5 &= -3s_1 + 4s_2 + 3\alpha_3^2 + 2\alpha_4^2 + \alpha_5^2 \\
&\vdots \\
s_i &= -(i-2)s_1 + (i-1)s_2 + (i-2)\,\alpha_3^2 + (i-3)\,\alpha_4^2 + \ldots + \alpha_i^2 \\
s_{i+1} &= -(i-1)s_1 + is_2 + (i-1)\,\alpha_3^2 + (i-2)\,\alpha_4^2 + \ldots + 2\alpha_i^2 - \alpha_{i+1}^2 \\
&\vdots \\
s_N &= -(N-2)s_1 + (N-1)\,s_2 + (N-2)\,\alpha_3^2 + (N-1)\,\alpha_4^2 + \ldots + (N-i+1)\alpha_i^2 \\
&\quad - (N-i)\,\alpha_{i+1}^2 - (N-i-1)\,\alpha_{i+2}^2 - \ldots - 2\,\alpha_{N-1}^2 - \alpha_N^2 .
\end{aligned}
\tag{22}
$$

It is convenient to rewrite (22) using vector notation. Define s to be the N dimensional column vector of the s_t, 1_N to be a vector of ones, e_1 to be the first unit vector and define $\tau^{(t)}$ to be an N dimensional column vector which has zeros in its first t - 1 components and then the remaining components go through the first N - (t - 1) integers starting at 1 for t = 2, 3, . . ., N. With these definitions, equations (22) may be rewritten as follows:

$$
s = [e_1 - \tau^{(3)}]s_1 + \tau^{(2)}s_2 + \Sigma_{t=3}^{i}\,\tau^{(t)}\alpha_t^2 - \Sigma_{t=i+1}^{N}\,\tau^{(t)}\alpha_t^2 . \tag{23}
$$

Finally, we reparameterize s_1, s_2 and the α_t as follows:

$$
\begin{aligned}
& s_1 \equiv \beta_1;\ s_2 \equiv \beta_1 + \beta_2;\ \alpha_t^2 \equiv \beta_t \quad \text{for} \quad t = 3, 4, \ldots, i; \\
& -\alpha_t^2 \equiv \beta_t \quad \text{for} \quad t = i+1, i+2, \ldots, N.
\end{aligned}
\tag{24}
$$

If we substitute (24) into (23) and make use of the identity $e_1 + \tau^{(2)} - \tau^{(3)} = 1_N$, then (23) becomes

$$
s = 1_N\beta_1 + \Sigma_{t=2}^{N}\,\tau^{(t)}\beta_t \equiv X\beta \tag{25}
$$

where $X \equiv [1_N, \tau^{(2)}, \ldots, \tau^{(N)}]$, $\beta \equiv [\beta_1, \beta_2, \ldots, \beta_N]^T$ and the components of the β vector satisfy the following restrictions:

(26) β_1 and β_2 are unrestricted; $\beta_t \geq 0$ for t = 3, 4, . . ., i and $\beta_t \leq 0$ for t = i + 1, i + 2, . . ., N.

Thus the general representation of a discrete series $\{s_t$: t = 1, 2, . . ., N], whose linear extension is a convex function of over the interval [1, i] and is a concave function over the interval [i - 1, N], is given by (25) and (26).

Now we can return to the problem of smoothing a given series $y \equiv [y_1, y_2, \ldots, y_N]^T$. Suppose that we want the smoothed series $s \equiv [s_1, s_2, \ldots, s_N]^T$ to be convex over the first i observations and concave over observations i - 1, i, . . ., N. Then our s must have the form given by (25) and (26). It is natural to pick s to be the member of this set of points which is closest in some sense to the given y. We choose to measure closeness by the sum of squared residuals since this choice of metric is consistent with maximum likelihood estimation of the parameters under certain conditions. Thus our smoothing function f in the present situation is defined as $\hat{s} = f(y) \equiv X\hat{\beta}$ where $\hat{\beta}$ is the unique[26] solution to the following constrained least squares minimization problem:

(27) $$\min_\beta\{(y - X\beta)^T(y - X\beta): \beta \equiv [\beta_1, \ldots, \beta_N]^T, \\ \beta_t \geq 0 \text{ for } t = 3, 4, \ldots, i;\ \beta_t \leq 0 \text{ for } t = i + 1, \ldots, N\}$$

where the N by N matrix X was defined below (25).

The Kuhn-Tucker conditions which characterize the optimal solution $\hat{\beta}$ to (27) are:

(28) $$- X^T y + X^T X\hat{\beta} + u = 0_N \quad \text{where} \quad u \equiv [u_1, u_2, \ldots, u_N]^T;$$

(29) $$\hat{\beta}^T u = 0$$

(30) $$u_1 = 0$$

(31) $$u_2 = 0$$

(32) $$\hat{\beta}_t \geq 0 \quad \text{and} \quad u_t \leq 0 \quad \text{for } t = 3, 4, \ldots, i;$$

(33) $$\hat{\beta}_t \leq 0 \quad \text{and} \quad u_t \geq 0 \quad \text{for } t = i + 1, i + 2, \ldots, N.$$

These conditions can be used to establish some properties of the smoothed values, $\hat{s} = X\hat{\beta}$.

If we replace $X\hat{\beta}$ in (28) by $\hat{s}$ and look at the resulting first equation, then making use of the definition of X and (30), we have

(34) $1_N \cdot y = 1_N \cdot \hat{s}$.

Thus the sum of the smoothed values, $\Sigma_{t=1}^{N} \hat{s}_t$, is equal to the sum of the unsmoothed values, $\Sigma_{t=1}^{N} y_t$, and hence our smoother f satisfies the sum preserving test (8). Geometrically, this means that the areas under the curves traced out by the linear extensions of the two series are equal.

Now look at the second equation in (28), replacing $X\hat{\beta}$ by $\hat{s}$. Using (31) and the definition of X, we have, after a bit of rearrangement:

(35) $\tau^{(2)} \cdot y = \tau^{(2)} \cdot \hat{s}$ or $\Sigma_{t=2}^{N}(t - 1)\, y_t = \Sigma_{t=2}^{N}(t - 1)\, \hat{s}_t$.

Using (34), it can be seen that (35) implies

(36) $\Sigma_{t=1}^{N}\, ty_t = \Sigma_{t=1}^{N}\, t\hat{s}_t$

and hence our smoothing method also satisfies the first moment preserving test, (9).

Finally, premultiply both sides of (28) by $\hat{\beta}^T$. Using $\hat{s} = X\hat{\beta}$ and (29), the resulting equation simplifies to:

(37) $\hat{s} \cdot \hat{s} = \hat{s} \cdot y$

Since the right hand side of (37) is equal to or less than $(\hat{s} \cdot \hat{s})^{1/2}(y \cdot y)^{1/2}$ by the Cauchy-Schwarz inequality, we obtain the following inequality:

$$\begin{aligned} \hat{s} \cdot \hat{s} &\le (\hat{s} \cdot \hat{s})^{1/2}(y \cdot y)^{1/2} \quad \text{or} \\ (\hat{s} \cdot \hat{s})^{1/2} &\le (y \cdot y)^{1/2} \quad \text{or} \\ \text{(38)} \quad \hat{s} \cdot \hat{s} &\le y \cdot y \end{aligned}$$

Hence our smoothing method satisfies the diminishing variation test (14).

The reader can verify that the above smoother also satisfies the identity test (10), the linear trend test (11) and the smoothing invariance test (15).

It should be noted that the nonlinear regression problem (27) is very closely related to the nonparametric regression models pioneered by Hildreth

(1954) and Wagner (1962). Our model more or less specializes to Hildreth's concave regression problem if we change the constraints in (32) to $\beta_t \leq 0$ for $t = 3, 4, \ldots, i$. Our problem (27) is included in the general class of models considered briefly by Wagner (1962; 576-577).[27]

An efficient algorithm to solve problems of the form (27) is outlined in Appendix 1.

Restricted least squares minimization problems of the form (27) are the basic building blocks that we use to construct our smoothing method which we shall describe in the following section.

5. Flexible Smoothing Procedures

Recall the definition of a k smooth series at the beginning of the previous section. Let k be a positive integer. We say that a smoothing method f is flexible of order k if it possesses the following property:

(39) y is k smooth implies $y \quad f(y)$;

i.e., if $y \equiv (y_1, \ldots, y_N)$ is a k smooth rough, then y also belongs to the set of smoothed series generated by the smoothing method f.

Consider now the problem of finding a flexible of order 1 smoothing procedure. If $y \equiv (y_1, \ldots, y_N)$ is 1 smooth, then $\Delta^2 y_t$ has 0 sign changes; i.e., either $\Delta^2 y_t \geq 0$ for all t or $\Delta^2 y_t \leq 0$ for all t ($\Delta^2 y_t$ attaining 0 does not count as a sign change). In order to define a flexible of order 1 smoothing procedure, we need to solve the following two constrained least squares minimization problems which are similar to (27):

(40) $\min_\beta\{(y - X\beta)^T(y - X\beta)\colon\ \beta \equiv [\beta_1, \ldots, \beta_N]^T;\ \beta_t \geq 0 \text{ for } t = 3, 4, \ldots, N\}$;

(41) $\min_\beta\{(y - X\beta)^T(y - X\beta)\colon\ \beta \equiv [\beta_1, \ldots, \beta_N]^T;\ \beta_t \leq 0 \text{ for } t = 3, 4, \ldots, N\}$.

where X was defined below (25).

Let $\hat\beta^1$ and $\hat\beta^2$ denote the unique solutions to (40) and (41) respectively and define

(42) $s^1 \equiv X\hat\beta^1 \quad ; \quad s^2 \equiv X\hat\beta^2$.

If the minimum for (40) is less than the minimum for (41), define the smoothed vector to be $s^1 \equiv f(y)$; if the minimum for (41) is less than the minimum for (40), define $s^2 \equiv f(y)$; if the minima are equal, define the set of smoothed vectors to be $\{s^1, s^2\} \equiv f(y)$. Thus in the empirically rare case where the minima are equal, our smoothing procedure cannot distinguish between s^1 and s^2 and gives both vectors as solutions to the smoothing problem.

It is easy to see that the above smoothing procedure satisfies test (39) for $k = 1$; i.e., if the input vector y into our suggested smoothing procedure is 1 smooth (or Sprague smooth of order 1), then the smoothed vector will coincide with the input vector. Thus the above smoothing method is flexible of order 1. It also can be verified that our suggested flexible of order 1 smoothing procedure satisfies tests (8) - (12) and (14); i.e., the only test that the method fails is test (13), the cubic trend test.[28]

We now consider the problem of finding a flexible of order 2 smoothing procedure; i.e., we want to find a smoothing procedure that will generate the input vector y as the vector of smoothed values if y is 2 smooth (or Sprague smooth of order 2) so that the linear extension of y has at most one turning point or inflection point.[29]

In addition to the two least squares minimization problems (40) and (41) (which will allow for the possibility that the input vector y has no turning points), we need to solve minimization problems of the form (27) for all positive choices of the turning point i. Problems of the form (27) will reproduce an input curve which is initially convex up to observation i and then is concave. We require $N \geq 4$ and $3 \leq i \leq N - 1$. Thus there are $N - 3$ possible choices for the turning point i. Denote the β solution to (27) for $i = 3, 4, \ldots, N - 1$ by $\hat{\beta}^i$ and define $s^i \equiv X\hat{\beta}^i$ for $i = 3, 4, \ldots, N - 1$. The vectors s^1 and s^2 are still defined by (42).

