ON ALGORITHMS FOR GENERALISED SMOOTHING SPLINES

M. R. OSBORNE[1] AND TANIA PRVAN[2]

(Received June 18, 1986; revised February 2, 1987)

Abstract

Recently, considerable interest has been shown in the connection between smoothing splines and a particular class of stochastic processes. Here the connection with an equivalent class of least squares problems is used to develop algorithms, and properties of the solution are examined. We give an estimate of the condition number of the solution process and compare this with an estimate for the condition number of the Reinsch algorithm in its conventional implementation.

1. Introduction

Let data values y_i be given at points t_i, $i = 1, 2, \ldots, n$. A smoothing spline of order $2p - 1$ provides a smooth curve through the data by minimising

$$S(f) = \sum_{i=1}^{n}(y_i - f(t_i))^2 + \mu \int_{t_1}^{t_n} f^{(p)^2} \, dt \qquad (1.1)$$

where the smoothing parameter μ provides a compromise between fidelity of the approximation as measured by the first term in (1.1), and smoothness of the approximation as measured by the second. The computation of the smoothing spline, given μ, is generally regarded as straightforward for smll p (for example, an implementation of the algorithm of Reinsch [17] is given in de Boor [5] for the case $p = 2$), and most recent interest has centered on the questions of how to set μ and how to attach confidence intervals to f. Perhaps the most popular method for choosing μ based on the data, is the procedure of generalised cross

[1]Department of Statistics, Research School of Social Sciences, Australian National University, G.P.O. Box 4, Canberra, A.C.T. 2601, Australia.
[2]Department of Pure and Applied Mathematics, Washington State University, Pullman, Washington 99164, U.S.A.

validation championed by Wahba (for example [23]). This now has strong the-
oretical support [20], and can be implemented with small numerical cost—see
[21], [18], and [6] for possible approaches offering different compromises between
simplicity and generality. Explicit in these considerations is the assumption that
the data can be decomposed as

$$y_i = \eta(t_i) + \varepsilon_i \tag{1.2}$$

where $\eta(t)$ can be thought of as an unknown signal which is to be estimated
by minimising $S(f)$, and ε_i is a noise term. Here it is assumed that the ε_i
are independent and distributed as $N(0, \sigma^2)$. This nonparametric estimation
problem has a solution which is a piecewise polynomial of degree $2p - 1$, and
within this framework of piecewise polynomial functions it is possible to interpret
smoothing splines as the mean of a posterior distribution given the data under
the assumption of an appropriately tailored form of prior distribution [19]. This
stochastic setting has the advantage that it makes it possible to attach confidence
intervals to the smoothing spline.

An alternative route to a similar type of result has been given by Wahba [22].
In her approach the signal is modelled by the stochastic differential equation

$$\frac{d^p x}{dt^p} = \sigma \sqrt{\lambda} \frac{dw}{dt} \tag{1.3}$$

where $w(t)$ is a unit scale Wiener process (Billingsley [4]) and λ is a scale pa-
rameter (corresponding to $1/\mu$ in (1.1)) which is to be determined. She assumes
a diffuse prior distribution on the vector of initial conditions on (1.3)

$$\mathbf{x}_0 = [x(t_1), x^{(1)}(t_1), \ldots, x^{(p-1)}(t_1)]^T$$

which is made precise by setting $\mathbf{x}_0 \sim N(0, \gamma^2 I)$ and letting $\gamma^2 \to \infty$, and shows
that

$$f(t) = \lim_{\gamma \to \infty} E\{x(t) | y_1, y_2, \ldots, y_n\}. \tag{1.4}$$

This result has stimulated other work. For example, Weinert, Byrd and Sidhu
[25] observe that a natural setting for characterising a smoothing spline is an ap-
propriate reproducing kernel Hilbert space, and that the reproducing kernel can
be interpreted as a covariance function in an isomorphic and isometric Hilbert
space of random variables. This provides them with a stochastic interpretation
of the smoothing spline which is readily generalised to L_g splines, and they also
describe filtering and smoothing algorithms for its computation. An alternative
approach is adopted by Wecker and Ansley [24]. They assume $\mathbf{x}_0$ is a constant
vector and use the Markov property of the Wiener process to set up an equiv-
alent discrete stochastic process describing the evolution of $\mathbf{x}(t_i)$. This permits
them to use the Kalman filter to construct a likelihood function for the prob-
lem, and they maximise this to estimate $\mathbf{x}_0, \lambda, \sigma^2$. Kohn and Ansley [13] show

how to modify the Kalman filter in order to handle the computations involving the diffuse prior in the Wahba model (other authors, for example [10], have worked explicitly with large values of γ^2), and discuss equivalences between the approaches based on stochastic models.

Here we consider a stochastic differential equation model which includes (1.3) as a special case (our aim is to escape from purely deterministic equations for compartmental modelling, but that is another story). We summarise filtering and smoothing algorithms available for estimating the solution given the data, and make an application of these procedures to derive smoothness results for the solution, which provide a neat generalisation of the variational characterisation of smoothing splines. To ensure efficient and numerically stable computation, we have investigated the use of a novel information-filter formulation of the Kalman filter due to Paige and Saunders [15] which has the advantage that the smoothing step, which is a major component of the other algorithms, comes practically for nothing in the sense that it is conceptually an integral part of the algorithm. Both the Wahba, and Wecker and Ansley, models can be estimated by this approach, and in the Wahba model the starting procedure needed to take account of the diffuse prior leads to an initialisation which is identical to that used by Weinert, Byrd and Sidhu. Although both approaches must be equivalent in the sense that both lead to the same smoothing spline, the mechanism of the equivalence is not trivial [3].

The remaining question considered is, which of the numerical algorithms should be used? This question is settled easily if efficiency is the main criterion, for then there would seem to be no serious competition to the Reinsch algorithm. This appears to be the case, for we argue in Section 5 that the sensitivity of both types of algorithm is similar in having the same asymptotic dependence on n. More serious is the size of the estimate, which suggests that numerical problems are likely to be encountered for small values of p and modest values of n with both approaches, and this is confirmed for the information-filter algorithm by presenting the results of calculations.

2. The stochastic model

Let $T(t, \xi)$ satisfy the initial value problem

$$\frac{d}{dt}T(t, \xi) = M(t)T(t, \xi), \qquad T(\xi, \xi) = I \tag{2.1}$$

where $M(t)\colon R^p \to R^p$ is sufficiently smooth to permit subsequent manipulations. Then T is a fundamental matrix for the associated system of first order linear differential equations. Properties of T that will be used frequently include

$$\text{(i)} \qquad T(t, s)T(s, \xi) = T(t, \xi) \tag{2.2a}$$

and

$$\text{(ii)} \quad \frac{d}{ds}T(t,s) = -T(t,s)M(s). \tag{2.2b}$$

Given the inhomogeneous problem

$$\frac{d\mathbf{x}}{dt} = M\mathbf{x} + \mathbf{g}(t), \qquad \mathbf{x}(\xi) = \mathbf{x}_\xi, \tag{2.3}$$

then knowledge of T permits an explicit form for the solution to be written down. This is

$$\mathbf{x}(t) = T(t,\xi)\mathbf{x}_\xi + \int_\xi^t T(t,s)\mathbf{g}(s)\,ds. \tag{2.4}$$

