SMOOTHNESS AND CONDITIONING
IN GENERALISED SMOOTHING SPLINE CALCULATIONS

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

Abstract

We consider a generalisation of the stochastic formulation of smoothing splines, and discuss the smoothness properties of the resulting conditional expectation (generalised smoothing spline), and the sensitivity of the numerical algorithms. One application is to the calculation of smoothing splines with less than the usual order of continuity at the data points.

1. Introduction

Let data y_i corresponding to the signal plus noise model

$$y_i = \mathbf{h}^T \mathbf{x}(t_i) + \varepsilon_i, \qquad i = 1, 2, \dots, n, \tag{1.1}$$

be given, where $\mathbf{h}, \mathbf{x}(t_i) \in R^p$ and $y_i, \varepsilon_i \in R$. Here, for simplicity, $\mathbf{h}$ is constant and the data points t_i equispaced ($t_i = (i-1)/(n-1) = (i-1)\delta$). Also the $\varepsilon_i \sim N(0, \sigma^2)$ are assumed to be independent, while $\mathbf{x}(t)$ satisfies the stochastic differential equation

$$\frac{d\mathbf{x}}{dt} = M(t)\mathbf{x} + \sigma\sqrt{\lambda}\frac{d\mathbf{w}}{dt}, \tag{1.2}$$

where $M: R^p \to R^p$ is assumed smooth enough for our purposes, $\mathbf{w}(t)$ is a Wiener process independent of the $\varepsilon_i, i = 1, 2, \dots, n$, satisfying, for positive semi-definite $V: R^p \to R^p$ given,

$$E\{(w_i(t+\delta) - w_i(t))(w_j(t+\delta) - w_j(t))\} = V_{ij}\delta \tag{1.3}$$

[1]Dept. of Statistics, Institute of Advanced Studies, Australian National University, GPO Box 4, Canberra, A.C.T. 2601, Australia.
[2]Dept. of Pure and Applied Mathematics, Washington State University, Pullman 99164-2930, WA, USA.

and λ serves to define the relative scale of the two noise processes. We note that, although determining λ is often an important part of the estimation problem, it is not considered in detail here. Our aim is to compute the conditional expectation $\mathbf{x}(t\,|\,n)$ of $\mathbf{x}(t)$ given the data (1.1), and to explore the relations between:

(i) the smoothness attainable in $\mathbf{x}(t\,|\,n)$,

(ii) the role of the control quantities $\mathbf{h}$ and V, and

(iii) the conditioning of the algorithms for computing $\mathbf{x}(t\,|\,n)$.

Let $T(t, \xi)$ be the fundamental matrix defined by

$$dT/dt = MT, \qquad T(\xi, \xi) = I. \tag{1.4}$$

Then the solution to (1.2) satisfying the initial condition $\mathbf{x}(\xi) = \mathbf{x}_\xi$ can be written

$$\mathbf{x}(t) = T(t, \xi)\mathbf{x}_\xi + \sigma\sqrt{\lambda} \int_\xi^t T(t, s)\frac{d\mathbf{w}}{ds}ds. \tag{1.5}$$

It follows that we can associate with the observational equations (1.1) the state-space dynamics equations

$$\mathbf{x}(t_{i+1}) = T(t_{i+1}, t_i)\mathbf{x}(t_i) + \mathbf{u}(t_{i+1}, t_i), \tag{1.6}$$

where (using the assumption that $\mathbf{w}(t)$ is a Wiener process)

$$\mathbf{u}(t_{i+1}, t_i) = \sigma\sqrt{\lambda} \int_{t_i}^{t_{i+1}} T(t_{i+1}, s)\frac{d\mathbf{w}}{ds}ds \sim N(0, \Omega(t_{i+1}, t_i)), \tag{1.7}$$

and

$$\Omega(t_{i+1}, t_i) = \sigma^2\lambda \int_{t_i}^{t_{i+1}} T(t_{i+1}, s)VT(t_{i+1}, s)^T ds. \tag{1.8}$$

An immediate consequence of the formulation (1.1), (1.6)–(1.8) is that the Kalman filter can be used to compute conditional expectations. Assume that $\mathbf{x}(t_i\,|\,i)$, the conditional expectation of $\mathbf{x}(t_i)$ given $y_j, j = 1, 2, \ldots, i$, and

$$\text{var}\{\mathbf{x}(t_i) - \mathbf{x}(t_i\,|\,i)\} = S_{i|i} \tag{1.9}$$

are known; problems involved in the specification of the initial conditions are discussed in [7]. Then

$$\mathbf{x}(t_{i+1}\,|\,i) = T(t_{i+1}, t_i)\mathbf{x}(t_i\,|\,i) \tag{1.10}$$

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

$$\varsigma_{i+1} = y_{i+1} - \mathbf{h}^T\mathbf{x}(t_{i+1}\,|\,i) \tag{1.11}$$

is uncorrelated with the past. It follows that the projection theorem (for example, Luenberger [6]) can be used to give

$$\mathbf{x}(t_{i+1}\,|\,i+1) = \mathbf{x}(t_{i+1}\,|\,i) + \frac{\varsigma_{i+1}}{\sigma^2 + \mathbf{h}^T S_{i+1\,|\,i}\mathbf{h}} S_{i+1\,|\,i}\mathbf{h}, \tag{1.12}$$

and

$$S_{i+1|i+1} = S_{i+1|i} - \frac{S_{i+1|i}\mathbf{h}\mathbf{h}^T S_{i+1|i}}{\sigma^2 + \mathbf{h}^T S_{i+1|i}\mathbf{h}}, \tag{1.13}$$

where

$$S_{i+1|i} = T(t_{i+1}, t_i)S_{i|i}T(t_{i+1}, t_i)^T + \Omega(t_{i+1}, t_i). \tag{1.14}$$