Now we need to consider least squares minimization problems which will reproduce input curves which are initially concave (up to observation i) and then are convex. Thus consider the following $N - 3$ minimization problems for $i = 3, 4, \ldots, N - 1$:

$$(43) \quad \min_\beta\{(y - X\beta)^T(y - X\beta): \ \beta \equiv [\beta_1, \ldots, \beta_N]^T; \beta_t \leq 0 \text{ for } t = 3, 4, \ldots, i; \beta_t \geq 0 \text{ for } i + 1, i + 2, \ldots, N\}$$

where as usual X was defined below (25). Denote the solution to (43) by β^{N-3+i} and define $s^{N-3+i} \equiv X\hat{\beta}^{N-3+i}$ for i = 3, 4, . . ., N - 1. Thus in all, there are 2 + N - 3 + N - 3 = 2N - 4 constrained minimization problems to be solved and 2N - 4 candidate smoothed vectors s^i, i = 1, . . ., 2N - 4, are generated. Now find the minimum of the 2N - 4 minimized objective functions. We define our solution set of smoothed vectors f(y) to be the set of s^i which correspond to the set of minimization problems i which attained the overall minimized objective function. In most practical applications of our suggested smoothing technique, the overall minimum will be attained by only one of the 2N - 4 minimization problems and under these conditions, the optimal smoothed vector will be unique.

It can be seen that the above method is a flexible of order 2 smoothing procedure; i.e., if the input vector y has at most one turning point, then this vector will emerge unaltered as a smoothed vector. Moreover, it can be verified that our suggested flexible of order 2 smoothing procedure satisfies all of the tests (8) - (15) listed in section 3 above.[30] Thus this smoothing method has attractive axiomatic properties.

By now, the reader should be able to see how a flexible of order 3 smoothing procedure could be constructed. We need to solve the 2N - 4 minimization problems that occurred in the flexible of order 2 procedure plus the following two sequences of minimization problems for all possible choices of i and j where $N \geq 5$ and $3 \leq i < j \leq N - 1$:

$$(44) \quad \min_\beta\{(y - X\beta)^T(y - X\beta): \beta \equiv [\beta_1, . . ., \beta_N]^T; \beta_t \leq 0 \text{ for } t = 3, 4, . . ., i; \beta_t \geq 0; \text{ for } t = i + 1, i + 2, . . ., j; \beta_t \leq 0 \text{ for } t = j + 1, . . ., N\};$$

$$(45) \quad \min_\beta\{(y - X\beta)^T(y - X\beta): \beta \equiv [\beta_1, . . ., \beta_N]^T; \beta_t \geq 0 \text{ for } t = 3, 4, . . ., i; \beta_t \leq 0; \text{ for } t = i + 1, i + 2, . . ., j; \beta_t \geq 0 \text{ for } t = j + 1, . . ., N\}.$$

Thus in theory, in order to generate a flexible of order 3 smoothing procedure, we would have to solve 2N - 4 + (N - 3)(N - 4)/2 + (N - 3)(N - 4)/2 = 2N - 4 + (N - 3) (N - 4) constrained least squares minimization problems. Fortunately, in practice, it will not be necessary to solve all of these problems, if the investigator resorts to graphical or visual plotting techniques. The investigator need only plot the data and draw what appears to be the best fitting curve through the data where the curve has only two inflection points. If the curve starts with a hill, then the location of the two inflection points will determine the integers i and j which appear in (44) and the resulting restricted least squares problem can be solved. Solve additional restricted least squares regressions of the form (44) where the choices of the turning points i and j are chosen to be close to the initial choices. Then the minimum of the minimized objective functions for these regressions can be taken and the resulting $s = X\hat{\beta}$ is the flexible of order 3 smooth. If the graphed curve starts with a valley, then a few restricted least squares problems of the form (45) will be solved in order to generate the flexible of order 3 smooth. Thus our suggested smoothing procedure can be viewed as a scientific[31] implementation of Sprague's graphical method, and moreover, our method satisfies all of the tests for smoothing methods that Sprague originally proposed.[32]

It should be evident how the above procedure can be generalized to define a flexible of order k smoothing method.

Note that the choice of k, the maximum number of distinct regions of curvature or the maximum number of hills and valleys that we allow an output smooth to have, can be regarded as a smoothing parameter[33] for our smoothing method: the smaller k is, the more smooth we regard the output smooth as being, according to Sprague's definition of smoothness.

How should k be chosen in empirical applications of our method? Either from a priori scientific considerations or by visually examining a plot of the rough, the investigator should determine a maximal k that he or she is willing to accept. Having specified this maximal number, k* say, we may want to know whether a lower k smooth can satisfactorily explain the nonrandom part of our data. More specifically, suppose k* = 3 is specified and the best fitting restricted least squares regression turns out to be of the form (44) for some i*, j* such that $3 < i^* < j^* < N - 1$. The resulting smooth is concave up to i*, convex over the interval between i* and j* and then concave over the interval between j* and N. Thus a Sprague smooth of order 3 is used to explain the systematic part of the rough. We now ask whether a Sprague smooth of order 1 is adequate to explain the systematic part of the data. To answer this question, we solve the restricted least squares problem (41) which eliminates the convex portion of the initial smooth between i* and j*. Comparing the estimated $\hat{\beta}_t^k$ coefficients for the two regressions indexed by k = 1 and 2, we will find that the second regression has more $\hat{\beta}_t^2$ set equal to zero than was the case for the initial $\hat{\beta}_t^1$, since in the second regression, the smooth will jump over the valley between i* and j* in a linear fashion. Thus an approximate test for the null hypothesis that the valley can be eliminated can be obtained by performing a classical F test, treating the initial smooth of order 3 model as a linear regression model in the nonzero $\hat{\beta}_t$, and testing whether the appropriate subset of these coefficients is significantly different from zero. We shall illustrate this approximate[34] testing methodology in section 7 below.

Another method for determining the appropriate number of hills and valleys for our output smooth can be obtained by using the nonparametric runs test of Wald and Wolfowitz (1940; 151). Suppose that the rough is y and that our best restricted least squares smooth s has been determined, allowing for k distinct regions of curvature. Should we recalculate our smooth, adding another bump or dip; i.e., should we increase k to k + 2? To answer this question, calculate the residuals $e_i = y_i - s_i$ for i = 1, . . . , N. If the addition of an extra hill (valley) is warranted, there should be a substantial run of positive (negative) residuals. In either case, the number of runs in the residual series should be lower than would be the case with random residuals. Thus if the test statistic is sufficiently low, we add another hill or valley. This method for determining k is also illustrated in section 7.[35] The difficulty with the use of the Wald and Wolfowitz Runs Test is that it has rather low power against many alternatives.

A final method that could be used to determine k is the use of cross validation; i.e., k is chosen to be that model which generates the smallest sum of squared prediction errors, dropping one observation at a time.[36] We did not pursue this method in the present paper.

Our suggested flexible of order k smoothing procedure can be viewed as a rather complicated method for selecting a linear regression model. Once the model has been chosen, we can simply assume that the β_t that correspond to the zero $\hat{\beta}_t$ are zero as well and we obtain a classical linear regression model and the usual procedures can be used in order to obtain confidence intervals for the estimated smooth, $\hat{s} = X\hat{\beta}$ say. However, these confidence intervals are only approximately valid.[37]

Before we illustrate our suggested smoothing method empirically, we indicate in the next section how our method can be modified to deal with end effects.

6. End Point Problems

Simulation experiments as well as theoretical considerations have shown that many nonparametric smoothing methods are less accurate in fitting the true smooth at the end points of the data set.[38] Unfortunately, our suggested smoothing method is particularly subject to end point bias. Since we are fitting a concave or convex function near the end points of our data set, our method will tend to add the first and last errors to the systematic parts, s_1 and s_N, about 50% of the time, assuming symmetrically distributed errors, e_1 and e_N. The end point bias problem will be particularly troublesome if the ratio of noise to signal is high; i.e., if the variance of the e_t is large.

Our solution to the end point bias problem is quite conventional: we simply force our estimated smooth to be linear near the end points.[39] To force linearity over three observations at each end, we add the following two restrictions to the constrained least squares minimization problems that were described in the previous two sections:

(46) $\beta_3 = 0; \beta_N = 0.$

To force linearity over i+2 observations at each end, add the following 2i restrictions on the β_t:

(47) $\beta_{2+1} = 0, \ldots, \beta_{2+i} = 0; \beta_{N-i+1} = 0; \beta_{N-i+2} = 0, \ldots, \beta_N = 0.$

At this point, we are unable to suggest any formal rules on how large i should be for any given empirical application. For our empirical example to be described in the next section, the error variance seemed to be small and so we chose i to be zero. For the sawtooth function example to be described in section 8 below, the error variance was moderate and we chose i to be 3. Obviously, there is room for further research on this problem.

Although imposing the restrictions (46) or (47) on our smoothing method will reduce the variance of our estimated smooth near the endpoints, this reduction in variance will not be purchased without a cost. The cost is that our endpoint modified smoothing method will no longer satisfy tests (12), (13) and (39). Put another way, if our true smooth is sufficiently curvilinear near the endpoints, then our modified smoothing method will incorrectly force the estimated smooth to be linear near the endpoints. Thus our modified smoothing method will no longer automatically pass through such smooth functions as quadratics, cubics and general concave and convex functions—only subclasses of these functions (which are suitably linear near the end points) will pass through our endpoint modified method unchanged.

We turn now to our empirical examples.

7. The Smoothing of Mortality Data

In the first empirical illustration of our method, the underlying series to be smoothed consists of U.S. mortality data, which have the property that the series is reasonably smooth initially. We use our algorithm to further smooth this series, and we test for the appropriate number of break points.