In the case that $\mathbf{g}$ corresponds to a stochastic forcing term (compare (1.3))

$$\mathbf{g} = \sigma\sqrt{\lambda}\frac{dw}{dt}\mathbf{b} \tag{2.5}$$

where $\mathbf{b}$ is a constant vector, then (2.4) can be written in recurrence relation form as

$$\mathbf{x}(t_{i+1}) = T(t_{i+1},t_i)\mathbf{x}(t_i) + \mathbf{u}(t_{i+1},t_i) \tag{2.6}$$

where

$$\mathbf{u}(t_{i+1},t_i) = \sigma\sqrt{\lambda}\int_{t_i}^{t_{i+1}} T(t_{i+1},s)\mathbf{b}\frac{dw}{ds}\,ds. \tag{2.7}$$

By definition $\mathbf{u}(t_{i+1},t_i)$ is independent of $\mathbf{x}(t)$, $t \le t_i$, and

$$\mathbf{u}(t_{i+1},t_i) \sim N(0,\Omega(t_{i+1},t_i))$$

where

$$\Omega(t_{i+1},t_i) = \sigma^2\lambda\int_{t_i}^{t_{i+1}} T(t_{i+1},s)\mathbf{b}\mathbf{b}^T T(t_{i+1},s)^T\,ds. \tag{2.8}$$

Corresponding to assumption (1.2) we write

$$y_i = \mathbf{h}^T\mathbf{x}(t_i) + \varepsilon_i, \qquad i = 1,2,\ldots,n, \tag{2.9}$$

where (in part to simplify presentation) we take $\mathbf{h}$ to be a constant vector, and where we assume also that the ε_j are independent of $\mathbf{u}(t_{i+1},t_i)$.

The problem considered is that of computing $\mathbf{x}(t|n)$, the conditional expectation of $\mathbf{x}(t)$ given $y_1,y_2,\ldots,y_n$. An estimate of the data is then given by $\mathbf{h}^T\mathbf{x}(t_i|n)$, $i = 1,2,\ldots,n$, and we refer to $\mathbf{h}^T\mathbf{x}(t|n)$ as a generalised smoothing spline. This computation requires a strategy for imposing the initial conditions $\mathbf{x}_\xi$ needed to specify the solution to (2.3); and here two possible approaches are considered.

(i) The vector of initial conditions $\mathbf{x}_0$ is constant. This means that $\mathbf{x}(t)$ can be written in signal-plus-noise form

$$\mathbf{x}(t) = T(t,t_1)\mathbf{x}_0 + \mathbf{z}(t) \tag{2.10}$$

where $z(t)$ provides the stochastic dependence. As x_0 is constant, $z(t)$ satisfies

$$E\{z(t_1)\} = 0, \quad \text{var}\{z(t_1)\} = 0. \tag{2.11}$$

The estimation problem can now be treated by writing the observational equations in the form (subtracting out the signal)

$$y_i - h^T T(t_i, t_1)x_0 = h^T z(t_i) + \varepsilon_i, \qquad i = 1, 2, \ldots, n, \tag{2.12}$$

developing the likelihood for this system, and them maximising the likelihood to estimate x_0, λ and σ^2. This is the program advocated by Wecker and Ansley [24].

(ii) A diffuse prior distribution is assumed for $x(t_1|0)$ and is treated by setting $x(t_1|0) \sim N(0, \gamma^2 I)$ and seeking limits for the conditional expectations as $\gamma^2 \to \infty$. In the particular case of (1.3) this is Wahba's model for smoothing splines.

In both cases, the estimation problem is specified by a dynamics equation (2.6) which describes the evolution of the state vector $x(t)$, and an observation equation (2.9) or (2.12) providing information about the particular realisation of the process. It is a consequence of the assumptions made concerning the independence of the noise terms that the Kalman filter is an appropriate device for computing the conditional expectations. Assume that $x(t_i|i)$, the conditional expectation of $x(t_i)$ given y_j, $j = 1, 2, \ldots, i$, and

$$\text{var}\{x(t_i) - x(t_i|i)\} = S_{i|i} \tag{2.13}$$

are known. Then

$$x(t_{i+1}|i) = T(t_{i+1}, t_i)x(t_i|i) \tag{2.14}$$

is the best available estimate of $x(t_{i+1})$, and the innovation

$$\varsigma_{i+1} = y_{i+1} - h^T x(t_{i+1}|i) \tag{2.15}$$

is uncorrelated with the past. It follows that the projection theorem (see, for example, Luenberger [14]) can be used to give

$$x(t_{i+1}|i+1) = x(t_{i+1}|i) + \frac{\varsigma_{i+1}}{\sigma^2 + h^T S_{i+1|i}h}S_{i+1|i}h \tag{2.16}$$

and

$$S_{i+1|i+1} = S_{i+1|i} - \frac{S_{i+1|i}hh^T S_{i+1|i}}{\sigma^2 + h^T S_{i+1|i}h}, \tag{2.17}$$

where

$$S_{i+1|i} = T(t_{i+1}, t_i)S_{i|i}T(t_{i+1}, t_i)^T + \Omega(t_{i+1}, t_i). \tag{2.18}$$

To estimate the dependence of $x(t_i)$ on all of the data requires also a backward recurrence (the fixed interval smoothing algorithm [1]). For $t_{i-1} \le t \le t_i$ this gives

$$x(t|n) = T(t, t_{i-1})x(t_{i-1}|i-1) + A(t_i, t)(x(t_i|n) - x(t_i|i-1)) \tag{2.19}$$

where

$$A(t_i, t) = \{T(t, t_{i-1})S_{i-1|i-1}T(t_i, t_{i-1})^T + \Gamma(t_i, t)\}S_{i|i-1}^{-1}, \qquad (2.20)$$

and

$$\Gamma(t_i, t) = \Omega(t, t_{i-1})T(t_i, t)^T. \qquad (2.21)$$

The corresponding formula for $S(t|n) = \mathrm{var}\{\mathbf{x}(t) - \mathbf{x}(t|n)\}$ is

$$\begin{aligned} S(t|n) = {}& \Omega(t, t_{i-1}) + T(t, t_{i-1})S_{i-1|i-1}T(t, t_{i-1})^T \\ & - A(t_i, t)(S_{i|i-1} - S_{i|n})A(t_i, t)^T. \end{aligned} \qquad (2.22)$$

Because the ς_i are independent, the output from the Kalman filter permits the log likelihood to be written down. For the Wecker and Ansley model this gives

$$L_{WA} = \sum_{i=1}^{n} L(\varsigma_i) = \sum_{i=1}^{n} \frac{1}{2} \left\{ \frac{\varsigma_i^2}{\sigma^2 + \mathbf{h}^T S_{i|i-1}\mathbf{h}} + \log(\sigma^2 + \mathbf{h}^T S_{i|i-1}\mathbf{h}) \right\} + \mathrm{const} \qquad (2.23)$$

where $L(\varsigma_i)$ is the contribution to the log likelihood by ς_i. A similar formula is available in the case of the Wahba model but it must interpreted in a limiting sense as $\gamma^2 \to \infty$ [13].