A backward recurrence called the fixed-point smoothing algorithm [10] is needed to incorporate the dependence on all the data. For $t_{i-1} \leq t \leq t_i$, this gives

$$\mathbf{x}(t\,|\,n) = T(t, t_{i-1})\mathbf{x}(t_{i-1}\,|\,i-1) + A(t_i, t)(\mathbf{x}(t_i\,|\,n) - \mathbf{x}(t_i\,|\,i-1)), \tag{1.15}$$

where

$$A(t_i, t) = \{T(t, t_{i-1})S_{i-1|i-1}T(t_i, t_{i-1})^T + \Gamma(t_i, t)\}S_{i|i-1}^{-1}, \tag{1.16}$$

and

$$\Gamma(t_i, t) = \Omega(t, t_{i-1})T(t_i, t)^T. \tag{1.17}$$

This recursion reduces to the "Discrete-time Fixed-interval Smoother" when $t = t_i$. The derivation of this smoother given in [1] can be extended without difficulty to give (1.15)–(1.17).

Several important algorithms for carrying out the Kalman filter numerically (essentially the class of information filter algorithms) make use of a local rescaling based on a square-root factorisation of $\Omega(t_{i+1}, t_i)$ as an important step. The method of Paige and Saunders [9] is one such procedure. It makes use of an equivalent least-squares formulation of the estimation problem due to Duncan and Horn [3], and applies stable computing procedures based on orthogonal transformations to this least-squares form of the problem. This approach was followed in [7], where it is shown that if

$$V = \mathbf{b}\mathbf{b}^T, \tag{1.18}$$

$$\mathbf{v}_1 \rightarrow \mathbf{h}, \delta \rightarrow 0, \tag{1.19}$$

where $\mathbf{v}_1$ is the eigenvector associated with the smallest eigenvalue of $\Omega(t + \delta, t)$, and

$$\mathbf{b}^T P_i(M)\mathbf{h} = 0, \qquad i = 0, 1, \ldots, p - 2, \tag{1.20}$$

where

$$P_0(M) = I, \qquad P_i(M) = dP_{i-1}/dt - MP_{i-1}, \qquad i = 1, 2, \ldots \tag{1.21}$$

then

(i) $\mathbf{x}(t\,|\,n)$ has $p - 1$ continuous derivatives,

(ii) the condition number of the Paige and Saunders algorithm is $O(\delta^{-2p+1}/\lambda)$, and

(iii) this condition number is of the same order in δ as the condition number of the standard implementation of the generalised Reinsch algorithm [7].

Here the smoothness result is generalised to our more general situation, and interest centres on the relation of $\mathbf{h}$ to the eigendecomposition of V. We also consider whether it is possible to improve the condition-number results. Our point of departure is that these results are both disappointing and somewhat surprising. The Reinsch algorithm is analogous to using a high-order difference approximation to integrate a high-order differential equation for a linear functional of x, and it is now widely appreciated that this approach leads to larger condition numbers than do methods which work with the corresponding first-order system directly. Here the right analogue of the first-order system would seem to be the Kalman filter applied to (1.1), (1.6). We show that the problem is in the local scaling, by using an alternative approach suggested by Paige [8] for the generalised least-squares problem, which overcomes the bulk of the difficulty.

If this device is applied to the full least-squares problem then it appears difficult to maintain sparsity, but the resulting algorithm does retain the main advantage of the Paige and Saunders method, in that it provides the smoothed values $\mathbf{x}(t_i \,|\, n)$, $i = 1, 2, \ldots, n$, without the need for a separate smoothing step. The order of magnitude of the condition number for this modified procedure can be estimated, and proves to be both independent of p and almost always more favourable than the estimates obtained in [7]. Paige's device can also be applied to generate a recursive algorithm which proves to be a fairly standard square-root covariance form of the Kalman filter. This would appear to support the claim of superiority made for the covariance form over the information form of the Kalman filter made by Ansley and Kohn [2] in similar applications.

2. Smoothness properties

In this section the dependence of the smoothness of $\mathbf{x}(t \,|\, n)$ on $\mathbf{h}$ and V is explored.