The mortality data are drawn from Table 3.2, Life Tables for the Total Population: United States, 1979-81, that appear in Bowers, Gerber, Hickman, Jones and Nesbitt (1986). These data are based on estimates of probabilities of death given survival to various ages, derived from the experience of the entire U.S. population in the years around the 1980 Census. The data consist of information on the number of individuals living at the beginning of each age interval and the number dying during the interval, for age intervals 0-1 though 109-110 years. We define the mortality rate for the age interval t-1 to t as the number dying during that age interval divided by the number alive at the beginning of that age interval, and denote it as the mortality rate at age t. These mortality rates are presented in Table 1—the mortality rate falls until about age 11 or 12 then rises monotonically until age 110. The rates are plotted in Figure 1 (the solid line) together with the predicted mortality rates obtained by applying our new procedure to these data assuming a convex curve throughout (the dashed line). In applying our procedure we minimize the weighted sum of squared residuals (SSR), where a residual is defined as the difference between the actual and predicted mortality rate for a given age, and the weight is the square root of the sample size, that is, the number at risk for that age.[40]

From Figure 1 it is clear that the smoothed rates are close to the actual rates except for large discrepancies that occur in the last ten to fifteen years, and for smaller differences that occur in approximately the 15-25 year range. Since the mortality function is generally convex, allowing for concave segments in the interior of the age range requires the introduction of an even number of break points. Of course allowing for a concave segment at the beginning or end of the range requires only a single break point. Thus we apply our procedure allowing for the possibility of up to four break points, two of which we expect to occur in the mid-teens to mid-twenties range, and one or two of which we expect to occur near the end of the sample age range. As in the theoretical discussion above, a break point is defined as a point at which the second difference of the mortality function changes sign. For each of these cases (that is, conditional on the number of breaks), we select the break points by minimizing the weighted SSR using a grid search procedure.[41]

Table 2 contains the weighted SSR values together with the break points, the number of nonzero parameters and a goodness of fit measure, for models with zero up to four break points. As would be expected from the evidence of Figure 1, the weighted SSR declines sharply when allowance is made for a concave segment near the end of the age range, but falls only slightly more when an additional break point is added. However, with three break points, the concave section in the teens and twenties age range is modelled more appropriately, and the weighted SSR drops sharply again compared with the case of only one or two breaks. Finally, introduction of the fourth break point allows for the possibility of a convex segment in the last few years.

Conventional F tests may be used to statistically compare some of the entries in Table 2, and thus test the significance of adding break points. Of course such F tests are only valid for nested models, which in our case means that the variables (ages) in the model with the smaller number of break points must be a subset of the variables in the model with the larger number of break points. Although information on the included variables is not provided in Table 2, it is the case that, for results reported there, all models with fewer breaks are nested in those with more breaks, except for the model with two break points, which is not nested in the one with three breaks.

We begin with the most general model that we estimate (four breaks) and ask if restrictions imposed by reducing the number of breaks results in a significant reduction in the weighted SSR. Without presenting detailed calculations, we can report that in all cases reducing the number of break points from four results in a significant reduction. For example, a change from four to three breaks yields an F statistic of 14.0, while the critical F at the one percent level is 6.1, and a change from four to two breaks yields an F of 148.3, with the critical level being 3.3. Indeed all nested comparisons in Table 1 yield highly significant F values except for the change from two breaks to one break, where the F values is .39 and the critical five percent level is 4.1.

Figure 2 contains a plot of the actual mortality data (solid line) together with the values predicted by our four break model (dashed line). Clearly the fit is extremely good; indeed for all intents and purposes the two lines appear to coincide. Although not shown here the divergence between actual and predicted mortality rates is zero (to four decimals) for all ages between 1 and 100 years. This is consistent with the R^2 value of unity (to five decimals) that appears in Table 2. We conclude that there is little to be gained by introducing more break points; and that our model with four break points, and thus two concave segments in an otherwise convex function, adequately represents a "smoothed" mortality function for these data.

As mentioned in section 5, an alternative analysis that can be used to shed some light on the question of the appropriate number of break points involves the study of runs in the residuals. In columns 2 through 5 of Table 3 we present the basic information, obtained from our estimations, that can be used to calculate the runs statistic given in column 6. Since this statistic is asymptotically distributed as a standard normal, we conclude that in the case of no breaks one break, or two breaks, there are significantly fewer runs than would be expected with random residuals. This suggests that we add at least one additional hill or valley. For three or four breaks there are significantly more runs than would be expected. This suggests that it would be inappropriate to add another hill or valley to the three or four break models. Putting these results together and drawing on our F criterion results, leads us to prefer the model with 4 break points.

8. <u>Fitting a Sawtooth Function</u>

An attractive feature of our new procedure is that it should prove useful for smoothing even when the data to be smoothed are generated by an underlying function that contains discontinuities. We use the sawtooth function studied by McDonald and Owen (1986) and Breiman and Peters (1992) to illustrate this point. These authors define a function s(t) over the (0,1] interval as follows:

$$(48) \qquad s(t) \equiv \begin{cases} 2t & \text{for} \quad 0 < t \le 1/2 \\ 2t - 1 & \text{for} \quad 1/2 < t \le 1. \end{cases}$$

This simple sawtooth function consists of two line segments essentially rising from 0 to 1, with a discontinuity at $t = 1/2$. To introduce a nonstochastic element we select 100 equispaced points in the interval [.01,1], and at each point add a random normal disturbance e_t, which has mean zero and standard deviation one half the standard deviation of s.[42] This gives a sample of 100 observations on the series y_t, defined as $y_t = s_t + e_t$:

$$(49) \qquad y_t = \begin{cases} 2t + e_t & , \quad 0 < t \le 1/2; \\ 2t - 1 + e_t & , \quad 1/2 < t \le 1. \end{cases}$$

We smooth the y_t using our new procedure assuming a concave segment for the first half of the sample, a single break point, and then a convex segment for the remainder. It may be noted that if we were fitting the nonstochastic series $\{s_t\}$ rather than $\{y_t\}$, then this smoothing procedure would result in a perfect fit, that is, it would replicate the $\{s_t\}$ series exactly, as is appropriate.

To summarize the performance of our smoothing procedure in a situation involving a discontinuity of this type, we carry out a simple Monte Carlo experiment. The experiment consists of 500 trials in each of which we generated, for the 100 fixed t values, a new series of e_t values, and hence of y_t values, which we smoothed using our new approach. In each case the single breakpoint occurred at t = .5. For each trial we recorded the R squared value, defined as the fraction of the variance of y explained by the smoothed series, as well as the number of parameters estimated as part of the smoothing algorithm.

In Figures 3 and 4 we plot the means of the smoothed y_t values (solid lines), and the means of the unsmoothed y_t values (dashed lines). The difference between the two figures is that in Figure 3 no endpoint corrections have been made, whereas in Figure 4, we used our endpoint modified smoothing method as explained in section 6 where the first three smoothed observations are restricted to lie on a straight line, as are the last three. The effects of these restrictions are evident from the figures; the smoothed functions are almost identical, except at the beginning and end of the sample range. Consistent with is this is the finding that the R squared values averaged over the 500 trials are almost the same in the two cases, namely .815 and .814 for the unrestricted and restricted models respectively. The average number of nonzero estimated parameters is also very similar in the two cases, namely 8.4 and 7.9, and consequently in the discussion below we consider only the smoothed series in which the endpoint restrictions have been imposed.

In Figure 4 we compared the means of the smoothed and unsmoothed series, while in Figure 5 we compare the means of the smoothed series with the underlying nonstochastic function. The figures are of course very similar since the independent random disturbance used in generating the unsmoothed series has mean zero, and thus the pointwise means of the unsmoothed series converge to the nonstochastic function as the number of trials increases. In Figure 6 we present evidence on the variability as well as the mean of our smoothed series. For each t observation we calculate the standard deviation of the 500 corresponding smoothed y values. The solid line in Figure 6 reflects the means of the smoothed values, while the dotted lines represent two standard deviations on either side of these means, and thus may be interpreted as approximate 95 percent confidence bands for our procedure.

This same information is presented in Table 4, which contains in columns 1 through 5 respectively (and is continued in columns 6 through 10 respectively) the t values, the nonstochastic function values s_t, the means of the unsmoothed data, the means of the smoothed data, and the standard deviations of the smoothed data. The standard deviations appear to be uniformly small for most of the sample, with larger values occurring, as would be expected, at the beginning, middle and end of the t range. Although the standard deviations are largest around the point of discontinuity, they are still reasonably small. On the basis of Figures 4–6 and the information in Table 4 we conclude that our procedure performs well in smoothing this sawtooth function.

9. Conclusion

Most nonparametric smoothing methods produce rather wiggly output series. Using Sprague's definition of a smooth series or curve, our flexible of order k nonparameteric smoothing method produces output series which are limited to k separate regions of convexity and concavity. Thus our suggested smoothing method appears to be a reasonable formulation of Sprague's carefully crafted graphical smoothing method. Our proposed method also has very desirable axiomatic properties.

However, there are some difficulties associated with the use of our method: (i) our procedure may estimate a rather large number of nonzero parameters relative to the number of degrees of freedom; (ii) the series $s_t \equiv \cos \pi t$ for t = 1, 2, . . ., N is very smooth from some points of view but it is very rough using Sprague's definition of smoothness (basically that the number of changes in sign of its second differences be "small");[43] (iii) our suggested methods for finding confidence intervals and testing for the degree of flexibility are only approximate and subject to some sample selection bias; (iv) we have not subjected our method to a wide variety of Monte Carlo experiments; and (v) we have not explored very many of the implications of the axiomatic approach to smoothing.

Even though our proposed method of smoothing is subject to the above difficulties, we feel that its tremendous flexibility and good axiomatic properties will make it a useful method to try in many empirical applications. We also hope that our many references to the ancient actuarial literature will help to make researchers aware of its large influence on the modern nonparametric smoothing literature.

Table 1
U.S. Mortality Rates, 1979-1981

Age	Mortality	Age	Mortality
1	0.01260	56	0.00902
2	0.00093	57	0.00978
3	0.00065	58	0.01060
4	0.00050	59	0.01151
5	0.00041	60	0.01254
6	0.00037	61	0.01368
7	0.00034	62	0.01493
8	0.00030	63	0.01628
9	0.00026	64	0.01768
10	0.00023	65	0.01911
11	0.00019	66	0.02058
12	0.00019	67	0.02217
13	0.00024	68	0.02389
14	0.00038	69	0.02586
15	0.00053	70	0.02806
16	0.00068	71	0.03052
17	0.00084	72	0.03314
18	0.00096	73	0.03594
19	0.00104	74	0.03882
20	0.00112	75	0.04184
21	0.00121	76	0.04507
22	0.00127	77	0.04867
23	0.00132	78	0.05273
24	0.00134	79	0.05743
25	0.00134	80	0.06275
26	0.00132	81	0.06883
27	0.00130	82	0.07551
28	0.00130	83	0.08278
29	0.00130	84	0.09042
30	0.00131	85	0.09841
31	0.00132	86	0.10726
32	0.00135	87	0.11710
33	0.00137	88	0.12719
34	0.00143	89	0.13709
35	0.00149	90	0.14725
36	0.00160	91	0.15868
37	0.00170	92	0.17173
38	0.00183	93	0.18564
39	0.00197	94	0.20020
40	0.00213	95	0.21498
41	0.00232	96	0.22982
42	0.00254	97	0.24331
43	0.00279	98	0.25655
44	0.00306	99	0.26865
45	0.00334	100	0.28035
46	0.00366	101	0.29130
47	0.00401	102	0.30061
48	0.00441	103	0.31053
49	0.00488	104	0.32061
50	0.00538	105	0.32959
51	0.00590	106	0.33520
52	0.00642	107	0.34454
53	0.00698	108	0.34615
54	0.00762	109	0.35294
55	0.00830	110	0.36364

Table 2
Goodness of Fit Statistics For U.S. Mortality Data

Number of Breaks	SSR (Weighted)	Break Points	Number of Nonzero Parameters	R^2 (Unweighted)
0	1018.34	--	63	.99468
1	92.59	97	76	.99998
2	91.52	97,109	77	.99999
3	1.72	16,28,97	91	.99999
4	.65	16,28,97,109	93	1.00000

Notes

1. Entries in column 2 have been multiplied by 1000.

2. The R^2 values are defined using unweighted data as one minus the variance of the residuals divided by the variance of the mortality rate.