3. Smoothness properties

Justification for calling $\mathbf{h}^T\mathbf{x}(t|n)$ a generalised smoothing spline is provided by considering the smoothness properties of $\mathbf{x}(t|n)$. The key formula to investigate is the interpolation formula (2.19). The first step is to differentiate $A(t_i, t)$ which in turn depends on $\Gamma(t_i, t)$ which can be written (using (2.2), (2.8))

$$\begin{aligned} \Gamma(t_i, t) &= \sigma^2\lambda \int_{t_{i-1}}^{t} T(t, s)\mathbf{b}\mathbf{b}^T T(t, s)^T T(t_i, t)^T \, ds \\ &= \sigma^2\lambda \int_{t_{i-1}}^{t} T(t, s)\mathbf{b}\mathbf{b}^T T(t_i, s)^T \, ds. \end{aligned} \qquad (3.1)$$

Differentiating (3.1) gives

$$\frac{d\Gamma}{dt} = M\Gamma + \sigma^2\lambda\mathbf{b}\mathbf{b}^T T(t_i, t)^T \qquad (3.2)$$

from which we obtain

$$\frac{d}{dt}A(t_i, t) = MA + \sigma^2\lambda\mathbf{b}\mathbf{b}^T T^T(t_i, t)S_{i|i-1}^{-1} \qquad (3.3)$$

and

$$\frac{d}{dt}\mathbf{x}(t|n) = M\mathbf{x}(t|n) + \sigma^2\lambda\mathbf{b}\mathbf{b}^T T^T(t_i, t)S_{i|i-1}^{-1}(\mathbf{x}(t_i|n) - \mathbf{x}(t_i|i-1)). \qquad (3.4)$$

An immediate consequence is that between data points $\mathbf{x}(t|n)$ has the same smoothness as solutions of the homogeneous differential equation. It follows that it is the data points which are of most interest in investigating the smoothness of $\mathbf{x}(t|n)$.

REMARK 3.1. Consider the pth order linear differential equation

$$\mathcal{L}_p x_1 = g(t) \tag{3.5}$$

where $g(t)$ is a stochastic forcing term (compare (1.3)). If this is written as a first order system using the standard substitution

$$dx_{i-1}/dt = x_i, \qquad i = 2, \ldots, p \tag{3.6}$$

then the natural choices for $\mathbf{b}$ and $\mathbf{g}$ are

$$\mathbf{b} = \mathbf{e}_p, \qquad \mathbf{g} = g(t)\mathbf{b}. \tag{3.7}$$

Taking the scalar product of (3.4) with $\mathbf{e}_p$ gives

$$\mathcal{L}_p x_1 = \sigma^2 \lambda \mathbf{e}_p^T T^T(t_i, t)\mathbf{d} \tag{3.8}$$

where

$$\mathbf{d} = S_{i|i-1}^{-1}(\mathbf{x}((t_i|n) - \mathbf{x}(t_i|i-1))).$$

But it is immediately verified that

$$\mathcal{L}_p^+ \mathbf{e}_p^T T^T(t_i, t)\mathbf{d} = 0 \tag{3.9}$$

for every $\mathbf{d}$ where $\mathcal{L}_p^+$ is the formal adjoint of $\mathcal{L}_p$, and this shows that between the data points x_1 satisfies

$$\mathcal{L}_p^+ \mathcal{L}_p x_1 = 0. \tag{3.10}$$

This result is well known for smoothing splines. The above derivation is essentially that given in [12].

Continuity of $\frac{d}{dt}\mathbf{x}(t|n)$ at $t = t_i$ requires

$$\mathbf{B}^T S_{i|i-1}^{-1}(\mathbf{x}(t_i|n) - \mathbf{x}(t_i|i-1)) = \mathbf{B}^T T^T(t_{i+1}, t_i)S_{i+1|i}^{-1}(\mathbf{x}(t_{i+1}|n) - \mathbf{x}(t_{i+1}|i)). \tag{3.11}$$

Using (2.19) gives

$$S_{i|i}T(t_{i+1}, t_i)^T S_{i+1|i}^{-1}(\mathbf{x}(t_{i+1}|n) - \mathbf{x}(t_{i+1}|i)) = \mathbf{x}(t_i|n) - \mathbf{x}(t_i|i).$$

Also, inverting both sides of (2.17),

$$S_{i|i}^{-1} = S_{i|i-1}^{-1} + \sigma^{-2}\mathbf{h}\mathbf{h}^T,$$

so that the difference between the two sides of (3.11) becomes

$$D = \mathbf{b}^T \{S_{i|i-1}^{-1}(\mathbf{x}(t_i|i) - \mathbf{x}(t_i|i-1)) - \sigma^{-2}\mathbf{h}\mathbf{h}^T(\mathbf{x}(t|n) - \mathbf{x}(t_i|i))\}.$$

Simplifying using the Kalman filter equation (2.16) gives

$$D = \mathbf{b}^T \mathbf{h} \left\{ \frac{\varsigma_i}{\sigma^2 + \mathbf{h}^T S_{i|i-1} \mathbf{h}} - \frac{1}{\sigma^2} \mathbf{h}^T (\mathbf{x}(t_i|n) - \mathbf{x}(t_i|i)) \right\} \qquad (3.12)$$

and this vanishes provided $\mathbf{b}^T \mathbf{h} = 0$.

To extend this result to higher derivatives note that

(a) it is necessry only to consider derivatives of the term involving $\mathbf{b}\mathbf{b}^T T^T (t_i, t)$ in (3.4) to find the first occurrence of a discontinuity, and

(b) successive derivatives of $T(t_i, t)$ can be represented by

$$d^j T(t_i, t)/dt^j = T(t_i, t) P_j(M) \qquad (3.13)$$

where the P_j satisfy the recurrence

$$P_0 = I, \qquad P_j = dP_{j-1}/dt - MP_{j-1}, \quad j = 1, 2, \ldots. \qquad (3.14)$$

Paralleling the above argument now gives the result that the first k derivatives of $\mathbf{x}(t|n)$ are continuous at t_i provided

$$\mathbf{b}^T P_{j-1}(M)^T \mathbf{h} = 0, \qquad j = 1, 2, \ldots, k. \qquad (3.15)$$

If the successive vectors $P_j(M)^T \mathbf{h}$, $j = 0, 1, \ldots, p-1$ are linearly independent, then at most the first $(p-1)$ derivatives can be continuous, as any vector which is orthogonal to p linearly independent vectors must vanish, leading to a contradiction in (3.15) in the case $p = k$.

REMARK 3.2. Returning to the example discussed in Remark 3.1 note that in this case $P_i(M)$ can be written

$$P_i(M) = \begin{bmatrix} 0 \cdots 0 | I_{p-i} \\ \hline \mathbf{a}_{i1}^T \\ \cdots \\ \mathbf{a}_{ii}^T \end{bmatrix}$$

where the $\mathbf{a}_{ij}$ are functions of the coefficients in $\mathcal{L}_p$. Clearly

$$\begin{aligned} \mathbf{e}_p^T P_i(M)^T \mathbf{e}_1 &= 0, & i = 1, \ldots, p-2, \\ &= 1, & i = p-1, \end{aligned} \qquad (3.16)$$

showing that the choice

$$\mathbf{b} = \mathbf{e}_p, \qquad \mathbf{h} = \mathbf{e}_1$$

makes x_1 together with its first $2p - 2$ derivatives continuous at the data points. The discontinuity in the $(2p - 1)$'st derivative is obtained by evaluating the bracketed term in (3.12). Simplifying using (2.16) gives

$$x_1^{(2p-1)}(t_i+) - x_1^{(2p-1)}(t_i-) = \lambda(y_i - \mathbf{h}^T \mathbf{x}(t_i|n)). \qquad (3.17)$$

A direct comparison with the variational argument given in Reinsch [17] shows

that $x_1(t|n)$ is exactly the L_g smoothing spline provided
   (a) $\lambda = 1/\mu$, and
   (b) the natural boundary conditions

$$\mathcal{L}_p x_1(t|n) = 0, \qquad t < t_i, \ t > t_n \tag{3.18}$$

are satisfied.