LEMMA 2.1. *If $V \neq 0$ then the largest eigenvalue of $\Omega(t + \delta, t)$ is $O(\lambda \delta)$ as $\delta \to 0$ and the corresponding eigenvector is asymptotic to the orthogonal complement of the kernel of V. If there exists $\mathbf{v}$ such that*

$$V P_j(M)\mathbf{v} = 0, \qquad j = 0, 1, \ldots, k - 1, \tag{2.1}$$

then the smallest eigenvalue of $\Omega(t + \delta, t)$ is $O(\lambda \delta^{2k+1})$.

PROOF. The first step is to make a Taylor-series expansion of $\Omega(t + \delta, t)$. This is derived using

$$d^j T(t, s)/ds^j = T(t, s)P_j(M), \tag{2.2}$$

where $P_j(M)$ is defined in (1.21). We obtain

$$\Omega(t+\delta,t) = \sigma^2\lambda \int_t^{t+\delta} \sum_{i,j}(s-(t+\delta))^{i+j}P_i(M)VP_j(M)^T ds. \qquad (2.3)$$

The result now follows by making Rayleigh-quotient estimates [5] for the extreme eigenvalues.

The smoothness properties of $\mathbf{x}(t\,|\,n)$ are obtained by differentiating the interpolation formula (1.15). This gives, for $t_{i-1} \le t \le t_i$.

$$\frac{d}{dt}\mathbf{x}(t\,|\,n) = M\mathbf{x}(t\,|\,n) + \sigma^2\lambda VT^T(t_i,t)S_{i\,|\,i-1}^{-1}(\mathbf{x}(t_i\,|\,n) - \mathbf{x}(t_i\,|\,i-1)). \qquad (2.4)$$

It shows immediately that the interesting points are the data points t_i, $i = 1, 2, \ldots, n$. Here we have

$$\begin{aligned}
D &= \frac{d}{dt}\mathbf{x}(t_i + \,|\,n) - \frac{d}{dt}\mathbf{x}(t_i - \,|\,n) \\
&= \sigma^2\lambda V\{T(t_{i+1},t_i)^T S_{i+1\,|\,i}^{-1}(\mathbf{x}(t_{i+1}\,|\,n) - \mathbf{x}(t_{i+1}\,|\,i)) \\
&\qquad - S_{i\,|\,i-1}^{-1}(\mathbf{x}(t_i\,|\,n) - \mathbf{x}(t_i\,|\,i-1))\},
\end{aligned}$$

and this reduces, using the filtering and smoothing equations, to

$$D = -\lambda V\mathbf{h}\{y_i - \mathbf{h}^T\mathbf{x}(t_i\,|\,n)\}, \qquad (2.5)$$

so that $\mathbf{x}(t\,|\,n)$ can be continuously differentiable only if $V\mathbf{h} = 0$.

To extend this result to higher derivatives, note that the first occurrence of a discontinuity in $\mathbf{x}(t\,|\,n)$ will appear in differentiating the term involving $VT(t_i,t)^T$ in (2.4). Paralleling the above argument shows that the first k derivatives of $\mathbf{x}(t\,|\,n)$ are continuous only if

$$VP_j(M)^T\mathbf{h} = 0, \qquad j = 0, 1, \ldots, k-1. \qquad (2.6)$$

In particular, it follows that

$$k \le \dim\ker\{V\}. \qquad (2.7)$$

REMARK 2.1. The above discussion has the following consequences.

(i) If V is nonsingular, then (2.7) implies that $k = 0$, so that $\mathbf{x}(t\,|\,n)$ has a derivative that is discontinuous at the data points irrespective of the choice of $\mathbf{h}$. In this case the eigenvalues of $\Omega(t+\delta,t)$ are all $O(\lambda\delta)$.

(ii) If there exists $\mathbf{v}$ satisfying (2.1), then the choice $\mathbf{h} = \mathbf{v}$ ensures that $\mathbf{x}(t\,|\,n)$ has k continuous derivatives.

REMARK 2.2 The condition-number estimate for the Paige and Saunders algorithm given in [7] was based on a Rayleigh-quotient estimate using a comparison vector constructed to stimulate the small eigenvalues of $\Omega(t+\delta,t)$. This argument can be paraphrased here to show that if $\mathbf{h}$ is chosen to give optimal smoothness (say k continuous derivatives) to $\mathbf{x}(t\,|\,n)$, then, using Lemma 2.1,

the condition number of the Paige and Saunders algorithm is $O(\delta^{-2k-1}/\lambda)$. If V is nonsingular then $k = 0$ and the Paige and Saunders algorithm is comparatively well conditioned. In the polynomial smoothing spline case the resulting piecewise polynomial has $(p-1)$ derivatives continuous at the data points.

REMARK 2.3 The freedom of choice in both $\mathbf{h}$ and V makes it possible to construct algorithms for generalised splines with a range of smoothness possibilities at the data points. For example, if

$$M = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \quad V = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{h} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix},$$

corresponding to $p = 3$, then $x_1(t \,|\, n)$ is a quintic piecewise polynomial with its first three derivatives continuous at the data points.