Table 3
Runs Statistics for U.S. Mortality Data

Number of Breaks	N_1	N_2	n	E(n)	Runs Test Statistic
0	21	33	9	26.7	-5.11
1	23	28	17	26.3	-2.64
2	23	32	15	24.8	-2.88
3	15	19	24	17.8	2.20
4	15	17	24	16.9	2.55

Notes

1. N_1 = number of observations with a positive residual.

2. N_2 = number of observations with a negative residual.

3. n = number of runs.

4. E(n) = expected number of runs, given by $2N_1N_2/(N_1 + N_2) + 1$.

5. The statistic is defined as $(n - E(n))/\sigma_n$, where $\sigma_n^2 = \dfrac{2N_1N_2(2N_1N_2 - N_1 - N_2)}{(N_1 + N_2)^2(N_1 + N_2 - 1)}$, and is asymptotically normally distributed with mean zero and unit variance.

6. The difference between the total number of observations used in the estimations (110) and $N_1 + N_2$ represents the number of observations with a zero residual.

Table 4
Monte Carlo Results

0.01	**0.01**	**0.010**	**-0.014**	**0.0335**	**0.51**	**0.01**	**0.014**	**0.050**	**0.0468**
0.02	**0.02**	**0.017**	**0.001**	**0.0285**	**0.52**	**0.02**	**0.023**	**0.038**	**0.0293**
0.03	**0.03**	**0.022**	**0.015**	**0.0241**	**0.53**	**0.03**	**0.027**	**0.041**	**0.0247**
0.04	**0.04**	**0.046**	**0.030**	**0.0208**	**0.54**	**0.04**	**0.039**	**0.048**	**0.0223**
0.05	**0.05**	**0.047**	**0.044**	**0.0191**	**0.55**	**0.05**	**0.050**	**0.055**	**0.0209**
0.06	**0.06**	**0.058**	**0.056**	**0.0178**	**0.56**	**0.06**	**0.061**	**0.063**	**0.0199**
0.07	**0.07**	**0.071**	**0.068**	**0.0168**	**0.57**	**0.07**	**0.070**	**0.071**	**0.0191**
0.08	**0.08**	**0.083**	**0.079**	**0.0162**	**0.58**	**0.08**	**0.081**	**0.080**	**0.0184**
0.09	**0.09**	**0.093**	**0.090**	**0.0155**	**0.59**	**0.09**	**0.093**	**0.089**	**0.0178**
0.10	**0.10**	**0.103**	**0.101**	**0.0149**	**0.60**	**0.10**	**0.095**	**0.098**	**0.0172**
0.11	**0.11**	**0.108**	**0.112**	**0.0145**	**0.61**	**0.11**	**0.114**	**0.107**	**0.0166**
0.12	**0.12**	**0.116**	**0.122**	**0.0140**	**0.62**	**0.12**	**0.120**	**0.117**	**0.0161**
0.13	**0.13**	**0.129**	**0.133**	**0.0136**	**0.63**	**0.13**	**0.127**	**0.126**	**0.0156**
0.14	**0.14**	**0.135**	**0.143**	**0.0133**	**0.64**	**0.14**	**0.133**	**0.136**	**0.0151**
0.15	**0.15**	**0.150**	**0.154**	**0.0130**	**0.65**	**0.15**	**0.148**	**0.145**	**0.0146**
0.16	**0.16**	**0.164**	**0.164**	**0.0127**	**0.66**	**0.16**	**0.161**	**0.155**	**0.0142**
0.17	**0.17**	**0.174**	**0.175**	**0.0126**	**0.67**	**0.17**	**0.172**	**0.165**	**0.0138**
0.18	**0.18**	**0.180**	**0.185**	**0.0124**	**0.68**	**0.18**	**0.184**	**0.174**	**0.0135**
0.19	**0.19**	**0.191**	**0.195**	**0.0123**	**0.69**	**0.19**	**0.186**	**0.184**	**0.0131**
0.20	**0.20**	**0.199**	**0.206**	**0.0123**	**0.70**	**0.20**	**0.204**	**0.194**	**0.0128**
0.21	**0.21**	**0.209**	**0.216**	**0.0123**	**0.71**	**0.21**	**0.207**	**0.204**	**0.0125**
0.22	**0.22**	**0.218**	**0.226**	**0.0124**	**0.72**	**0.22**	**0.220**	**0.214**	**0.0123**
0.23	**0.23**	**0.232**	**0.236**	**0.0125**	**0.73**	**0.23**	**0.230**	**0.224**	**0.0122**
0.24	**0.24**	**0.245**	**0.246**	**0.0126**	**0.74**	**0.24**	**0.241**	**0.234**	**0.0121**
0.25	**0.25**	**0.248**	**0.256**	**0.0126**	**0.75**	**0.25**	**0.253**	**0.244**	**0.0120**
0.26	**0.26**	**0.261**	**0.266**	**0.0125**	**0.76**	**0.26**	**0.262**	**0.253**	**0.0121**
0.27	**0.27**	**0.267**	**0.276**	**0.0125**	**0.77**	**0.27**	**0.273**	**0.263**	**0.0121**
0.28	**0.28**	**0.279**	**0.286**	**0.0126**	**0.78**	**0.28**	**0.287**	**0.273**	**0.0123**
0.29	**0.29**	**0.288**	**0.296**	**0.0127**	**0.79**	**0.29**	**0.295**	**0.283**	**0.0124**
0.30	**0.30**	**0.300**	**0.306**	**0.0128**	**0.80**	**0.30**	**0.297**	**0.293**	**0.0127**
0.31	**0.31**	**0.313**	**0.315**	**0.0130**	**0.81**	**0.31**	**0.307**	**0.304**	**0.0129**
0.32	**0.32**	**0.320**	**0.325**	**0.0133**	**0.82**	**0.32**	**0.317**	**0.314**	**0.0132**
0.33	**0.33**	**0.327**	**0.335**	**0.0135**	**0.83**	**0.33**	**0.325**	**0.324**	**0.0135**
0.34	**0.34**	**0.338**	**0.345**	**0.0138**	**0.84**	**0.34**	**0.340**	**0.334**	**0.0139**
0.35	**0.35**	**0.349**	**0.354**	**0.0141**	**0.85**	**0.35**	**0.355**	**0.344**	**0.0143**
0.36	**0.36**	**0.358**	**0.364**	**0.0144**	**0.86**	**0.36**	**0.354**	**0.355**	**0.0146**
0.37	**0.37**	**0.370**	**0.373**	**0.0147**	**0.87**	**0.37**	**0.373**	**0.365**	**0.0150**
0.38	**0.38**	**0.380**	**0.383**	**0.0150**	**0.88**	**0.38**	**0.379**	**0.375**	**0.0154**
0.39	**0.39**	**0.394**	**0.392**	**0.0154**	**0.89**	**0.39**	**0.383**	**0.386**	**0.0158**
0.40	**0.40**	**0.390**	**0.401**	**0.0157**	**0.90**	**0.40**	**0.400**	**0.397**	**0.0163**
0.41	**0.41**	**0.413**	**0.411**	**0.0160**	**0.91**	**0.41**	**0.414**	**0.407**	**0.0167**
0.42	**0.42**	**0.413**	**0.420**	**0.0164**	**0.92**	**0.42**	**0.416**	**0.418**	**0.0172**
0.43	**0.43**	**0.428**	**0.429**	**0.0169**	**0.93**	**0.43**	**0.426**	**0.429**	**0.0176**
0.44	**0.44**	**0.441**	**0.438**	**0.0176**	**0.94**	**0.44**	**0.435**	**0.441**	**0.0183**
0.45	**0.45**	**0.449**	**0.446**	**0.0184**	**0.95**	**0.45**	**0.448**	**0.452**	**0.0190**
0.46	**0.46**	**0.456**	**0.454**	**0.0198**	**0.96**	**0.46**	**0.458**	**0.465**	**0.0200**
0.47	**0.47**	**0.471**	**0.462**	**0.0215**	**0.97**	**0.47**	**0.470**	**0.479**	**0.0290**
0.48	**0.48**	**0.480**	**0.469**	**0.2027**	**0.98**	**0.48**	**0.479**	**0.494**	**0.0234**
0.49	**0.49**	**0.489**	**0.473**	**0.0229**	**0.99**	**0.49**	**0.494**	**0.508**	**0.0270**
0.50	**0.50**	**0.501**	**0.461**	**0.0459**	**1.00**	**0.50**	**0.495**	**0.522**	**0.0314**

Columns

1,6 t values

2,7 nonstochastic function values, s_t

3,8 means of unsmoothed data

4,9 means of smoothed data

5,10 standard deviations of smoothed data

Appendix 1: Minimizing a Quadratic Function Subject To Nonnegativity Constraints

All of the least squares minimization problems that we are required to solve in order to implement our flexible of order k smoothing procedure have the following structure: minimize a strictly convex quadratic function subject to some nonnegativity restrictions on the variables. In order to solve problems of this type, it is convenient to first study the following specific quadratic minimization problem involving M + 1 variables where $M \geq 3$.

Define the objective function f as a function of the decision variables x_0 and $x \equiv [x_1, \ldots, x_M]^T$ by

$$\text{(A1)} \quad f(x_0, x) \equiv - [a_0, a^T] \begin{bmatrix} x_0 \\ x \end{bmatrix} + (1/2)\, [x_0, x^T] \begin{bmatrix} b_0, & b^T \\ b, & B \end{bmatrix} \begin{bmatrix} x_0 \\ x \end{bmatrix}$$

where a_0 and b_0 are constants, $a^T \equiv [a_1, \ldots, a_M]$ and $b^T \equiv [b_1, \ldots, b_M]$ are vectors of constants and $B \equiv [b_{ij}]$ is a given symmetric positive definite matrix. We also assume that the bordered B matrix in (A1) is positive definite. Our goal is to solve:

$$\text{(A2)} \quad \min_{x_0,\, x}\{f(x_0, x):\ x_0 \geq 0, x_3 \geq 0, x_4 \geq 0, \ldots, x_M \geq 0\}.$$

Note that x_1 and x_2 are unrestricted in (A2). This corresponds to the fact that β_1 and β_2 are unrestricted in the least squares minimization problems that appeared in sections 4 and 5.

We first minimize $f(x_0, x)$ with respect to the components of x leaving $x_0^* \geq 0$ fixed temporarily. A candidate x* to solve this minimization problem can be obtained by solving the first order necessary conditions for minimizing $f(x_0^*, x)$ with respect to x, ignoring the nonnegativity constraints on the components of x:

$$\text{(A3)} \qquad \nabla_x f(x_0^*, x^*) = -a + bx_0^* + Bx^* = 0_M \quad \text{or}$$

$$\text{(A4)} \qquad x^* = B^{-1}[a - bx_0^*].$$

We now <u>assume</u> that the last M - 2 components of x* defined by (A4) are <u>positive</u>; i.e., (A5) $x_m^* > 0$, m = 3, 4, . . ., M.

