SOME PROBLEMS IN RECURSIVE ESTIMATION.

TANIA PRVAN.


A thesis submitted for the degree of Doctor of Philosophy of
The Australian National University.
Errata and Addenda

p.5 1 7 ... requires that \( f(t) \) be

p.21,22 COROLLARY

p.22 1-8 \[ \beta = (w^TQ^{-1}w+R^{-1})^{-1}w^TQ^{-1}y \]

p.32 1-8 ... are given

equation should be (2.3.11)

p.34 1-8 Also Weinart, Byrd and Sidhu (1980)

p.35 \( T(t_{i}, t_{i-1}), u(t_{i}, t_{i-1}) \) correspond to \( T_{i}, u_{i} \) in (2.3.3)

p.51 1-5 ... denote

p.78 Wecker and Ansley (1983) is a general reference to this chapter. An extension in a somewhat different direction is given in Kohn and Ansley (1985), but there is some overlap with the results presented here.

p.81 in (6.2.9) \( T(t_{i+1}, t_{i}), \) in (6.2.11) \( \Omega(t_{i+1}, t_{i}), \sigma^{2} \lambda \)

p.85 1-3 ... Problem 2.3.1 and Theorem 2.3.3

p.93 1 12 ... (6.5.12)

p.103 1 12 ... (7.2.10)

p.104 1 8 ... Recollect

p.107 1-7 the term marginal likelihood is also used

p.116 1-0 a similar result is given in Ansley and Kohn (1985) {also p.121, 1 0}

p.123 1-7 \( d\omega/ds \)

p.127 (8.2.9) \( \sigma^{2} \lambda \) {also p.136, (9.2.5)}
p.161 in caption 'formal' 95% {also p.162}
DECLARATION.

Much of the material in Chapters 5, 7, 8, and 9 is contained in Osborne and Prvan (1987a), (1987b) and Prvan and Osborne (1987), and represents joint work with my supervisor M.R. Osborne. Except where otherwise stated, the work presented in this thesis is my own.

T. Prvan.
I thank my supervisor Mike Osborne for having me as a student, for his assistance and encouragement throughout my course. I also thank Robert Kohn and Craig Ansley for making their results freely available to Mike Osborne and thank Peter Thompson for some helpful input at the beginning of my studies at ANU.
ABSTRACT.

Smoothing Splines provide curves which smooth discrete, noisy data. The connection between Smoothing Splines and the solution to a particular stochastic differential equation conditioned on all of the available data has given rise to new approaches for obtaining such smoothed curves. These approaches exploit a state space formulation that is provided by the solution to the stochastic differential equation and the observation equation. This stochastic setting has the added advantage that it permits confidence intervals to be attached to Smoothing Splines. A new approach for determining the smoothing parameter based on maximum likelihood estimation has also resulted. The Kalman Filter, the Fixed interval, Discrete time Smoother and Interpolation Smoother are invaluable tools in this stochastic approach. They are introduced in Chapters 2 and 3 and stable recursive implementations of them are the main thrust of Chapters 4 and 5. Chapter 6 extends Wecker and Ansley's (1983) stochastic approach to Lg Smoothing Splines while Chapter 7 presents other stochastic approaches.

In Chapter 8 a generalization of the stochastic formulation of Smoothing Splines is developed. This framework produces smoothed curves which can possess less than the usual order of continuity at the data points.

Numerical algorithms for Generalized Smoothing Splines and the inherent sensitivity of these algorithms are considered in Chapter 9. The connection between the stochastic differential equation approach to smoothed curves and an equivalent class of least squares problems is used to develop the algorithms. The condition number of the Reinsch algorithm in its conventional implementation is used as the benchmark
with which other algorithms are compared. Numerical results which support the conclusions based on the sensitivity analysis of the algorithms are presented.

It is worthwhile mentioning here that even though determining the optimal smoothing parameter is important it will not be considered in much detail in this thesis. The problem of determining the order of \( p \) is not addressed.
NOTATION.

A convention followed in this thesis is that capital case letters
are reserved for matrices and bold lower case letters denote vectors
unless otherwise stated. The sign ' := ' indicates ' equal by
definition '. A list of notation encountered in this thesis is given
below.