3. Algorithms

The signal-plus-noise model (1.1) and (1.6) can be rewritten as

$$\mathbf{0} = \mathbf{x}(t_{i+1}) - T(t_{i+1}, t_i)\mathbf{x}(t_i) - \mathbf{u}(t_{i+1}, t_i),$$
$$y_i = \mathbf{h}^T\mathbf{x}(t_i) + \varepsilon_i,$$

which can be expressed as the following system of equations

$$\mathbf{f} = C\mathbf{x} + \mathbf{r}. \tag{3.1}$$

To describe the terms in this equation explicitly, let

$$\mathbf{x} = \mathbf{x}_1 \times \mathbf{x}_2 \times \cdots \times \mathbf{x}_n,$$
$$J = \operatorname{diag}\{J_i, \ i = 1, 2, \ldots, n\}, \tag{3.1a}$$

where

$$J_1 = \begin{bmatrix} S_{1|0} & \\ & \sigma^2 \end{bmatrix}, \quad J_i = \begin{bmatrix} \Omega(t_i, t_{i-1}) & \\ & \sigma^2 \end{bmatrix}, \quad i = 2, \ldots, n, \tag{3.1b}$$

$$\mathbf{r} = \mathbf{r}_1 \times \mathbf{r}_2 \times \cdots \times \mathbf{r}_n,$$

where

$$\mathbf{r}_i^T = [-\mathbf{u}(t_i, t_{i-1})^T, \varepsilon_i],$$
$$C = \operatorname{bidiag}\left\{ \begin{bmatrix} I \\ \mathbf{h}^T \end{bmatrix}, \ i = 1, \begin{bmatrix} -T(t_i, t_{i-1}) & I \\ \mathbf{0}\mathbf{h}^T \end{bmatrix}, \ i = 2, \ldots, n \right\}, \tag{3.1c}$$

and

$$\mathbf{f} = \mathbf{f}_1 \times \mathbf{f}_2 \times \cdots \times \mathbf{f}_n,$$

where

$$\mathbf{f}_1 = \begin{bmatrix} \mathbf{x}_{1|0} \\ y_1 \end{bmatrix}, \quad \mathbf{f}_i = \begin{bmatrix} \mathbf{0} \\ y_i \end{bmatrix}, \quad i = 2, \ldots, n. \tag{3.1d}$$

Here J is the covariance of $\mathbf{r}$. Setting $J = SS^T$, the Paige and Saunders linear least-squares formulation can be obtained by premultiplying (3.1) by S^{-1}, which gives us the Duncan and Horn form of the generalised least-squares problem which is

$$\min_{\mathbf{x}} \mathbf{r}^T J^{-1} \mathbf{r}. \tag{3.2}$$

However, Paige [8] points out that the generalised least-squares problem can be posed as

$$\min_{\mathbf{x}, \mathbf{s}} \mathbf{s}^T \mathbf{s}; \; S\mathbf{s} = C\mathbf{x} - \mathbf{f}, \tag{3.3}$$

where $\mathbf{r} = S\mathbf{s}$. This has the form of an equality constrained sum of squares, and avoids the potentially ill-conditioned inversion of S. To solve this problem, let Z be a matrix of maximum rank such that

$$Z^T C = 0. \tag{3.4}$$

Then (3.3) is equivalent to the problem

$$\min_{\mathbf{s}} \mathbf{s}^T \mathbf{s}; \; Z^T S\mathbf{s} = -Z^T \mathbf{f}. \tag{3.5}$$

The key point of this development is that (3.5) can be well conditioned even when J is exactly singular. Once $\mathbf{s}$ is known, then the constraints on (3.3) are consistent for determining $\mathbf{x}$, and can be solved by any convenient and stable method.

To apply this procedure to (3.1), note first that Z is found easily. A suitable form is

$$Z = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n], \tag{3.6}$$

where

$$\mathbf{z}_i = \mathbf{z}_{i1} \times \mathbf{z}_{i2} \times \cdots \times \mathbf{z}_{in},$$

$$\mathbf{z}_{ii} = \begin{bmatrix} \mathbf{h} \\ -1 \end{bmatrix}, \quad \mathbf{z}_{ij} = \begin{bmatrix} T^T(t_{j+1}, t_j)\mathbf{h} \\ 0 \end{bmatrix}, \; j < i, \qquad \text{and } \mathbf{z}_{ij} = 0, j > i. \tag{3.6a}$$

It is readily seen that $S^T Z$ contains the submatrix σI, and so is of full rank. To solve (3.5) for $\mathbf{s}$, note that if $S^T Z$ is factored in the form

$$S^T Z = [Q_1 \mid Q_2] \begin{bmatrix} U \\ 0 \end{bmatrix}, \tag{3.7}$$

where $Q = [Q_1 \mid Q_2]$ is orthogonal and U upper triangular, then

$$\mathbf{s} = -Q_1 U^{-T}(Z^T \mathbf{f}). \tag{3.8}$$

Also, $S^T Z$ is block upper Hessenberg, so some economies of storage can be obtained by exploiting this structure. For example, if plane rotations are used to build up Q, then it is possible to economise on the elements that must be held

by carrying out the forward substitution in (3.8) as the diagonal elements of U are calculated and applying the result to collapse the area in which Q_1 is being accumulated. But this still appears to need at least $n^2/4$ locations. To complete the computation, note that, from (3.1b),

$$\mathbf{x}(t_i \mid n) = T(t_i, t_{i-1})\mathbf{x}(t_{i-1} \mid n) + \Omega^{1/2}(t_i, t_{i-1})\mathbf{s}_i \tag{3.9}$$

where $\mathbf{s}$ is partitioned to conform with $\mathbf{r}$, and $\mathbf{s}_i$ consists of the first p components of the i'th partition. This is a block forward substitution which can be expected to be stable in either stationary or transient (decaying) problems.