Since $x_0^* \geq 0$ by assumption, we see that $[x_0^*, x^*]$ is feasible for the minimization problem defined by (A2). Our final assumption on problem (A2) is:

(A6) $\partial f(x_0^*, x^*) / \partial x_0 = -a_0 + b_0 x_0^* + b^T x^* < 0.$

Assumption (A6) means that increasing x_0 from its initial level of x_0^* would decrease the objective function f. Thus our initial feasible solution $[x_0^*, x^*]$ is not an optimal solution for (A2).

We now attempt to solve (A2) under the above conditions.

We first solve the following unconstrained minimization problem:

(A7) $\min_{x_0, x}\{f(x_0, x)\colon\ x_0, x \text{ unrestricted}\}.$

In view of our positive definiteness assumptions, the solution to (A7) is determined by the following first order conditions:

(A8) $$f(\hat{x}_0, \hat{x}) = -\begin{bmatrix} a_0 \\ a \end{bmatrix} + \begin{bmatrix} b_0, & b^T \\ b, & B \end{bmatrix}\begin{bmatrix} \hat{x}_0 \\ \hat{x} \end{bmatrix} = 0_{M+1} \text{ or}$$

(A9) $$\begin{bmatrix} \hat{x}_0 \\ \hat{x} \end{bmatrix} = \begin{bmatrix} b_0, & b^T \\ b, & B \end{bmatrix}^{-1}\begin{bmatrix} a_0 \\ a \end{bmatrix}.$$

Using partitioned matrices, the inverse in (A9) can be expressed as

(A10) $$\begin{bmatrix} b_0, & b^T \\ b, & B \end{bmatrix}^{-1} = \begin{bmatrix} [b_0 - b^T B^{-1} b]^{-1} & , -[b_0 - b^T B^{-1} b]^{-1} b^T B^{-1} \\ -B^{-1} b [b_0 - b^T B^{-1} b]^{-1}, & B^{-1} + B^{-1} b [b_0 - b^T B^{-1} b]^{-1} b^T B^{-1} \end{bmatrix}.$$

Since the inverse matrix is positive definite, its diagonal elements are positive and hence

(A11) $[b_0 - b^T B^{-1} b]^{-1} > 0.$

Substituting (A10) into (A9) yields:

(A12) $\hat{x}_0 = [b_0 - b^T B^{-1} b]^{-1} [a_0 - b^T B^{-1} a]$ and

(A13) $\hat{x} = B^{-1}a - B^{-1}b\hat{x}_0$.

Substituting (A4) into (A6) yields:

(A14) $-a_0 + b_0 x_0^* + b^T B^{-1}[a - bx_0^*] = -a_0 + b^T B^{-1} a + [b_0 - b^T B^{-1} b] x_0^* < 0.$

The inequality in (A14) implies that

(A15) $a_0 - b^T B^{-1} a > [b_0 - b^T B^{-1} b] x_0^* \geq 0$

where the last inequality follows from (A11) and the assumption that $x_0^* \geq 0$. Using (A12), (A11) and (A15),

(A16) $\hat{x}_0 = [b_0 - b^T B^{-1} b]^{-1} [a_0 - b^T B^{-1} a] > x_0^*$.

We now consider two cases.

Case (i): $\hat{x}_m \geq 0$ for m = 3, 4, . . ., M.

In this case, $[\hat{x}_0, \hat{x}]$ solves the unrestricted minimization problem (A7) and is feasible for the restricted minimization problem (A2). Hence $[\hat{x}_0, \hat{x}]$ also solves (A2) and we are done.

Case (ii): $\hat{x}_m < 0$ for at least one m, m = 3, 4, . . ., M.

In this case, instead of increasing x_0 from x_0^* to $\hat{x}_0$, we increase x_0 just enough to make one component of $\tilde{x} \equiv B^{-1}a - B^{-1} b \tilde{x}_0$ equal to zero while keeping the remaining components nonnegative. Thus define the vectors c and d by

(A17) $c \equiv B^{-1}a$ and $d \equiv B^{-1}b$

and define $\tilde{x}_0$ by

(A18) $\tilde{x}_0 \equiv \min_m\{c_m / d_m$: m = 3, 4, . . ., M and $\hat{x}_m < 0\}$.

By (A4) and (A17), $x^* = c - dx_0^*$ and by (A13) and (A17), $\hat{x} = c - d\hat{x}_0$. Since $x_m^* > 0$ for m = 3, 4, . . ., M and $\hat{x}_0 > x_0^* \geq 0$ by (A16), if $\hat{x}_m < 0$ for any m = 3, 4, . . ., M then we must have $c_m > 0$ and $d_m > 0$. Hence the minimum in (A18) is positive. It is also evident that $\tilde{x}_0 > x_0^*$. Hence, in case (ii), we have

(A19) $0 \leq x_0^* < \tilde{x}_0 < \hat{x}_0$.

Now use (A13) to define the vector $\tilde{x}$ that corresponds to $\tilde{x}_0$:

(A20) $\tilde{x} \equiv B^{-1}a - B^{-1}\tilde{x}_0 = c - d\tilde{x}_0$

where the last equality in (A20) follows using (A17). Using (A20) and (A18), we see that the M dimensional vector $\tilde{x}$ has at least one zero component. In fact, for all indexes m which attain the minimum in (A18) (call this set of indexes S), we have $\tilde{x}_m = 0$. It is also the case that $\tilde{x}_m \geq 0$ for m = 3, 4, . . ., M. Hence $(\tilde{x}_0, \tilde{x})$ is feasible for (A2).

We now show that

(A21) $f(\tilde{x}_0, \hat{x}) < f(x_0^*, x^*)$.

In fact, we now show that f is monotonically decreasing along the line segment joining (x_0^*, x^*) to $(\hat{x}_0, \hat{x})$. Using (A4) and (A13), we see that this path defined by

(A22) $x_0(t) \equiv t$; $x(t) \equiv B^{-1}a - B^{-1}bt$; $x_0^* \leq t \leq \hat{x}_0$.

The gradient vector of f with respect to x along this path is

$$\begin{aligned} {}_x f[x_0(t), x(t)] &= -a + bx_0(t) + bx(t) && \text{differentiating (A1)} \\ &= -a + bt + B[B^{-1}a - B^{-1}bt] && \text{using (A22)} \\ \text{(A23)} \qquad &= 0_M. \end{aligned}$$

The equations (A23) and the positive definiteness of B imply that, conditional on $x_0(t)$ being fixed, $x(t)$ globally minimizes $f[x_0(t), x]$ with respect to the components of x. We also have the following partial derivatives of f with respect to x_0 at the endpoints of the path defined by (A22):

(A24) $\partial f(x_0^*, x^*) / \partial x_0 < 0$ by (A6);

(A25) $\partial f(\hat{x}_0, \hat{x}) / \partial x_0 = 0$ by the first equation in (A8).

Since $f(x_0, x)$ is a strictly convex quadratic function, the linearity of the path defined by (A22) and (A23) - (A25) imply that the directional derivative of f along the interior of the path is negative and hence $f[t, x(t)]$ decreases monotonically as t goes from x_0^* to $\hat{x}_0$. Since by (A19), $t = \tilde{x}_0$ is in the interior of this interval, (A21) follows.

This completes our discussion of case (ii).

We now explain how the above analysis can be used to either solve the inequality constrained minimization problem (A2) or repeatly decrease the objective function. First, we solve the unconstrained minimization problem (A7). If we terminate in case (i), we have solved (A2). If we terminate in case (ii), then we attempt to solve the following inequality constrained minimization problem (A26):

(A26) $\min_{x_0,x}\{f(x_0, x)$: $x \equiv [x_1, \ldots, x_M]^T$; x_1, x_2 unrestricted; $x_m \geq 0$,
$m = 3, 4, \ldots, M$; $x_m = 0$ for $m \quad S\}$.

Recall that S is the set of indexes m which attain the minimum in (A18). Thus (A26) has the same general form as (A2), except there is at least one less nonnegative free variable in the new problem (the number is exactly one if the minimum in (A18) is uniquely attained). The inequality (A6) is now replaced by

(A27) $\partial f(\tilde{x}_0, \hat{x}) / \partial x_0 < 0$

which follows from the fact that $\partial f[x_0(t), x(t)] / \partial x_0 < 0$ for all t such that $x_0^* \leq t < \hat{x}_0$, where the path $[x_0(t), x(t)]$ is defined by (A22).

Again, we obtain cases (i) and (ii) for (A26). In both cases, the objective function is decreased. If we end up in case (i), we terminate. If we end up in case (ii), we can drop at least one additional nonnegative variable from the minimization problem after computing the counterpart to (A18). We then obtain a new minimization problem similar to (A26) except that at least one more nonnegative x_m is set equal to zero.

We keep iterating the above process until we finally end up in case (i). The process must terminate in at most M - 2 iterations, because after M - 2 of the x_m for $m = 3, 4, \ldots M$ have been set equal to zero, we must end up in case (i).

We now use the above material to suggest a finite algorithm for a quadratic programming problem with nonnegativity restrictions on the decision variables.

We want to solve problems of the following form:

(A28) $\min_\beta\{(y - X\beta)^T(y - X\beta)$: $\beta \equiv [\beta_1, \beta_2, \ldots, \beta_N]^T$; $\beta_n \geq 0$ for $n = 3, 4, \ldots, N\}$.

In our particular case, $X \equiv [1_N, \tau^{(2)}, \delta_3\tau^{(3)}, \delta_4\tau^{(4)}, \ldots, \delta_N\tau^{(N)}]$ where $\delta_n \equiv \pm 1$, but the algorithm will work as long as X has full column rank.

The Kuhn-Tucker necessary and sufficient conditions for (A28) are given by (28) - (31) and the following conditions:

(A29) $u_n \leq 0$ for $n = 3, 4, \ldots, N$;

(A30) $\beta_n \geq 0$ for $n = 3, 4, \ldots, N$.

We now present a sketch of our algorithm.

General Interation 0:

Do a least squares regression for (A28) using only β_1 and β_2, the unrestricted variables. Define $X^{(0)} \equiv [1_N, \tau^{(2)}]$ and

(A31) $\left[\beta_1^{(0)}, \beta_2^{(0)}\right]^T \equiv [X^{(0)T} X^{(0)}]^{-1} X^{(0)T} y$.

Define $\beta_n^{(0)} \equiv 0$ for $n = 3, 4, \ldots, N$ and $\beta^{(0)} \equiv \left[\beta_1^{(0)}, \beta_2^{(0)}, \beta_3^{(0)}, \ldots, \beta_N^{(0)}\right]^T$. Then the vector of residuals $e^{(0)}$ and the vector of Kuhn-Tucker multipliers can be defined as

(A32) $e^{(0)} \equiv y - X\beta^{(0)}$;

(A33) $u^{(0)} \equiv X^T e^{(0)} = X^T y - X^T X \beta^{(0)}$.

Terminate if

(A34) $u_n^{(0)} \leq 0$ for $n = 3, 4, \ldots, N$.