$\mathbb{R}$ := the set of real numbers

$O(\cdot)$ := is order of (\cdot)

$\| \cdot \|_2$ := Euclidean norm

$\sim$ := is approximately

$\sim N(\mu, \Sigma)$ := is normally distributed with mean $\mu$ and covariance $\Sigma$

$\delta_{ij} := \begin{cases} 1 & i=j, \\ 0 & i \neq j \end{cases}$

$e_i$ := vector with 1 in i'th position and zeroes elsewhere

$x_i := x(t_i)$

$z_i := z(t_i)$

$u_i := u(t_i, t_{i-1})$

$\epsilon_i := e(t_i)$

$y_i := y(t_i)$

$\Omega_i := \Omega(t_i, t_{i-1})$

$x_{i|k} := x(t_i | k)$

$S_{i|k} := S(t | k)$ the covariance of $x(t | k)$ unless otherwise stated

$CV$ := Cross Validation

$GCV$ := Generalized Cross Validation

$GLS$ := Generalized Least Squares

$LS$ := Least Squares

$MLE$ := Maximum Likelihood Estimation
RSS := Residual Sum of Squares
RQ := Rayleigh quotient

\[ D^p := \frac{d^p}{dt^p} \]

\[ L^+ := \text{formal adjoint of the differential operator } L \]

\[ \text{tr} := \text{trace} \]

\[ \sigma_i(C) := i^{th} \text{ largest singular value of } C \]

\[ \text{cond}(C) := \text{condition number of } C \]

\[ t^+ := t + \delta \text{ where } \delta > 0 \text{ and } \delta \to 0 \]

\[ t^- := t - \delta \text{ where } \delta > 0 \text{ and } \delta \to 0 \]
TABLE OF CONTENTS.

DECLARATION. ii.

ACKNOWLEDGEMENTS. iii.

ABSTRACT. iv.

NOTATION. vi.

CHAPTER 1: INTRODUCTION.

1.1 Smoothing Splines. 1.

1.2 Cubic Smoothing Splines. 3.

1.3 Choosing the smoothing parameter. 7.

1.4 Material covered in this thesis. 10.

CHAPTER 2: THE KALMAN FILTER.

2.1 Introduction. 13.

2.2 Some Estimation Problems. 17.

2.3 The Kalman Filter. 23.

CHAPTER 3: SMOOTHERS.

3.1 Introduction. 34.

3.2 Discrete time, Fixed point Smoother. 35.

3.3 RTS Smoother. 38.

3.4 Interpolation Smoother. 40.
CHAPTER 4: IMPLEMENTING THE KALMAN FILTER.

4.1 Introduction. 45.
4.2 Covariance and Information Filters. 47.
4.3 Information Filter. 48.
4.4 Covariance Filter. 55.

CHAPTER 5: A SQUARE ROOT FIXED-INTERVAL, DISCRETE TIME SMOOTHER.

5.1 Introduction. 62.
5.2 Preliminaries. 62.
5.3 Bierman's Algorithm. 64.
5.4 A square root algorithm. 70.
5.5 Examples. 72.

CHAPTER 6: EXTENDING THE ANSLEY AND WECCKER APPROACH TO SMOOTHING SPLINES TO HANDLE LG SMOOTHING SPLINES.

6.1 Introduction. 78.
6.2 State space formulation. 80.
6.3 Maximum Likelihood Estimation. 82.
6.4 The Kalman Filter as a computational tool. 85.
6.5 Smoothness properties. 90.

CHAPTER 7: OTHER STOCHASTIC APPROACHES TO LG SMOOTHING SPLINES.

7.1 Introduction. 96.
7.2 Stochastic derivation of generalized Reinsch algorithm. 97.
7.3 Working explicitly with the diffuse prior. 105.
7.4 An alternative derivation of Wienert, Byrd and Sidhu's Lg Smoothing Spline. 108.
1.1. Smoothing Splines.