But (b) is a consequence of the extrapolation formula (2.14) for $t > t_n$ and also for $t < t_1$ provided that $\mathbf{u}(t, t_1)$ is uncorrelated with both $\mathbf{u}(s, t_1)$ for $s > t_1$ and $\mathbf{x}(t_1|0)$. This condition is satisfied for both the stochastic models considered here. Thus it follows from the uniqueness theorem for L_g splines that both models give the same $\mathbf{x}(t|n)$ for any prescribed value of λ.

REMARK 3.3. An alternative characterization of $\mathbf{b}$ and $\mathbf{h}$ can be obtained by considering the eigen-decomposition of $\Omega(t + \delta, t)$ in the limit as $\delta \to 0$. Expanding $T(t + \delta, s)$ in (2.8) using (3.13) gives

$$\Omega(t + \delta, t) = \sigma^2 \lambda \int_t^{t+\delta} \sum_{i,j} \frac{(s - (t+\delta))^{i+j}}{i! j!} P_i(M) \mathbf{b} \mathbf{b}^T P_j(M)^T \, ds. \tag{3.19}$$

Because successive powers of δ are incommensurable as $\delta \to 0$, an application of (3.16) to the evaluation of the Rayleigh quotient [9] gives:
   (i) The largest eigenvalue of $\Omega(t + \delta, t)$ is

$$\omega_p = \sigma^2 \lambda \delta \mathbf{b}^T \mathbf{b} + O(\delta^2), \qquad \delta \to 0 \tag{3.20}$$

and is associated with an eigenvector which tends to $\mathbf{b}$ as $\delta \to 0$.
   (ii) The smallest eigenvalue of $\Omega(t + \delta, t)$ is

$$\omega_1 = \frac{\sigma^2 \lambda}{((p-1)!)^2} \frac{(\mathbf{h}^T P_{p-1}(M) \mathbf{b})^2}{\mathbf{h}^T \mathbf{h}} \frac{\delta^{2p-1}}{2p-1} + O(\delta^{2p}) \tag{3.21}$$

and is associated with an eigenvector which tends to $\mathbf{h}$ as $\delta \to 0$.
   (iii) The remaining eigenvalues are contained between these asymptotic limits. Any vector asymptotically orthogonal to $\mathbf{b}$ must give a Rayleigh quotient estimate which is $o(\sigma^2 \lambda \delta)$ as $\delta \to 0$ and, in particular, this estimate must hold for ω_{p-1}.

4. Computation

It is not usual for the Kalman filter to be implemented directly in the recursive form (2.14)–(2.18) which can apparently lead to scaling-related numerical problems, and preferred strategies work either with a square root factorisation of the covariance matrix or of its inverse (the information matrix). Our approach has been to use a method of the latter kind due to Paige and Saunders [15] which has the particular advantage that the backward recursion, which gives

the smoothed values $x(t_i|n)$, is an integral part of the algorithm rather than a separate procedure as in the case of the fixed interval smoother.

The starting point for this algorithm is an alternative formulation of the estimation problem given by Duncan and Horn [7]. In this approach the problem becomes a linear least squares problem with independent noise terms having unit variance, and Paige and Saunders suggest the use of orthogonal matrix transformations to take the data matrix to upper triangular form (the standard stable algorithm). However, they also show that the factorisation can be organised to take advantage of the problem structure and to reveal the close relationship between this method and the Kalman filter. In fact the Kalman filter recursion can be deduced from the steps of the orthogonal factorisation and the smoothing procedure can be deduced from the back substitution.

To specify the least squares problem, it is convenient to define

$$\text{var}\{x(t_1) - x(t_1|0)\}^{-1} = L_1 L_1^T, \qquad \Omega(t_i, t_{i-1})^{-1} = L_i L_i^T, \qquad i = 2, \ldots, n,$$
$$\overline{T}_i = L_i^T T(t_i, t_{i-1}), \qquad \overline{u}_i = L_i^T u(t_i, t_{i-1}).$$

The key step is to transform (2.6) by premultiplying by L_i^T to give an equation involving the barred quantities just introduced, in which the noise components $\overline{u}_i$ are uncorrelated and have unit variance. In terms of this notation the required sum of squares is

$$H(x) = \|L_1^T(x(t_1) - x(t_1|0))\|^2 + \sigma^{-2}(y_1 - h^T x(t_1))^2$$
$$+ \sum_{i=2}^{n} \{\|L_i^T x(t_i) - \overline{T}_i x(t_{i-1})\|^2 + \sigma^{-2}(y_i - h^T x(t_i))^2\}. \tag{4.1}$$

Actually our models correspond to the special cases $L_1 = \infty$ (x_0 a constant vector), and $L_1 = 0$, the limiting case as $\gamma^2 \to \infty$ of the prior distribution $x(t_1|0) \sim N(0, \gamma^2 I)$. In the first case, the first two terms are omitted from the sum of squares, but not from the likelihood calculation. However, the case $L_1 = 0$ leads to a singular least squares problem, but it is possible to identify limiting values by running the filter for an increasing sequence of values $\{\gamma_i^2\} \to \infty$. The key result is that $S_{p|p}$ is bounded as $\gamma^2 \to \infty$ provided

$$\dim\{v_i; v_i = T^T(t_i, t_1)h, \ i = 1, 2, \ldots, k\} = k, \qquad k = 1, 2, \ldots, p. \tag{4.2}$$

This result makes it possible to initialise the information filter at $t = t_p$ instead of $t = t_1$ and then to proceed in the usual way to compute $x(t_i|n)$, $i \geq p$. A smoothing step is needed to compute $x(t_i|n)$ for $i < p$.

The required result is demonstrated inductively by showing that for $k \leq p$

$$S_{k|k-1} = T(t_k, t_1)\{\gamma^2(I - Q_{k-1}) + E_k\}T(t_k, t_1)^T \tag{4.3}$$

where $E_k = O(1)$, $\gamma^2 \to \infty$, and Q_k is the orthogonal projection of rank k defined by

$$Q_k V_k = V_k = [v_1, v_2, \ldots, v_k]. \tag{4.4}$$

The result is clearly true when $k = 1$, so it is necessary to verify it for $k := k + 1$ contingent on (4.3). Let

$$\mathbf{w}_k = (I - Q_{k-1})\mathbf{v}_k$$

so that

$$Q_k = Q_{k-1} + \mathbf{w}_k \mathbf{w}_k^T / \|\mathbf{w}_k\|^2.$$

Substituting in (2.17) gives

$$S_{k|k} = T(t_k, t_1) \left\{ \gamma^2(I - Q_{k-1}) + E_k - \frac{(\gamma^2 \mathbf{w}_k + E_k \mathbf{v}_k)(\gamma^2 \mathbf{w}_k + E_k \mathbf{v}_k)^T}{\sigma^2 + \mathbf{v}_k^T \{\gamma^2(I - Q_{k-1}) + E_k\}\mathbf{v}_k} \right\}$$
$$\times T(t_k, t_1)^T$$
$$= T(t_k, t_1)\{\gamma^2(I - Q_k) + R_k\}T(t_k, t_1)^T$$