If conditions (A34) are not satisfied, define $\max_n\{u_n^{(0)}: n = 3, 4, \ldots, N\} > 0$ and let i_1 be the smallest index which attains this maximum. Go to the next general iteration where we will allow the variable β_{i_1} to be nonzero.

General Iteration 1:

Do a least squares regression for (A28) using only β_1, β_2 and β_{i_1}. Define $X^{(1)} \equiv [1_N, \tau^{(2)}, \delta_{i_1}\tau^{(i_1)}]$ and

(A35) $\left[\beta_1^{(1)}, \beta_2^{(1)}, \beta_{i_1}^{(1)}\right]^T \equiv [X^{(1)T} X^{(1)}]^{-1}X^{(1)T}y.$

Let the remaining $\beta_n^{(1)} \equiv 0$ and let $\beta^{(1)}$ be the resulting N dimensional vector. Define

(A36) $e^{(1)} \equiv y - X\beta^{(1)};$

(A37) $u^{(1)} \equiv X^Te^{(1)} = X^Ty - X^TX\beta^{(1)}.$

Terminate if

(A38) $u_n^{(1)} \leq 0$, $n = 3, 4, ..., N$.

If conditions (A38) are not satisfied, define $\max_n\{u_n^{(1)}$: $n = 3, 4, . . ., N\} > 0$ and let i_2 be the largest index which attains this maximum. Go to the next general iteration where we will allow the nonnegative variables β_{i_1} and β_{i_2} to be positive.

General iteration k:

We need to define the notation for the termination of general iteration k - 1 which sends us to general iteration k. At the end of general iteration k - 1, we have say L positive variables, β_{i_1}, β_{i_2} ,. . ., β_{i_L}. Define the matrix $X^{(k-1)} \equiv [1_N, \tau^{(2)}, \delta_{i_1}\tau^{(i_1)}, . . ., \delta_{i_L}\tau^{(i_L)}]$ with $\left[\beta_1^{(k-1)}, \beta_2^{(k-1)}, \beta_{i_1}^{(k-1)}, .. ., \beta_{i_L}^{(k-1)}\right]^T \equiv$ $\{X^{(k-1)T}X^{(k-1)}]^{-1}X^{(k-1)T}y$. Define the remaining $\beta_n^{(k-1)} = 0$ and let $\beta^{(k-1)}$ be the resulting N dimensional vector. Define $e^{(k-1)} \equiv y - X\beta^{(k-1)}$ and $u^{(k-1)} \equiv X^Te^{(k-1)}$. Finally, at the end of the general iteration k-1, we have

(A39) $\max_n\{u_n^{(k-1)}$: $n = 3, 4, . . ., N\} > 0$.

Let i_{L+1} be the smallest index which attains the maximum in (A39). Now we are ready to start general iteration k.

Subiteration:

Do a least squares regression using only β_1, β_2, β_{i_1}, . . ., β_{i_L}, $\beta_{i_{L+1}}$. Define $X^{(k)} \equiv [1_N, \tau^{(2)}, \delta_{i_1}\tau^{(i_1)}, \ldots, \delta_{i_L}\tau^{(i_L)}, \delta_{i_{L+1}}\tau^{(i_{L+1})}]$ and

(A40) $[\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_{i_1}, \ldots, \hat{\beta}_{i_L}, \hat{\beta}_{i_{L+1}}]^T \equiv [X^{(k)T}X^{(k)}]^{-1}X^{(k)T}y.$

Define the remaining $\hat{\beta}_n = 0$ and let $\hat{\beta}$ be the resulting N dimensional vector. Define

(A41) $\hat{e} \equiv y - X\hat{\beta}$;

(A42) $\hat{u} \equiv X^T\hat{e}$.

Case (i): $\hat{\beta}_n \geq 0$ for n = 3, 4, . . ., N.

Case (i) (a): $\max_n\{\hat{u}_n$: n = 3, 4, . . ., N$\} \leq 0$.

In this case, the Kuhn-Tucker conditions for (A28) are satisfied, and we terminate the algorithm.

Case (i) (b): $\max_n\{\hat{u}_n$: n = 3, 4, . . ., N$\} > 0$.

In this case, define $\beta^{(k)} = \hat{\beta}$, $e^{(k)} \equiv y - X\beta^{(k)}$, $u^{(k)} \equiv X^Te^{(k)}$, terminate the subiterations and go to general iteration k + 1.

Case (ii): $\hat{\beta}_n < 0$ for at least one n = $i_1, i_2, \ldots, i_L$.

In this case, we have the following counterparts to (A4), (A13), and (A16):

(A43) $\hat{\beta} = \beta^{(k-1)} - d\hat{\beta}_{i_{L+1}}$:

(A44) $\hat{\beta}_{i_{L+1}} > 0$.

Using (A43) and (A44), the components of the vector d have the following representation:

(A45) $d_n = - [\hat{\beta}_n - \beta_n^{(k-1)}] / \hat{\beta}_{i_{L+1}}$; $n = 1, 2, 3, . . , N$.

The counterpart to (A18) is now

(A46) $\tilde{\beta}_{i_{L+1}} \equiv \min_n\{\beta_n^{(k-1)} \hat{\beta}_{i_{L+1}} / [\beta_n^{(k-1)} - \hat{\beta}_n]:\ n = i_1, i_2, . . , i_L \text{ and } \hat{\beta}_n < 0\}$
> 0,

where the inequality follows from (A19). Thus $\tilde{\beta}_{i_{L+1}}$ corresponds to $\tilde{x}_0$ in (A19) and the counterpart to (A20) is

(A47) $\tilde{\beta} \equiv \beta^{(k-1)} - d\, \tilde{\beta}_{i_{L+1}}$.

It can also be shown that

(A48) $\tilde{\beta}_n \geq 0$ for $n = 3, 4, . . ., N$.

In this case, we return to subiteration but we drop columns in the $X^{(k)}$ matrix that correspond to the indexes n which attained the minimum in (A46). If we end up in case (ii) of the new subiteration, then the $\beta^{(k-1)}$ which appears in (A43) - (A47) is replaced by the $\tilde{\beta}$ defined by (A47) in the present subiteration.

The subiterations continue as long as we remain in case (ii). We eventually emerge from case (ii) since there can be at most L subiterations.

This completes the description of our algorithm. It can be seen that our algorithm must converge after a finite number of iterations because there is only a finite number of choices for nonzero β_n. At each iteration and subiteration of our algorithm, the objective function strictly decreases.

Note that the only computationally intensive parts of our algorithm are the matrix inversions in (A31), (A35) and (A40). However, each of these inversions is a simple one column update (one column is either added or deleted) of the previous matrix inversion, and so standard updating formulae can be used to perform the inversions. Note also that our algorithm starts by inverting a two by two matrix and hence most of the inversions are performed on matrices of rather low dimension. In fact, if columns are dropped only infrequently (which was the case in our empirical work) and if we use the updating formulae, our entire algorithm requires only a little more work than a single M by M matrix inversion, where M is the number of nonzero variables that we end up with.

Appendix 2: Unequally Spaced Data and Multiple Data Points

Suppose that the independent variable t is no longer equally spaced. Suppose that there are N distinct t values and they are ordered as follows: $t_1 < t_2 < \ldots < t_N$. In this case, the vectors $\tau^{(n)}$ which occur in the least squares problems described above are redefined as follows: for n = 2, 3, . . ., N,

$$\text{(A49)}\quad \tau^{(n)} \equiv [0^T_{n-1}, t_n - t_{n-1}, t_{n+1} - t_{n-1}, \ldots, t_N - t_{n-1}]^T$$

where 0_{n-1} is a column vector of n-1 zeros. The discrete smooth $s \equiv [s_1, \ldots, s_N]^T$ that corresponds to the rough $y \equiv [y_1, \ldots, y_N]$ is still $s \equiv 1_N\beta_1 + \Sigma_{n=2}^N \tau^{(n)}\beta_n$ and the linear extension of s is defined as follows:

$$\text{(A50)}\quad s(t) = \begin{cases} \beta_1 + (t - t_1)\beta_2 & \text{for } t_1 \leq t \leq t_2; \\ \beta_1 + (t - t_1)\beta_2 + (t - t_2)\beta_3 & \text{for } t_2 \leq t \leq t_3; \\ \vdots & \\ \beta_1 + (t - t_1)\beta_2 + \;.\;\;.\;\;.\; + (t - t_{N-1})\beta_N & \text{for } t_{N-1} \leq t \leq t_N . \end{cases}$$

The remainder of our analysis is unchanged; e.g., our suggested flexible of order 2 smoothing procedure still satisfies tests (8) - (15) and (39) for k = 2. However, the statements of some of the tests must be changed slightly. For example, test (9) becomes: if s f(y), then $\Sigma_{n=1}^N t_n s_n = \Sigma_{n=1}^N t_n y_n$, and test (11) becomes: if $y_n = \alpha + \beta t_n$ for n = 1, 2, . . ., N, and $y \equiv [y_1, \ldots, y_N]^T$, then $f(y) \equiv \{y\}$.

Our smoothing method can also be modified to deal with the case where there are multiple observations for each distinct value of the independent variable. Thus for each distinct t_n value, suppose that there are M_n y values for n = 1, 2, . . ., N. In this case, we define y_n to be the <u>mean</u> of the y values that correspond to t_n for n = 1, . . ., N. Define the N dimensional vector $y \equiv [y_1, y_2, \ldots, y_N]^T$ to be this vector of means. Define the N by N diagonal weighting matrix W with diagonal elements $w_{nn} = M_n$ for n = 1, 2, . . ., N. The objective function for our least squares minimization problems now becomes $(y - X\beta)^T W(y - X\beta)$. Our suggested smoothing algorithm will still work in this multiple observations framework and it will have the same axiomatic properties after the axioms are modified a bit. The counterparts to axioms (8), (9) and (14) are:

(A51) if s f(y), then $1_N^T\, Ws = 1_N^T\, Wy$;

(A52) if $s \quad f(y)$, then $\tau^{(2)T} Ws = \tau^{(2)T} Wy$;

(A53) if $s \quad f(y)$, then $s^T Ws \leq y^T Wy$.

The other axioms are either unchanged or trivially modified as indicated in the modifications of tests (9) and (11) that we presented below (A50).

Finally, if the errors e_t are heteroskedastic or serially correlated, with known positive definite variance-covariance matrix Ω, then the weighting matrix W should be chosen to be proportional to Ω^{-1}.

Footnotes

* The authors are Professors of Economics at The University of British Columbia, Vancouver, Canada. We are indebted to John Cragg, David Donaldson, Kevin Fox, James Nason, Joris Pinske, Margaret Slade and Kenneth White for valuable discussions. This research was supported by a SSHRC strategic grant. Our thanks to Tina Corsi for typing a complex manuscript.

[1] Recent articles that contain further references to the vast literature on smoothing or nonparametric regression are Borgan (1979), Cleveland (1979), Craven and Wahba (1979), Wegman and Wright (1983), Breiman and Friedman (1985) (with discussion), Silverman (1985) (with discussion), McDonald and Owen (1986), Ramsay (1988) (with discussion), Tibshirani (1988), Buja, Hastie and Tibshirani (1989) (with discussion), Friedman and Silverman (1989) (with discussion), Altman (1992) and Breiman and Peters (1992). The articles with discussion are particularly informative and reference rich.