One possible way of avoiding the storage problem is by applying Paige's device recursively. The idea is to start with $\mathbf{x}(t_i \mid i)$ and compute $\mathbf{x}(t_{i+1} \mid i+1)$ by using the information from the $(i+1)$st observation to set up the corresponding generalised least-squares problem. This problem is

$$\min_{\mathbf{x}} \mathbf{r}^T J^{-1} \mathbf{r}, \tag{3.10}$$

where

$$J = \operatorname{diag}\{S_{i\mid i}, \Omega(t_{i+1}, t_i), \sigma^2\}, \tag{3.10a}$$

and

$$\mathbf{r}^T = [\mathbf{x}(t_i) - \mathbf{x}_{i\mid i}, \mathbf{x}(t_{i+1}) - T(t_{i+1}, t_i)\mathbf{x}(t_i), \mathbf{h}^T\mathbf{x}(t_{i+1}) - y_{i+1}]. \tag{3.10b}$$

Here Z reduces to the vector

$$\mathbf{z}^T = [\mathbf{h}^T T(t_{i+1}, t_i), \mathbf{h}^T, -1], \tag{3.11}$$

and the constraints on $\mathbf{s}$ corresponding to (3.5) become

$$\mathbf{z}^T S\mathbf{s} = -\mathbf{z}^T \mathbf{f}, \tag{3.12}$$

giving the least-norm solution

$$\mathbf{s} = -\frac{\mathbf{z}^T \mathbf{f}}{\mathbf{z}^T J\mathbf{z}} S^T \mathbf{z}. \tag{3.13}$$

Substituting $\mathbf{s}$ into the equation corresponding to (3.9) just gives the Kalman filter equations, essentially in square-root covariance form. These are well conditioned, in contrast to the corresponding recursive development for the information filter; the argument used in [7] is readily adapted to show this. To proceed to the next step requires $S_{i+1\mid i+1}^{1/2}$, which is conveniently kept in lower triangular form, and standard techniques exist for computing this [5]. For example, this can be done by first making the orthogonal transformation

$$Q^T \begin{bmatrix} (S_{i\mid i}^{1/2})^T T(t_{i+1}, t_i) \\ (\Omega(t_{i+1}, t_i)^{1/2}))^T \end{bmatrix} \rightarrow \begin{bmatrix} (S_{i+1\mid i}^{1/2})^T \\ 0 \end{bmatrix},$$

and then donating by computing a further orthogonal transformation to obtain

$$P^T \begin{bmatrix} (S_{i+1\mid i}^{1/2})^T & \{S_{i+1\mid i}^{1/2}\}^T\mathbf{h} \\ 0 & \sigma \end{bmatrix} \rightarrow \begin{bmatrix} (S_{i+1\mid i+1}^{1/2})^T & 0 \\ \mathbf{u}^T & \{\sigma^2 + \mathbf{h}^T S_{i+1\mid i}\mathbf{h}\}^{1/2} \end{bmatrix}$$

In this case, a smoothing pass is required to complete the computation of $\mathbf{x}(t \mid n)$.

4. Sensitivity analysis

Two phases must be considered in analysing the generalised least-squares formulation of the estimation problem. These correspond to setting up the equality-constrained least-squares problem and to its solution. The first phase requires the square-root factorisation of J given in (3.1a) by, for example, factoring the individual blocks using the Choleski algorithm. As $\Omega(t + \delta, t)$ can be ill conditioned, this step could be suspect in general. However, in many cases of interest, the structure of Ω can be exploited. For example, in estimating polynomial smoothing splines, pre-and post-diagonal rescaling can be used to effectively remove this difficulty. For this reason, attention is directed to the solution phase of the calculation.