(4.5)

where, up to terms of order $1/\gamma^2$,

$$R_k = \left(I - \frac{\mathbf{w}_k \mathbf{v}_k^T}{\|\mathbf{w}_k\|^2}\right) E_k \left(I - \frac{\mathbf{w}_k \mathbf{v}_k^T}{\|\mathbf{w}_k\|^2}\right)^T + \frac{\sigma^2}{\|\mathbf{w}_k\|^4} \mathbf{w}_k \mathbf{w}_k^T. \qquad (4.6)$$

Note that $R_k = O(1)$ and is positive definite if E_k is. Now (2.18) gives

$$S_{k+1|k} = T(t_{k+1}, t_k)S_{k|k}T(t_{k+1}, t_k)^T + \Omega(t_{k+1}, t_k)$$
$$= T(t_{k+1}, t_1)\{\gamma^2(I - Q_k) + E_{k+1}\}T(t_{k+1}, t_1)^T$$

where

$$E_{k+1} = R_k + T(t_{k+1}, t_1)^{-1}\Omega(t_{k+1}, t_k)T(t_{k+1}, t_1)^{-T}$$
$$= R_k + \sigma^2\lambda \left\{ \int_{t_1}^{t_{k+1}} T(t_1, s)\mathbf{b}\mathbf{b}^T T(t_1, s)^T \, ds \right.$$

(4.7)

$$\left. - \int_{t_1}^{t_k} T(t_1, s)\mathbf{b}\mathbf{b}^T T(t_1, s)^T \, ds \right\}.$$

This establishes the induction. But now after p steps $Q_p = I$ so $I - Q_p = 0$ and

$$S_{p|p} = T(t_p, t_1)R_p T(t_p, t_1)^T = O(1). \qquad (4.8)$$

To estimate $\lim_{\gamma^2 \to \infty} \mathbf{x}(t_k, k)$ for $k \le p$, assume an expansion having the form

$$\mathbf{x}(t_k|k) = \mathbf{x}^0(t_k|k) + \frac{1}{\gamma^2}\mathbf{x}^1(t_k|k) + \cdots. \qquad (4.9)$$

Substituting in (2.16), setting

$$\mathbf{x}^0(t_k|k) = T(t_k, t_1)\xi_k, \qquad (4.10)$$

and collecting leading terms gives the recurrence

$$\xi_k = \xi_{k-1} + (y_k - \mathbf{v}_k^T \xi_{k-1})\mathbf{w}_k / \|\mathbf{w}_k\|^2$$
$$= (I - \mathbf{w}_k \mathbf{v}_k^T / \|\mathbf{w}_k\|^2)\xi_{k-1} + y_k \mathbf{w}_k / \|\mathbf{w}_k\|^2$$

(4.11)

which should be compared with (4.6).

REMARK 4.1. Equation (4.11) can also be written

$$\xi_k = (V_k^T)^+ \mathbf{y}_k \tag{4.12}$$

where $\mathbf{e}_i^T \mathbf{y}_k = y_i$, $i = 1, 2, \ldots, k$. It is important to note that ξ_k, Q_k can be built up using a stable algorithm, for example by transforming the successive V_k to upper triangular form using Householder transformations. This would appear to offer numerical advantages over calculations which essentially deflate $I - Q_k$ in recursive calculations. A procedure of this kind is given by Ansley and Kohn [2].

REMARK 4.2. Note that both $\mathbf{x}^0(t_p|p)$ and $\mathrm{var}\{\mathbf{x}(t_p) - \mathbf{x}^0(t_p|p)\}$ can be computed directly as V_p is invertible by assumption—the variance calculation uses, for $k = 1, 2, \ldots, p$,

$$-\varepsilon_k = \mathbf{v}_k^T T(t_p, t_1)^{-1} \left\{ \mathbf{x}(t_p) - \mathbf{x}^0(t_p|p) + \sigma\sqrt{\lambda} \int_{t_p}^{t_k} T(t_p, s)\mathbf{b}\frac{dw}{ds}\, ds \right\}.$$

This corresponds to the initialisation used by Weinert, Byrd, and Sidhu [25].

The limiting form of the smoothing recursion for $i < p$ is obtained in similar fashion. Consider

$$A(t_{i+1}, t_i) = S_{i|i} T(t_{i+1}, t_i) S_{i+1|i}^{-1}$$
$$= T(t_i, t_1)\{\gamma^2(I - Q_i) + R_i\}\{\gamma^2(I - Q_i) + E_{i+1}\}^{-1} T(t_{i+1}, t_1)^T. \tag{4.13}$$

The term involving γ^2 can be written (in an obvious notation)

$$\{\gamma^2(I - Q_i) + R_{i0} + \gamma^{-2}R_{i1} + \cdots\}$$
$$\times \{\gamma^2(I - Q_i) + E_{(i+1)0} + \gamma^{-2}E_{(i+1)1} + \cdots\}^{-1}$$
$$= \{\gamma^2(I - Q_i) + R_{i0} + \gamma^{-2}R_{i1} + \cdots\}\{X_0 + \gamma^{-2}X_1 + \cdots\} \tag{4.14}$$
$$= I + \Delta_i X_0 + O(1/\gamma^2)$$

where $\Delta_i = R_{i0} - E_{(i+1)0}$, and X_0 satisfies

$$X_0 = Q_i X_0 Q_i, \qquad Q_i E_{(i+1)0} X_0 = Q_i.$$

The corresponding recursion for the smoothed variance (2.22) requires a knowledge of the $O(1/\gamma^2)$ terms in (4.14) in the intermediate calculations. The contribution from the unbounded terms is (after some manipulation)

$$T(t_i, t_1)^{-1}\{S_{i|i} - A(t_{i+1}, t_i)S_{i+1|i}A(t_{i+1}, t_i)^T\}T(t_i, t_1)^{-T}$$
$$= -\Delta_i - \Delta_i X_0 \Delta_i + O(1/\gamma^2). \tag{4.15}$$

Once again there is a well determined limit as $\gamma^2 \to \infty$.

REMARK 4.3. For the class of problems discussed in Remarks 3.1, 3.2 there exists an alternative to using these limiting forms of the smoothing recursions to calculate $\mathbf{x}(t|n)$, $t < t_p$. The idea is to use the knowledge of the discontinuity

in the $(2p-1)$st derivative at the data points to successively generate terminal conditions for $\mathbf{x}(t|n)$ at $t_i = p, p-1, \ldots, 2$, and then to integrate (3.10) back to t_{i-1}. In this case the smoothing formula for interpolating $\mathbf{x}(t|n)$ between the data points provides an explicit solution to the boundary value problem for (3.10) with the boundary data $\mathbf{x}(t_{i-1}|n)$, $\mathbf{x}(t_i|n)$.

5. Implications

Thus far we have shown that the stochastic models can be estimated using what are regarded as very satisfactory numerical procedures. But this is not sufficient to guarantee a successful computation—for this it is necessary to known something of the intrinsic sensitivity of this approach to the estimation problem itself. This information is sought in this section.