[2] The tradeoff between fit and smoothness is usually determined by a smoothing parameter; see Buja, Hastie and Tibshirani (1989) for a nice catalogue of smoothing parameters for the commonly used methods.

[3] The number of sign changes plays the role of a smoothing parameter.

[4] The sawtooth function was used in the simulation studies of McDonald and Owen (1986) and Breiman and Peters (1992). Breiman and Peters (1992; 278) refer to the sawtooth function as the most challenging function for a smoothing method to fit.

[5] Hildreth's (1954; 616-619) proposed algorithm does not necessarily converge in a finite number of iterations since it is subject to zig-zagging or the oscillation phenomenon; see Wilde and Beightler (1967; 78-79) for an explanation of the problem.

[6] The English actuary Woolhouse (1866; 139-141) implicitly defined a series to be smooth if its 4th or 6th differences were zero while, the American

mathematician De Forest (1873; 325) defined smoothness in terms of the smallness of the absolute values of the series 4th differences. Woolhouse (1866; 137) used the term "graduated progression" to describe a smooth while De Forest (1873; 325) used the terms "comparative regularity of the graduation" and its "smoothness of adjustment". Woolhouse (1866; 175-176) appears to have been the first to advocate the least squares fitting of low order polynomials as a smoothing technique. Woolhouse (1870; 392) was also the first to derive a moving average smoother that was exact for a low order polynomial. On the other hand, De Forest (1873; 290-292) showed how moving average smoothers of varying window length could be derived that were exact for cubic functions. In Appendix I of his paper, De Forest (1873; 322-324) showed how the weights for his exact moving average estimators could be chosen to resemble kernel smoother weights while in Appendix II, he showed how the weights could be chosen to minimize the squared 4th differences of the smooth. For a more detailed account of De Forest's work and the early actuarial research on smoothing, see Wolfenden (1925).

7. See for example Camp (1955; 9). In the actuarial literature, the process of smoothing a mortality table was known as graduating the data; i.e., the little hills and valleys of the rough were to be graded into smoothness, just as in building a road over rough terrain. The road building analogy for smoothing was initiated by Higham (1886; 16) and Sprague (1887; 112).

8. Hodrick and Prescott (1980; 5) and Kydland and Prescott (1990; 9) recommended that $\lambda = 1600$ be used when smoothing quarterly economic data.

9. See Buja, Hastie and Tibshirani (1989; 459). The statistics literature addressed the problem of estimating the trade off parameter λ instead of imposing it a priori: Wahba and Wold (1975) used the cross validation technique (see Stone (1974) for a description) and Craven and Wahba (1979) invented the generalized cross validation technique to determine λ.

10. Bizley (1958; 136) pointed out that if "small" were interpreted to mean "zero", then the only smooth curves according to his definition would be straight lines and circles.

11. The modern literature on smoothing methods has made little use of Sprague's definition of smoothness. His definition has occasionally been resuggested in the actuarial literature. Thus Bizley (1956; 99) in his discussion of a paper by the British actuary Sir William Elderton suggested that an alternative definition of smoothness might be that the second differential y"(t) of the curve y(t) should change sign as few times as possible and Mr. L.V. Martin in discussing Bizley's (1958; 147) paper remarked that $\Delta^2 s_t$ should pass through zero as infrequently as possible. Sprague's definition has also been mentioned occasionally in the statistics literature. Thus Good and Gaskins (1971; 258) initially define the roughness of a curve as being proportional to the number of inflection points that it exhibits and J.F.C. Kingman in discussing Boneva, Kendall and Stefanov's (1971; 55) method for smoothing histograms suggested that the resulting smoothed curve should have no more local maxima than the original histogram.

12. Mr. C.D. Higham, who was a discussant of Sprague's paper, noted that a principle advantage of Woolhouse's (1870) moving average method of smoothing was its reproducibility, as the following quotation indicates: "Moreover, any number of people using such a formula as his [Woolhouse's], would bring out the same results, whereas by the graphic method there would be as many different graduations as operators". [C.D. Higham; see Sprague (1887; 116)].

13. We allow f to be a set valued function in the general case for the following reason: many smoothing methods (including ours) define the smooth to be the solution to an optimization problem. If the solution is not unique, then the smooth could be any member of the solution set.

14. Sprague was attempting to smooth a mortality table which is essentially a probability distribution. Thus if test (8) is satisfied and the rough is a probability distribution, then the smooth is also a probability distribution (provided that $s \geq 0_N$). Note that test (8) implies that the arithmetic mean of the smooth equals the arithmetic mean of the rough. Thus test (8) might be called the mean preserving test.

15. Whittaker (1923; 67) showed that his smoothing method, which was based on solving an analogue to (5) where $\Sigma_{t=3}^{N} \Delta^2 s_t$ was replaced by $\Sigma_{t=4}^{N} \Delta^3 s_t$, satisfied tests (8) and (9). Whittaker termed tests (8) and (9) theorems of the conservation of moments of order 0 and 1 respectively. Note that if we are smoothing a probability distribution and test (9) is satisfied, then the mean of the smooth will equal the mean of the rough.

16. Schoenberg (1946; 53) also used the tests (10) - (13) to evaluate smoothing methods. Schoenberg (1946; 48) was the first to use the term "spline curve" to describe a curve made up of polynomials defined over line segments which are joined up at the end points in a continuously differentiable manner. However, the concept of a spline curve appears to be due to the actuary Sprague (1891; 277). In the actuarial literature, the construction of splines is called osculatory interpolation; e.g., see Greville (1944).

17. Greville (1944; 205) used the term "the reproduction of a constant" property. Our terminology is motivated by the test approach to index number theory; e.g., see Eichhorn (1976) or Diewert (1992).

18. Greville (1944; 211) used the terms "degree of reproduction 2 and 3" to describe properties (12) and (13). In Schoenberg's (1946; 53) terminology, the smoothing formula is <u>exact</u> for degrees 2 and 3 if f satisfies (12) and (13).

19. Buja, Hastie and Tibshirani (1989; 465) term a smoothing method f that possesses property (14) a "shrinking smoother". Schoenberg (1946; 52) also proposed the following sequence of tests: $s \quad f(y)$ implies $\Sigma_t(\Delta^m s_t)^2 \leq \Sigma_t(\Delta^m y_t)^2$ for m = 1, 2, ...

20. To show that the Henderson smoother satisfies property (14), we follow Buja, Hastie and Tibshirani's (1989; 466) example and proceed as follows. Rewrite (5) in matrix notation as $\min_s\{(y - s)^T(y - s) + \lambda\, s^T D^T D s\}$ for $\lambda \geq 0$ where s^T denotes the transpose of s and D is an N - 2 by N dimensional matrix which is constructed so that Ds represents the vector of second differences of s. The first order necessary conditions for the minimization problem lead to the equation $[I_N + \lambda D^T D]s = y$. The Hessian matrix for the objective function is the matrix $[I_N + \lambda D^T D]$ which is positive definite

since $\lambda D^T D$ is positive semidefinite and I_N is positive definite. Thus $s = [I_N + \lambda D^T D]^{-1} y \equiv Ay$ and the eigenvalues of A, λ_i say, satisfy the inequalities $0 < \lambda_i \leq 1$ for i = 1, 2, . . ., N since the eigenvalues of $A^{-1} \equiv [I_N + \lambda D^T D]$ are all greater than or equal to 1. Similarly, the Whittaker smoother can be defined as the solution to $\min_s\{(y - s)^T(y - s) + \lambda s^T E^T Es\}$ for $\lambda \geq 0$ where E is an N - 3 by N dimensional matrix such that Es represents the vector of third differences of s.

21. This same criticism applies to many nonparametric smoothing methods. This problem has not gone unnoticed in the current smoothing literature as the following quotations indicate:

> "I would like to suggest that a desirable property of a method of smoothing histograms is that the resulting smooth curve should have no more local maxima than the original histogram. This property is not enjoyed by the histospline, nor indeed by Professor Daniel's alternative. But it might be possible to invoke the powerful theory of variation – diminishing transformations developed by I.J. Schoenberg and others to combine this requirement with conditions of smoothness and so produce a smoothing technique which had less of a tendency to produce 'rabbits'." [Professor J.F.C. Kingman discussing Boneva, Kendall and Stefanov (1971; 55)],
>
> "In fact, if one smoothes a genuinely smooth curve, such as a cubic polynomial, the running-line smoother can put wiggles in the output!" [Buja, Hastie and Tibshirani (1989; 464)].

22. Conversely, given any convex function f(t) defined over the interval [1, N], then $\Delta^2 f_t \geq 0$ for t = 3, 4, . . ., N where $f_t \equiv f(t)$ for t = 1, 2, . . ., N.

23. Strictly speaking, we also require at least one strict inequality in (18) and at least one in (19) to speak of a turning point.

24. A more appropriate (but longer) terminology would be "Sprague smooth of order k".

25. The bumps and dips terminology may be found in Good and Gaskins (1980; 42).

26. The uniqueness follows from the fact that the X matrix has full rank N. Thus the Hessian matrix of the objective function in (27), $2X^TX$, is positive definite and hence the objective function is a strictly convex function of β.

27. However, Hildreth and Wagner did not derive the fundamental reparameterization of s given by (25) and (26) and they did not present an effective algorithm for solving their nonlinear programming problems. Hildreth's (1954; 605) suggested algorithm is not finite and it is also subject to the zig-zagging phenomenon; see Wilde and Beightler (1967; 79). Finally, Hildreth and Wagner did not discuss their methods in the context of alternative definitions of smoothness.

28. The signs of the second differences of a quadratic function are constant and this property implies that the quadratic trend test (12) will be satisfied.

29. Equivalently, y is a 2 smooth if and only if its linear extension exhibits at most one hill and one valley or at most one bump and one dip or at most one region of convexity and one region of concavity.

30. The cubic trend test (15) is satisfied because a cubic function can have at most one turning point.

31. Scientific in the sense that our method is reproducible; different operators should end up with the same smooth.

32. Our flexible of order 3 smoothing method satisfied tests (8) - (15). Moreover, it also satisfies the <u>biquadratic trend test</u>: if $y_t = \alpha + \beta t + \gamma t^2 + \delta t^3 + \varepsilon t^4$ for t = 1, 2, . . ., N, then f(y) = {y}.

33. Buja, Hastie and Tibshirani (1989; 460) note that the <u>span size</u> is the smoothing parameter for running mean, running line and bin smoothers, the <u>number of regressors</u> (or the rank of $X(X^TX)^{-1}X^T$) is the smoothing parameter for regression based smoothers and the <u>penalty parameter</u> λ is the smoothing parameter for the Whittaker-Henderson class of smoothers.

34. It is very difficult to derive exact statistical distribution s (independent of s_t) for estimators of the s_t in (1), after making some assumptions about the distribution of the errors e_t in (1), since the s_t are unknown. De Forest (1873; 301) understood these difficulties very well as the following quotation indicates: "Not only is absolute accuracy unattainable, but we cannot even decide, by the method of least squares, that a certain result is the most probable of any; for the true form of the function being unknown, any particular residual error, or difference between the observed and computed values of a term, will in general be the aggregate of two errors, one of them due to the difference of form between the assumed function and the true one, and the other due to the error of observation or difference between the observed value and the true value".