Paige's analysis [8] shows that if $\|\mathbf{s}\| \neq 0$, and if $\mathrm{cond}\{Z^T S\}$ and $\mathrm{cond}\{A\}$ get large as $\delta \to 0$, then the relative change due to perturbation $\|\delta \mathbf{x}\|/\|\mathbf{x}\|$ will be proportional to

$$\kappa = \mathrm{cond}\{Z^T S\}\mathrm{cond}\{C\}^2 \|\mathbf{s}\|/\|\mathbf{q}\|, \tag{4.1}$$

where $\mathbf{q} = \mathbf{f} + \mathbf{r}$ is the estimate of the true signal. In particular, there is no direct dependence on $\mathrm{cond}\{S\}$. Here $\mathbf{s} = S^{-1}\mathbf{r}$ is a vector of estimates of independent, $N(0,1)$ random variables. It follows that, almost surely, $\|\mathbf{s}\| = O(\delta^{-1/2})$ and, as a consequence

$$\|\mathbf{s}\|/\|\mathbf{q}\| = O(1). \tag{4.2}$$

To compute $\mathrm{cond}\{Z^T S\}$, note that this depends on the extreme eigenvalues of

$$Z^T J Z = \sigma^2 I + n\Psi, \tag{4.3}$$

where

$$\Psi_{ij} = \psi(t_i, t_j) + O(\delta\|\Omega\|),$$

and

$$\psi(t_i, t_j) = \int_{t_1}^{\min(t_i, t_j)} \mathbf{h}^T T(t_i, s)\Omega T(t_j, s)^T \mathbf{h}\, ds.$$

As $\psi(s, t)$ can be expected to be a slowly varying function in general (for example, in the spline case where T has polynomial elements), it is reasonable to expect the order of magnitude of the largest eigenvalue of (4.3) to be picked up by forming the Rayleigh quotient with a test vector, each component of which is 1. This gives

$$\begin{aligned} RQ &\approx \sigma^2 + n^2 \int\int \psi(s, t)\, ds\, dt \\ &= \sigma^2(1 + O(\lambda\delta^{-1})), \end{aligned} \tag{4.4}$$

as $\|\Omega\| = \sigma^2 \lambda O(\delta)$. To estimate the order of magnitude of the small eigenvalues, it is necessary to introduce some cancellation into the Rayleigh-quotient sum. This can be done using a test vector with components $(-1)^i$, $i = 1, 2, \ldots, n$, which achieves a differencing of adjacent elements so that

$$RQ = \sigma^2 + n\{n^2(\text{double sum})\delta(\text{differencing})\sigma^2 \lambda \delta(\text{from } \Omega)\}/n(\text{normalising})$$
$$= \sigma^2(1 + O(\lambda)).$$

$$(4.5)$$

If, as in [7], we make the assumption that $\lambda = 0(\delta^\phi)$, $0 < \phi < 1$, then the estimates (4.4), (4.5) give (strictly a lower bound to the condition number)

$$\text{cond}\{Z^T S\} = 0(\lambda^{1/2}\delta^{-1/2}). \tag{4.6}$$

To estimate the order of magnitude of the condition number of C, we again use the Rayleigh-quotient argument. In this case, it follows from (3.1c) that

$$RQ = \frac{\|\mathbf{x}_1\|^2 + \sum_{i=2}^{n-1} \|\mathbf{w}_i\|^2 + \sum_{i=1}^{n} (\mathbf{h}^T \mathbf{x}_i)^2}{\sum_{i=1}^{n} \|\mathbf{x}_i\|^2}, \tag{4.7}$$

where

$$\mathbf{w}_i = \mathbf{x}_{i+1} - T(t_{i+1}, t_i)\mathbf{x}_i. \tag{4.8}$$

To make RQ small, it is necessary to make both $\mathbf{w}_i$ and $\mathbf{h}^T \mathbf{x}_i$ small while simultaneously making $\mathbf{x}_i$ large. Note that there is no restriction in scaling $\mathbf{w}_i$ so that $\max_i \|\mathbf{w}_i\| = 1$ and that

$$T(t_{i+1}, t_i) = I + \delta M(t_i) + O(\delta^2). \tag{4.9}$$

It follows that if $\mathbf{w}_i$ is chosen at each step to just about maximally increase $\|\mathbf{x}_i\|$, then $\mathbf{w}_i$ will be slowly varying (essentially representable by a continuous function so $\mathbf{w}(t_i) = \mathbf{w}_i$) and consequently will have norm close to 1. In addition, it is necessary to keep $\mathbf{h}^T \mathbf{x}_i$ small. Now if $\mathbf{h}^T \mathbf{x}_k = 0$, $k = 1, 2, \ldots, i+1$, then

$$\mathbf{h}^T \mathbf{x}_{i+1} = \mathbf{h}^T \{\mathbf{w}_i + T(t_{i+1}, t_i)\mathbf{x}_i\} = 0$$
$$= \mathbf{h}^T \{\mathbf{w}_i + \delta M(t_i)\mathbf{x}_i\} + O(\delta^2 \|\mathbf{x}_i\|). \tag{4.10}$$

This suggests that $\mathbf{w}_i$ be chosen recursively by

$$\mathbf{w}_i = -\delta M(t_i)\mathbf{x}_i + \mathbf{z}_i, \qquad \mathbf{h}^T \mathbf{z}_i = 0, \tag{4.11}$$

where $\mathbf{z}_i$ is chosen to increase $\mathbf{x}_{i+1}$ and the remarks made concerning the selection of $\mathbf{w}_i$ apply. In this case

$$\mathbf{x}_{i+1} = \mathbf{z} + i(T(t_{i+1}, t_i) - \delta M(t_i))\mathbf{x}_i$$
$$= \mathbf{x}_i + \mathbf{z}_i + O(\delta^2 \|\mathbf{x}_i\|)$$
$$= n \int_0^{t_{i+1}} \mathbf{z}(t)dt + O(1). \tag{4.12}$$