Our analysis is based on the assumptions that $\delta \to 0$, and $n\delta$ is bounded as $n \to \infty$. This corresponds to sampling on finer and finer subintervals of a fixed base interval and seems appropriate enough if the aim is to smooth a given set of data. It need not correspond to sampling with fixed δ unless δ is small in some natural scale.

Following Golub and Van Loan [9], the sensitivity analysis appropriate to the linear least squares problem

$$\min_{\mathbf{x}} \|\mathbf{r}\|^2; \qquad \mathbf{r} = C\mathbf{x} - \mathbf{c}$$

leads to the basic perturbation result

$$\|\bar{\mathbf{x}} - \mathbf{x}\|/\|\mathbf{x}\| \le \varepsilon\{2\operatorname{cond}\{C\}/\cos\theta + \tan\theta\operatorname{cond}\{C\}^2\} + O(\varepsilon^2) \qquad (5.1)$$

where $\bar{\mathbf{x}}$ is the solution to the perturbed problem, ε is a bound for the relative perturbations to C, $\mathbf{c}$, $\sin\theta = \|\mathbf{r}\|/\|\mathbf{c}\|$, and $\operatorname{cond}\{C\} = \pi_p/\pi_1$, the ratio of largest to smallest singular values of C. The application of this result to the Paige and Saunders algorithm needs an estimate of the extreme singular values of the data matrix $\tilde{H}$ associated with (4.1). Two further assumptions are made.
(i) It is assumed for simplicity that $t_{i+1} - t_i = \delta = 1/n - 1$, $i = 1, 2, \ldots, n-1$.
(ii) It is necessary to have an idea of the asymptotic behaviour of the scale parameter λ. Some information is available ([20], [23]), and it would appear to be compatible with this to assume that $\lambda \to 0$ like δ^ϕ, $0 < \phi < 1$ as $\delta \to 0$.

That this assumption is tenable can be seen by considering the extreme values of ϕ. When $\phi = 0$ the approximation term must dominate (1.1) for large n unless the spline tends to interpolate the data, while when $\phi = 1$ the smoothing term matches the growth of the approximation term for a general signal plus noise model. Numerical experience is reported in Section 5.

The following sequences prove useful. These are

$$\alpha_1 = \mathbf{h}, \quad \alpha_{i+1} = T(t_{i+1}, t_i)\alpha_i + \mathbf{h}, \qquad i = 1, 2, \ldots, n-1, \qquad (5.2)$$

$$\beta_1 = \mathbf{b}, \quad \beta_{i+1} = T(t_{i+1}, t_i)\beta_i + \mathbf{b}, \qquad i = 1, 2, \ldots, n-1, \qquad (5.3)$$

and they have the property that

$$\alpha_i \sim n\overline{\alpha}(t_i) + O(1), \qquad \overline{\alpha}(t) = \int_{t_1}^{t} T(t, s)\mathbf{h}\, ds, \qquad (5.4)$$

$$\beta_i \sim n\overline{\beta}(t_i) + O(1), \qquad \overline{\beta}(t) = \int_{t_1}^{t} T(t, s)\mathbf{b}\, ds. \qquad (5.5)$$

The idea is to estimate the Rayleigh quotient associated with the data matrix using the results already obtained concerning the extreme eigenvalues of $\Omega(t + \delta, t)$. The Rayleigh quotient is

$$RQ(\mathbf{x}) = \overline{H}(\mathbf{x}) \Big/ \sum_{i=1}^{n} \|\mathbf{x}_i\|^2 \qquad (5.6)$$

where

$$\overline{H}(\mathbf{x}) = \|L_1^T \mathbf{x}_1\|^2 + \sigma^{-2}(\mathbf{h}^T \mathbf{x}_1)^2 + \sum_{i=2}^{n} \left\{ \|L_i^T \mathbf{x}_i - \overline{T}_i \mathbf{x}_{i-1}\|^2 + \sigma^{-2}(\mathbf{h}^T \mathbf{x}_i)^2 \right\}. \tag{5.7}$$

Large values of RQ are obtained by exciting the smallest eigenvalue ω_1 of $\Omega(t + \delta, t)$ which has been estimated by (3.21). This can be done by choosing $\mathbf{x} = \alpha$ which picks up these small values for each i by (5.2) and gives

$$\overline{H}(\alpha) = \sum_{i=1}^{n} \left\{ \omega_1^{-1}(t_i) \|\mathbf{h}\|^2 (1 + O(\delta)) + \frac{1}{\sigma^2}(\mathbf{h}^T \alpha_i)^2 \right\}. \qquad (5.8)$$

Also

$$\sum_{i=1}^{n} \alpha_i^T \alpha_i \sim n^2 \sum_{i=1}^{n} \|\overline{\alpha}(t_i)\|^2 \sim n^3 \int_{t_1}^{t_n} \|\overline{\alpha}(s)\|^2\, ds, \qquad (5.9)$$

so that

$$RQ(\alpha) = O(\lambda^{-1}\delta^{-2p+3}) + O(1).$$

A second possibility is to choose $\mathbf{x}$ so that the small eigenvalues are excited in only a small number of terms in (5.7) and the $O(n^3)$ growth in the normalising term (5.9) avoided. The simplest possibility corresponds to the choice $\mathbf{x}_i = \mathbf{h}$, $\mathbf{x}_j = 0$, $j \neq i$ and gives

$$RQ(\mathbf{x}) = O(\lambda^{-1}\delta^{-2p+1}) + O(1). \qquad (5.10)$$

This estimate beats that above and it is clear that a larger order is not possible.

To obtain small values of RQ it is necessary to choose $\mathbf{x}$ to stimulate only the largest eigenvalue of $\Omega(t + \delta, t)$. In this case the choice $\mathbf{x}_i = \mathbf{b}$, $\mathbf{x}_j = 0$, $j \neq i$ does not work. It gives

$$\overline{H}(\mathbf{x}) = \omega_p^{-1}(t_i)\|\mathbf{b}\|^2(1 + O(\delta)) + \|\overline{T}_{i+1}\mathbf{b}\|^2 + O(1)$$

and in the second term other eigenvalues (exception ω_1) will be excited. It follows that the smallest $\overline{H}(\mathbf{x})$ can be, with this choice of $\mathbf{x}$, is $O(\delta^{-1}/\lambda)$. The alternative is to excite the large eigenvalues for each i by taking $\mathbf{x} = \boldsymbol{\beta}$. This gives

$$\overline{H}(\boldsymbol{\beta}) = \sum_{i=1}^{n}\{\omega_p^{-1}(t_i)\|\mathbf{b}\|^2(1 + O(\delta)) + \sigma^{-2}(\mathbf{h}^T\boldsymbol{\beta}_i)^2\}. \tag{5.11}$$

Now, by (5.3)
$$\mathbf{h}^T\boldsymbol{\beta}_{i+1} = \mathbf{h}^T\boldsymbol{\beta}_i + \delta\mathbf{h}^T M(t_i)\boldsymbol{\beta}_i + O(\delta^2\boldsymbol{\beta}_i)$$

$$\sim \sum_{j=1}^{i}\delta\mathbf{h}^T M(t_j)n\int_0^{t_j} T(t_j, s)\mathbf{b}\,ds$$

$$\sim \int_0^{t_i} \mathbf{h}^T M(t)\int_0^t T(t, s)\mathbf{b}\,ds.$$

It follows that the second term in $\overline{H}(\boldsymbol{\beta})$ contributes a term which is $O(n^3)$. But we also have

$$\sum_{i=1}^{n}\boldsymbol{\beta}_i^T\boldsymbol{\beta}_i \sim n^2\sum_{i=1}^{n}\|\overline{\boldsymbol{\beta}}(t_i)\|^2 \sim n^3\int_{t_i}^{t_n}\|\overline{\boldsymbol{\beta}}(s)\|^2\,ds$$

so that
$$RQ(\boldsymbol{\beta}) = O(\delta/\lambda) + O(1) = O(1). \tag{5.12}$$

Again it is clear that a smaller order is not possible.