35. We actually use the large sample normal approximation derived by Wald and Wolfowitz (1940; 151). According to Gibbons (1971; 57), this approximation is adequate provided the number of positive and negative residuals both exceed 10. Since our smoothing method will tend to lead to a substantial number of zero residuals, we follow the recommendation of Gibbons (1971; 58) by simply deleting observations that correspond to zero residuals.

36. For references to the cross validation literature, see Allen (1971), Stone (1974) and Geisser (1975).

37. Our suggested method for forming approximate confidence intervals is frequently used in the recent smoothing literature; i.e., see Friedman and Silverman (1989; 9). The situation is nicely summarized by the following quotation due to Leo Breiman in his discussion of Buja, Hastie and Tibshirani (1989; 513). "Once the final model is arrived at, forget that it has been gotten by data-directed variable selection; and compute the intervals in the classical manner based on the final model. Of course this is cheating, but the issue is whether these confidence intervals are reasonable approximations. That is under investigation".

38. See for example, McDonald and Owen (1986; 200-203), Friedman and Silverman (1989; 14) or Breiman and Peters (1992; 287). De Forest (1873; 310) noted that the "accuracy" of a moving average smooth was greatest at the middle of a series and least at its extremities.

39. Leo Breiman in his discussions of Ramsay (1988; 444) and of Buja, Hastie and Tibshirani (1989; 512) suggested that spline functions be linear at the lowest and highest values of the independent variable. Other papers which suggested linearization near the endpoints are Fuller (1969; 41) Friedman and Silverman (1989; 10), Breiman (1991; 132), Breiman and Peters (1992; 275), and Diewert and Wales (1993; 86). As usual, the idea appears to have originated in the actuarial smoothing literature where it is called extending the series by constant first differences; see Henderson (1924; 38) and Greville (1944; 239).

40. Thus instead of solving problems of the form (40), we solve problems of the following form: $\min_\beta\{(y - X\beta)^T W(y - X\beta)$: $\beta \equiv [\beta_1, \beta_2, \ldots, \beta_N]^T$, $\beta_t \geq 0$ for $t = 3, 4, \ldots, N\}$ where W is a positive definite weighting matrix. In the present context where at each age i, we have a different number of people n at risk, we take W to be a diagonal matrix with iith element $w_{ii} \equiv n_i$. Thus the mortality rates which are determined by larger samples get a larger weight in our smoothing procedure. In general, if the variance-covariance matrix of the errors e_t which appear in (1) is Ω, then the weighting matrix W should be chosen to be proportional to Ω^{-1}. The use of a weighting matrix means that our smoothing method no longer satisfies all of the tests listed in section 3. To see how the tests must be modified, see Appendix 2.

41. For all of these calculations we use the algorithm based on the material presented in Appendix 1. Even with a large number of parameters to estimate the algorithm converges rapidly. For example, with four break points and 93 parameters convergence requires approximately 40 seconds of CPU time on an IMB 3081-63. Further, since we did not use updating formulae in our matrix inversions, the efficiency of our algorithm could be greatly improved.

42. We have followed McDonald and Owen in relating the standard deviations of the e_t and y_t in this way. However, in order to reduce computational expense we have used 100 points rather than the 256 equispaced points that they used.

43. However, if a two sided runs test were used to determine k for this example, our chosen k would be appropriately large.

References

Allen, D.M. (1971), "Mean Square Error of Prediction as a Criterion for Selecting Variables", *Technometrics* 13, 469-475.

Altman, N.S. (1992), "An Introduction to Kernel and Nearest—Neighbor Nonparametric Regression", *The American Statistician* 46, 175-185.

Beale, E.M.L. (1955), "On Minimizing a Convex Function Subject to Linear Inequalities", *Journal of the Royal Statistical Society*, Series B 17, 173-184.

Beale, E.M.L. (1959), "On Quadratic Progamming", *Naval Research Logistics Quarterly* 6, 227-243.

Beale, E.M.L. (1967), "Numerical Methods", pp. 134-205 in *Nonlinear Programming*, J. Abadie (ed.), Amsterdam: North-Holland Publishing.

Bizley, M.T.L. (1956), "Discussion", *Journal of the Institute of Actuaries* 82, 96-100.

Bizley, M.T.L. (1958), "A Measure of Smoothness and Some Remarks on a New Principle of Graduation (with discussion)", *Journal of the Institute of Actuaries* 84, 125-165.

Boneva, L.I., D.E. Kendall and I. Stefanov (1971), "Spline Transformations: Three New Diagnostic Aids for the Data Analyst (with discussion)", *Journal of the Royal Statistical Society*, Series B, 33, 1-70.

Borgan, O. (1979), "On the Theory of Moving Average Graduation", *Scandinavian Actuarial Journal* 1979, 83-105.

Bowers, N.L., H.U. Gerber, J.C. Hickman, D.A. Jones and C.J. Nesbitt (1986), *Actuarial Mathematics*, Itasca, Illinois: The Society of Actuaries.

Breiman, L. (1991), The π Method for Estimating Multivariate Functions from Noisy Data (with discussion)", *Technometrics* 33, 125-160.

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

Breiman, L. and S. Peters (1992), "Comparing Automatic Smoothers (a Public Service Enterprise)", *International Statistical Review* 60, 271-290.

Buja, A., T. Hastie and R. Tibshirani (1989), "Linear Smoothers and Additive Models (with discussion)", *The Annals of Statistics* 17, 453-555.

Camp, K. (1955), "New Possibilities in Graduation", *Society of Actuaries Transactions* 7, 6-30.

Cleveland, W.S. (1979), "Robust Locally Weighted Regression and Smoothing Scatterplots", *Journal of the American Statistical Association* 74, 829-836.

Craven, P. and G. Wahba (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.

De Forest, E.L. (1873), "On Some Methods of Interpolation Applicable to the Graduation of Irregular Series, such as Tables of Mortality", *Annual Report of the Board of Regents of the Smithsonian Institution for the Year 1871*, 275-339, Washington, D.C.: Government Printing Office.

Diewert, W.E. (1992), "Fisher Ideal Output, Input and Productivity Indexes Revisited", *Journal of Productivity Analysis* 3, 211-248.

Diewert, W.E. and T.J. Wales (1993), "Linear and Quadratic Spline Models for Consumer Demand Functions", *Canadian Journal of Economics* 26, 77-106.

Eichhorn, W. (1976), "Fisher's Tests Revisited", *Econometrica* 44, 247-256.

Friedman, J.H. and B.W. Silverman (1989), "Flexible Parsimonious Smoothing and Additive Modeling (with discussion)", *Technometrics* 31, 3-39.

Fuller, W.A. (1969), "Grafted Polynominals as Approximating Functions", A*ustralian Journal of Agricultural Economics* 13, 35-46.

Geisser, S. (1975), "The Predictive Sample Reuse Method with Applications", *Journal of the American Statistical Association* 70, 320-328.

Gibbons, J.D. (1971), *Nonparametric Statistical Inference*, New York: McGraw-Hill Book Company.

Good, I.J. and R.A. Gaskins (1971), "Nonparametric Roughness Penalties for Probability Densities", *Biometrika* 58, 255-277.

Good, I.J. and R.A. Gaskins (1980), "Density Estimation and Bump-Hunting by the Penalized Likelihood Method Exemplified by Scattering and Meteorite Data (with discussion)", *Journal of the American Statistical Association*, 75, 42-69.

Greville, T.N.E. (1944), "The General Theory of Osculatory Interpolation", *Actuarial Society of America Transactions* 45, 202-265.

Henderson, R. (1924), "A New Method of Graduation", *Acturial Society of America Transactions* 25, 29-40.

Higham, J.A. (1986), "On the Graduation of Mortality Tables", *Journal of the Institute of Actuaries* 25, 15-24.

Hildreth, C. (1954), "Point Estimates of Ordinates of Concave Functions", *Journal of the American Statistical Association* 49, 598-619.

Hodrick, R.J. and E.C. Prescott (1980), "Post-War U.S. Business Cycles: An Empirical Investigation", Discussion Paper 451, Carnegie-Mellon University.

Kydland, F.E. and E.C. Prescott (1990), "Business Cycles: Real Facts and a Monetary Myth", *Federal Reserve Bank of Minneapolis Quarterly Review*, Spring, 3-18.

McDonald, J.A. and A.B. Owen (1986), "Smoothing with Split Linear Fits", *Technometrics* 28, 195-208.

Ramsay, J.O. (1988), Monotone Regression Splines in Action (with discussion)", *Statistical Science* 3, 425-461.

Schoenberg, I.J. (1946), "Contributions to the Problem of Approximation of Equidistant Data by Analytic Functions", *Quarterly of Applied Mathematics* 4, 45-99 and 112-141.

Silverman, B.W. (1985), "Some Aspects of the Spline Smoothing Approach to Nonparametric Regression Curve Fitting (with discussion)", *Journal of the Royal Statistical Society*, Series B, 47, 1-52.

Sprague, T.B. (1887), "The Graphic Method of Adjusting Mortality Tables (with discussion)", *Journal of the Institute of Actuaries* 26, 77-120.

Sprague, T.B. (1891), "Explanation of a New Formula for Interpolation", *Journal of the Institute of Actuarie*s 22, 270-285.

Stone, M. (1974), "Cross-Validatory Choice and Assessment of Statistical Predictions (with discussion)", *Journal of the Royal Statistical Society*, Series B, 36, 111-147.

Tibshirani, R. (1988), "Estimating Transformations for Regression via Additivity and Variance Stabilization", *Journal of the American Statistical Association* 83, 394-405.

Wagner, H.M. (1962), "Non-Linear Regression with Minimal Assumptions", *Journal of the American Statistical Association* 57, 572-578.

Wahba, G. and S. Wold (1975), "A Completely Automatic French Curve: Fitting Spline Functions by Cross Validation", *Communications in Statistics* 4, 1-17.

Wald, A. and J. Wolfowitz (1940), "On a Test Whether Two Samples Are From the Same Population", *Annals of Mathematical Statistics* 11, 147-162.

Wegman, E.J. and I.W. Wright (1983), "Splines in Statistics", *Journal of the American Statistical Association* 78, 351-365.

Whittaker, E.T. (1923), "On a New Method of Graduation", Proceedings of the Edinburgh Mathematical Society 41, 63-75.

Wilde, D.J. and C.S. Beightler (1967), *Foundations of Optimization*, Englewood Cliffs, New Jersey: Prentice-Hall Inc.

Wolfenden, H.H. (1925), "On the Development of Formulae for Graduation by Linear Compounding, with Special Reference to the Work of Erastus L. De Forest", *Actuarial Society of America Transactions* 26, 81-121.

Woolhouse, W.S.B. (1866), "On Interpolation, Summation, and the Adjustment of Numercial Tables ", *Journal of the Institute of Actuaries* 12, 136-176.

Woolhouse, W.S.B. (1870), "Explanation of a New Method of Adjusting Mortality Tables", *Journal of the Institute of Actuaries* 15, 389-410.