In particular, both $\|\mathbf{w}_i\| = O(1)$ as required and $\mathbf{h}^T \mathbf{x}_i = O(1)$. Substituting in (4.7) gives (again strictly a lower bound)

$$RQ = O(\delta^2). \tag{4.13}$$

The condition-number estimate requires information also on the largest singular value of C. That this is $O(1)$ follows from a straightforward application of Gershgorin's theorem to $C^T C$. It follows from (4.2), (4.6), and (4.13) that

$$\begin{aligned}
\kappa &= O(1 \times \lambda^{1/2} \delta^{-1/2} \times \delta^{-2}) \\
&= O(\lambda^{1/2} \delta^{-5/2}).
\end{aligned} \tag{4.14}$$

REMARK 4.1 The important feature of (4.14) is that it is independent of p. It is also significant that it is generally smaller than the condition-number estimate derived in [7]. For example, when $p = 2$ and $\mathbf{x}(t \mid n)$ is continuously differentiable (this includes the important case of cubic smoothing splines), the ratio of condition numbers is

$$(\delta^{-3}/\lambda)/(\lambda^{1/2}\delta^{-5/2}) = \delta^{-1/2}/\lambda^{3/2},$$

which favours the algorithm considered here provided the assumption $\lambda = O(\delta^\phi)$, $0 < \phi < 1$, is realistic. Of course, by the time $p = 3$ the margin in favour has become considerable. It is worthwhile noting that the condition numbers quoted are lower bounds.

5. Numerical results

The generalised recursive least-squares implementation of the generalised smoothing spline was used on two data sets, the Gallant data [4] (here $n = 72$), and the sunspot data [10, p. 487], (here $n = 176$) for $\mathbf{h} = \mathbf{e}_1$, for a range of values of p, λ and V. These examples are of interest in showing two extreme kinds of behaviour, in the sense that the Gallant data appears to require a low order smoothing spline (so that there is overfitting for higher values of p), while in the case of the sunspot data the smoothing spline appears to try very hard to interpolate the data. The estimates of the errors in the computations are obtained by running the algorithm in both single and double precision on a Univac 1100 computer which has 10^{-8} accuracy in single precision. The algorithm in [7] was generalised to cope with different V. Error estimates for this modified algorithm were obtained in a similar fashion.

TABLE I. Gallant data error estimates

p	V	λ	Paige error	Paige and Saunders error
2	$e_2 e_2^T$	0.0001	0.12×10^{-8}	0.24×10^{-6}
		0.001	0.18×10^{-8}	0.20×10^{-6}
		0.01	0.63×10^{-8}	0.31×10^{-7}
		0.1	0.44×10^{-8}	0.41×10^{-7}
	I_2	0.0001	0.75×10^{-8}	0.14×10^{-6}
		0.001	0.87×10^{-8}	0.11×10^{-6}
		0.01	0.61×10^{-8}	0.43×10^{-7}
		0.1	0.41×10^{-8}	0.34×10^{-7}
3	$e_3 e_3^T$	0.0001	0.23×10^{-7}	0.32×10^{-5}
		0.001	0.41×10^{-7}	0.20×10^{-5}
		0.01	0.25×10^{-7}	0.31×10^{-6}
		0.1	0.10×10^{-7}	0.36×10^{-7}
	$I_3 - e_1 e_1^T$	0.0001	0.26×10^{-7}	0.30×10^{-5}
		0.001	0.40×10^{-7}	0.20×10^{-5}
		0.01	0.22×10^{-7}	0.31×10^{-6}
		0.1	0.84×10^{-7}	0.36×10^{-7}
	I_3	0.0001	0.23×10^{-7}	0.13×10^{-5}
		0.001	0.31×10^{-7}	0.33×10^{-6}
		0.01	0.20×10^{-7}	0.24×10^{-6}
		0.1	0.87×10^{-8}	0.28×10^{-7}
4	$e_4 e_4^T$	0.0001	0.80×10^{-7}	0.53×10^{-3}
		0.001	0.31×10^{-7}	0.19×10^{-4}
		0.01	0.13×10^{-6}	0.37×10^{-3}
		0.1	0.46×10^{-6}	0.27×10^{-3}
	$e_3 e_3^T +$ $e_4 e_4^T$	0.0001	0.39×10^{-7}	0.14×10^{-3}
		0.001	0.27×10^{-7}	0.19×10^{-5}
		0.01	0.10×10^{-6}	0.20×10^{-4}
		0.1	0.39×10^{-6}	0.24×10^{-4}
	$I_4 - e_1 e_1^T$	0.0001	0.38×10^{-7}	0.43×10^{-5}
		0.001	0.22×10^{-7}	0.13×10^{-6}
		0.01	0.10×10^{-6}	0.22×10^{-5}
		0.1	0.39×10^{-6}	0.41×10^{-5}
	I_4	0.0001	0.33×10^{-7}	0.75×10^{-6}
		0.001	0.25×10^{-7}	0.64×10^{-7}
		0.01	0.13×10^{-6}	0.99×10^{-6}
		0.1	0.30×10^{-6}	0.16×10^{-5}