To use these results in the sensitivity inequality requires also a knowledge of θ. Here the assumptions of independence and unit variance give (for n effective degrees of freedom)

$$\sin\theta \sim \frac{\sqrt{n}}{\sqrt{\sum_{i=1}^{n}(y_i^2/\sigma_p^2)}} \sim \left\{\frac{1}{\sigma^{-2}\int_{t_1}^{t_n}\eta(t)^2\,dt + 1}\right\}^{1/2} \tag{5.13}$$

and this tends to a limit > 0 as $n \to \infty$. It follows from (5.10), (5.12) that $\text{cond}\{\tilde{H}\} \to \infty$, $\delta \to 0$ so that (5.1) must eventually be dominated by the term involving the square of the condition number. This means that asymptotically condPS, the condition number of the Paige Saunders algorithm, satisfies

$$\text{condPS} = O(\delta^{-2p+1}/\lambda), \tag{5.14}$$

provided the assumption regarding λ is correct.

REMARK 5.1. The occurrence of the square of the condition number of $\tilde{H}$ when $\delta \to 0$ suggests that this may give a generic dependence on n for sensitivity estimates of stable implementations of the Kalman filter applied to this problem which use some form of local scaling based on $\Omega(t_{i+1}, t_i)^{1/2}$. It may be that the local scaling makes a significant contribution to (5.14) as it is clearly an important aspect of the estimate (5.10) which, as noted above, is able to beat $RQ(\alpha)$ because of its purely local character. A method which avoids this explicit local scaling is being investigated.

The main algorithm used to construct smoothing splines is the Reinsch algorithm, and this generalises readily to give smoothed estimates of the observed data also for the somewhat more general model considered here. The argument follows Kohn and Ansley [13]. Let

$$\Delta_i^P = [\Delta_{i1}^P, \dots, \Delta_{i(p+1)}^P] \tag{5.15}$$

be defined by the conditions

$$\sum_{j=1}^{p+1} \Delta_{ij}^P \mathbf{h}^T T(t_{i+j-1}, t_i) = 0, \qquad \sum_{j=1}^{p+1} |\Delta_{ij}^P| = 1. \tag{5.16}$$

Then, using (2.4) with $\xi = t_i$,

$$z_i = \Delta_i^P y_i = \Delta_i^P \eta(t_i) + \Delta_i^P \varepsilon_i \tag{5.17}$$

is independent of $\mathbf{x}(t_i)$, and hence of the particular model of the initial conditions. The projection theorem gives the best estimate $\hat{\boldsymbol{\eta}}$ of the signal given the data as

$$\hat{\boldsymbol{\eta}} = \mathbf{y} - \text{cov}(\boldsymbol{\varepsilon}, \mathbf{z}) \text{var}\{\mathbf{z}\}^{-1} \mathbf{z}, \tag{5.18}$$

and in the particular case of the smoothing spline this is exactly the vector of values of the spline at the data points [13]. To evaluate the quantities in (5.18) let $J: R^n \to R^{n-p}$ be the upper triangular matrix formed from the Δ_i^P shifted so that $J_{ii} = \Delta_{i1}^P$. Then

$$\text{cov}\{\boldsymbol{\varepsilon}, \mathbf{z}\} = \sigma^2 J^T, \tag{5.19}$$
$$\text{var}\{\mathbf{z}\} = \sigma^2 J J^T + \sigma^2 \lambda \delta^{2p-1} V, \tag{5.20}$$

where V is the scaled covariance matrix of the vector $\mathbf{q}$ with components defined by

$$q_i = \sum_{j=2}^{p+1} \Delta_{ij}^P \mathbf{h}^T \int_{t_i}^{t_{i+j-1}} T(t_{i+j-1}, s) \mathbf{b} \frac{dw}{ds} \, ds. \tag{5.21}$$

There can be a contribution to $E\{\mathbf{q}\mathbf{q}^T\}$ only when intervals of integration in (5.21) overlap. It follows that V is a $2p - 1$ banded matrix. Also, as $\mathbf{h}$ is asymptotically the eigenvector associated with the smallest eigenvalue of

$\Omega(t+\delta,t)$ it follows that the nonzero elements in V are $O(1)$ as $\delta \to 0$. In simple examples they are also positive.

Implementations of the Reinsch algorithm (for example [5]) appear to construct and factorise var$\{z\}$ explicitly. Thus the obvious quantity to consider in attaching a condition number to this method is cond$\{$var$\{z\}\}$. To estimate this we note first that there is a strong analogy between (5.16) and the definition of divided differences. Let $D: R^{\dot{p}} \to R^p$ be nonsingular. Then

$$\Psi(t)^T = \mathbf{h}^T T(t, t_1) D \tag{5.22}$$

is a vector of linearly independent functions which are annihilated by Δ_i^p, $i = 1, 2, \ldots, n - p$. Sample values of the Ψ_j, $j = 1, 2, \ldots, p$, span the null space of $J^T J$, and any sufficiently smooth function $m(t)$ can be written as a linear combination of the Ψ_j plus a remainder term and it is then easy to show (Householder [11]) that

$$\sum_{j=1}^{p+1} \Delta_{ij}^p m(t_{i+j-1}) = O(\delta^p). \tag{5.23}$$

An immediate consequence is that $J^T J$ has a nonzero eigenvalue which is $O(\delta^{2p})$. But the nonzero eigenvalues of $J^T J$ are the eigenvalues of $J J^T$. Thus the Rayleigh quotient of var$\{z\}$ can be made small of order $\max\{\lambda \delta^{2p-1}, \delta^{2p}\}$ at most. Worked examples show that the correct order is $O(\lambda \delta^{2p-1})$, and this is to be expected because if V has positive elements then its contributions to the Rayleigh quotient will be small for the highly oscillatory eigenvectors of var$\{z\}$. But these eigenvalues give $O(1)$ contributions from $J J^T$ (case of $m(t)$ varying rapidly) and so are not relevant in estimating small eigenvalues.

The above argument indicates that the condition number of the algorithm (at least in its usual implementation) is given by

$$\text{cond}\{\text{var}\{z\}\} = O(\delta^{-2p+1}/\lambda). \tag{5.24}$$

However, there are other possibilities which perhaps should be considered. For example, cov$\{\varepsilon, z\}$var$(z)^{-1}J$ can be interpreted as the northwest corner of a projection matrix which can be computed using the orthogonal transformation determined by the QR factorisation of $[J | \sqrt{\lambda} \delta^{p-1/2} V^{1/2}]^T$.

REMARK 5.2. It follows from (5.14) and (5.24) that the Paige and Saunders and Reinsch algorithms have condition number estimates of the same order as $\delta \to 0$ and it is easily seen that this does not depend on the particular assumptions concerning λ. Thus the analysis presented here does not discriminate between them on the grounds of stability. More disturbing is the actual order in δ given by this sensitivity analysis, which suggests that both algorithms must be considered suspect for values of p as small as 5 (say) and modest values of n. This point is illustrated in Table 5.1 which gives estimates of the rounding error

in results obtained by applying the Paige and Saunders algorithm to be Gallant data [8] (here $n = 72$), and the sunspot data [16, p. 487], (here $n = 176$) for a range of values of p and λ. The values of λ have been chosen to straddle the values given by maximising the likelihood (2.23) as in [24]. For the Gallant data the maximising λ decreased rapidly with p (from $O(10^{-2})$ for $p = 2$ to $O(10^{-6})$ for $p = 4$). As this implies increasingly heavy smoothing it suggests overfitting for the higher values of p. For the sunspot data, values of the maximising λ were much larger (varying from 20.1 for $p = 2$ to 1.1 for $p = 4$), and in this case the spline is trying hard to interpolate the data values. Thus these examples illustrate two rather different types of behaviour. The estimates are obtained by computing $\{n^{-1} \sum_{i=1}^{n} \Delta\hat{\eta}(t_i)^2\}^{1/2}$ where $\Delta\hat{\eta}(t_i)$ is the difference between estimates of the signal computed in single and double precision on a Univac 1182 computer. The results clearly show a deterioration in accuracy as p is increased.

TABLE 5.1. Error estimates for fitted values.

p	λ	Gallant data	λ	Sunspot data
1	.1	$.13 \times 10^{-7}$	100	$.11 \times 10^{-6}$
	.01	$.21 \times 10^{-7}$	10	$.13 \times 10^{-6}$
	.001	$.67 \times 10^{-8}$	1	$.27 \times 10^{-6}$
	.0001	$.21 \times 10^{-8}$	.1	$.52 \times 10^{-6}$
2	.1	$.37 \times 10^{-7}$	100	$.17 \times 10^{-5}$
	.01	$.62 \times 10^{-7}$	10	$.14 \times 10^{-5}$
	.001	$.13 \times 10^{-6}$	1	$.18 \times 10^{-4}$
	.0001	$.24 \times 10^{-6}$	.1	$.24 \times 10^{-4}$
3	.1	$.25 \times 10^{-6}$	100	$.72 \times 10^{-3}$
	.01	$.72 \times 10^{-5}$	10	$.17 \times 10^{-2}$
	.001	$.80 \times 10^{-5}$	1	$.33 \times 10^{-3}$
	.0001	$.10 \times 10^{-4}$	.1	$.38 \times 10^{-3}$
4	.1	$.27 \times 10^{-3}$	100	$.41 \times 10^{-1}$
	.01	$.40 \times 10^{-3}$	10	$.11 \times 10$
	.001	$.89 \times 10^{-4}$	1	$.88$
	.0001	$.52 \times 10^{-3}$	.1	$.33 \times 10$
5	.1	$.25 \times 10^{-1}$		
	.01	$.28$		
	.001	$.28 \times 10^{-1}$		
	.0001	$.26 \times 10^{-1}$		

6. Acknowledgement

7. References

[1] B. D. O. Anderson and J. B. Moore, *Optimal filtering* (Prentice-Hall, Englewood Cliffs, 1979).
[2] C. F. Ansley and R. Kohn, "Estimation, filtering and smoothing in state space models with incompletely specified initial conditions", *Ann. Statist.* **13** (1985) 1286–1316.
[3] C. F. Ansley and R. Kohn, "On the equivalence of two stochastic approaches to spline smoothing", *J. Appl. Prob. Special volume* **23A** (1986) 391–406.
[4] P. Billingsley, *Probability and measure* (Wiley, 1979).
[5] C. de Boor, *A practical guide to splines* (Springer-Verlag, 1978).
[6] F. de Hoog and M. F. Hutchinson, "Two fast procedures for calculating smoothing splines", Report CMA-R27-85, Centre for Mathematical Analysis, Australian National University, 1985.
[7] D. B. Duncan and S. D. Horn, "Linear dynamic recursive estimation from the viewpoint of regression analysis", *J. Amer. Statist. Assoc.* **67** (1972) 815–821.
[8] T. M. Gerig and A. R. Gallant, "Computing methods for linear models subject to linear constraints", *J. Statist. Comput. Simulation* **3** (1975) 283–296.
[9] G. H. Golub and C. F. Van Loan, *Matrix computations* (Johns Hopkins University Press, 1983).
[10] A. C. Harvey and G. D. A. Phillips, "Maximum likelihood estimation of regression models with autoregressive-moving average disturbances", *Biometrika* **66** (1979) 49–58.
[11] A. S. Householder, *Principles of numerical analysis* (McGraw-Hill, 1953).
[12] R. Kohn and C. F. Ansley, "On the smoothness properties of the best linear unbiased estimate of a stochastic process observed with noise", *Ann. Statist.* **11** (1983) 1011–1017.
[13] R. Kohn and C. F. Ansley, "A new algorithm for spline smoothing based on smoothing a stochastic process", Working Paper 85-006, Australian Graduate School of Management, 1985.
[14] D. G. Luenberger, *Optimization by vector space methods* (Wiley, 1969).
[15] C. C. Paige and M. A. Saunders, "Least squares estimation of discrete linear dynamic systems using orthogonal transformations", *SIAM J. Numer. Anal.* **14** (1977) 180–193.
[16] S. M. Pondit and S. M. Wu, *Time series and system analysis with application* (Wiley, 1983).
[17] C. H. Reinsch, "Smoothing by spline functions", *Numer. Math.* **10** (1967) 177–183.
[18] B. W. Silverman, "A fast and efficient cross-validation method for smoothing parameter choice in spline regression", *J. Amer. Statist. Assoc.* **79** (1984) 584–589.
[19] B. W. Silverman, "Some aspects of the spline smoothing approach to nonparametric regression curve fitting", *J. Roy. Statist. Soc. Ser. B.* **47** (1985) 1–52.
[20] P. Speckman, "Spline smoothing and optimal rates of convergence in nonparametric regression models", *Ann. Statist.* **13** (1985) 970–983.
[21] D. F. Utreras, "Sur le choix du parametre d'adjustement dans le lissage par fonctions spline", *Numer. Math.* **34** (1980) 15–28.

[22] G. Wahba, "Improper priors, spline smoothing, and the problem of guarding against model errors in regression", *J. Roy. Statist. Soc. Ser. B.* **40** (1978) 364–372.

[23] G. Wahba, "A comparison of GCV and GML for choosing the smoothing parameter in the generalised spline smoothing problem", *Ann. Statist.* **13** (1985) 1378–1402.

[24] W. Wecker and C. F. Ansley, "The signal extraction approach to nonlinear regression and spline smoothing", *J. Amer. Statist. Assoc.* **78** (1983) 81–89.

[25] H. L. Weinert, R. H. Byrd and G. S. Sidhu, "A stochastic framework for recursive computation of spline functions: Part II, smoothing splines", *J. Optim. Theory Appl.* **30** (1980) 255–268.