Clearly the results in the accompanying tables illustrate that the recursive algorithm based on Paige's device [8] performs better than the algorithm based on Paige and Saunders' method [9] for different V. Results are given for a range of values of λ which straddle the optimal λ computed by the maximum likelihood estimate procedure suggested in [11]. As predicted by the theory for the Paige and Saunders algorithm, the accuracy improves when the continuity conditions are relaxed, which corresponds to increasing the rank of V. This does not happen when Paige's device is implemented. For the sunspot data the optimal λ is very large and is more or less independent of the continuity conditions. This is consistent with interpolating the data [11]. Evidence for

TABLE 2. Sunspot data error estimates

p	V	λ	Paige error	Paige and Saunders error
2	$e_2 e_2^T$	0.1	0.27×10^{-5}	0.27×10^{-4}
		1.0	0.14×10^{-5}	0.18×10^{-4}
		10.0	0.39×10^{-6}	0.14×10^{-5}
		100.0	0.13×10^{-5}	0.17×10^{05}
	I_2	0.1	0.18×10^{-5}	0.21×10^{-4}
		1.0	0.11×10^{-5}	0.66×10^{-5}
		10.0	0.91×10^{-6}	0.12×10^{-5}
		100.0	0.11×10^{-5}	0.15×10^{-5}
3	$e_e e_e^T$	0.1	0.21×10^{-4}	$0.75a \times 10^{-3}$
		1.0	0.83×10^{-4}	0.33×10^{-3}
		10.0	0.26×10^{-3}	0.17×10^{-2}
		100.0	0.60×10^{-3}	0.10×10^{-2}
	$I_3 - e_1 e_1^T$	0.1	0.10×10^{-4}	0.59×10^{-4}
		1.0	0.52×10^{-4}	0.15×10^{-3}
		10.0	0.19×10^{-3}	0.65×10^{-4}
		100.0	0.37×10^{-3}	0.96×10^{-4}
	I_3	0.1	0.10×10^{-4}	0.67×10^{-4}
		1.0	0.37×10^{-4}	0.95×10^{-4}
		10.0	0.10×10^{-3}	0.77×10^{-4}
		100.0	0.99×10^{-4}	0.10×10^{-3}
4	$e_4 e_4^T$	0.1	0.55×10^{-2}	0.32×10^{1}
		1.0	0.39×10^{-2}	0.88×10^{0}
		10.0	0.81×10^{-2}	0.39×10^{-1}
		100.0	0.28×10^{-1}	0.10×10^{0}
	$e_3 e_3^T +$	0.1	0.38×10^{-2}	0.26×10^{0}
	$e_4 e_4^T$	1.0	0.32×10^{-2}	0.62×10^{-1}
		10.0	0.87×10^{-2}	0.95×10^{-1}
		100.0	0.23×10^{-1}	0.68×10^{-1}
	$I_4 - e_1 e_1^T$	0.1	0.33×10^{-2}	0.48×10^{-1}
		1.0	0.24×10^{-2}	0.13×10^{-1}
		10.0	0.73×10^{-2}	0.12×10^{-1}
		100.0	0.12×10^{-1}	0.86×10^{-2}
	I_4	0.1	0.33×10^{-2}	0.19×10^{-1}
		1.0	0.24×10^{-2}	0.81×10^{-2}
		10.0	0.49×10^{-2}	0.44×10^{-2}
		100.0	0.57×10^{-2}	0.33×10^{-2}

oversmoothing with the Gallant data is provided by the optimal λ decreasing rapidly (from 10^{-2} to 10^{-6}) as p is increased from 2 to 4.

  In summary, the new algorithm proposed in this paper does appear to perform better than existing algorithms and has the additional advantage that the continuity conditions can be relaxed when we possess *a priori* information which suggests that having $2p - 2$ continuous derivatives is too stringent a constraint.

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] D. B. Duncan and S. D. Horn, "Linear dynamic recursive estimation from the viewpoint of regression analysis," *J. Amer. Statist. Assoc.* **67** (1972) 816–821.

[4] T. M. Gerig and A. R. Gallant, "Computing methods for linear models subject to linear constraints," *J. Statist. Comput.* Simulation (1975) 283–296.

[5] G. H. Golub and C. F. VanLoan, *Matrix Computations*, (John Hopkins University Press, 1983).

[6] D. G. Luenberger, *Optimization by Vector Space Methods,* (Wiley, 1969).

[7] M. R. Osborne and Tania Prvan, "On algorithms for generalised smoothing splines," *J. Austral. Math. Soc. Ser B* **29** (1988) 319–338.

[8] C. C. Paige, "Computer solution and perturbation analysis of generalised linear least squares problems," *Math. Comp.* **33** (1979), 171–184.

[9] 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.

[10] S. M. Pondit and S. M. Su, *Time Series and System Analysis with Applications.* (Wiley, 1983).

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