Suppose that the data $y_1, \ldots, y_n$ are given at the data points $t_1, \ldots, t_n$ respectively. In most situations the data $y_1, \ldots, y_n$ contain errors. It isn't unrealistic to think of the data as being represented by a regression model

$$y_i = f(t_i) + \varepsilon_i, \quad i = 1, \ldots, n. \quad (1.1.1)$$

where $\varepsilon_i$ is normally distributed with zero mean and variance $\sigma^2$. The $\varepsilon_i$ can be thought of as 'noise' and the function $f(t)$ as the 'signal'. If the underlying mechanism generating the data is not known then it doesn't make sense to prescribe a parametric form for $f$. A nonparametric form is desirable. This leads to the consideration of what sort of properties we would like $f(t)$ to possess.

Fitting a polynomial of degree $n-1$ which interpolates the data is not desirable because it could oscillate wildly (see for example Atkinson (1978)) while the underlying function generating the data could be smooth. This suggests that a continuous piecewise curve might be desirable. It makes sense to prescribe the continuous piecewise curves between the points $[t_{i-1}, t_i]$, $i=2, \ldots, n$. Interpolating the data by a continuous piecewise curve is only useful if sufficiently accurate data is available. This is not always the case, so one desirable property of $f(t)$ is that it be free of the interpolation restriction. This leaves the consideration of what continuity properties it would be desirable for $f(t)$ to possess at the data points. If it is assumed that the $p$'th derivative of $f(t)$ is square
integrable then minimizing the following functional

\[ \sum_{i=1}^{n} (f(t_i) - y_i)^2 + \mu \int_{t_1}^{t_n} (f^{(p)}(u))^2 \, du \]  

over \( f \) where \( \mu \) is constrained to be positive seems a useful compromise. The first term can be thought of as a measurement of the approximation to the data and the second term as a penalty function which prescribes the smoothness of the data. The term \( \mu \) is known as the smoothing parameter and can be thought of as controlling the trade off between "smoothness" and "approximation to the data".

Whittaker (1923) was the first person to perceive how to balance the two conflicting goals of accuracy and smoothing. He used a sum of squares of p'th differences in the penalty function and only considered equally spaced data points. Schoenberg (1964) is responsible for suggesting the functional (1.1.2). Reinsch (1967) showed that the function which minimized (1.1.2) for a given \( \mu \), in the special case \( p=2 \), was a cubic piecewise polynomial with two continuous derivatives at the data points \( t_2, \ldots, t_{n-1} \) and zero second and third derivatives at the end points. For general \( p \) the function which minimizes (1.1.2) is a piecewise polynomial of degree \( 2p-1 \) with \( 2p-2 \) continuous derivatives at the points \( t_2, \ldots, t_{n-1} \) and zero \( p' \)th to \( (2p-1)' \)th derivatives at the end points. This function is known as a Smoothing Spline and is unique for given \( \mu \). The problem of how to estimate the optimal smoothing parameter \( \mu \) will be addressed in section 1.3.

Anselone and Laurent (1968) gave a general method for constructing Splines and Smoothing Splines which exploited the Hilbert space setting they used.

Silverman (1985) showed that it was possible to interpret
Smoothing Splines as the mean of a posterior distribution given the data under the assumption of an appropriately tailored form of the prior distribution. Kimeldorf and Wahba (1970a), (1970b), (1971) explored particular relationships between Bayesian estimation and spline smoothing. Wahba (1978) provided a different formulation and generalization of the result in Kimeldorf and Wahba (1971). She showed that spline smoothing is equivalent to modelling the signal by the stochastic differential equation

\[
\frac{d^p x}{dt^p} = \sigma \lambda \frac{d\omega(t)}{dt}
\]

where \( \omega(t) \) is a Wiener process with unit dispersion parameter (see, for example, Billingsley (1979)) and \( \lambda \) is a scale parameter corresponding to the reciprocal of the smoothing parameter \( \mu \) which is to be determined. Let

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

be the vector of initial conditions on the above stochastic differential equation. She shows that if \( x(t_1) \) is allowed to have a diffuse prior; that is, setting \( x(t_1) \sim N(0, \gamma^2) \) and letting \( \gamma^2 \to \infty \); then

\[
f(t) = \lim_{\gamma \to \infty} E[x(t)|y(t_1), \ldots, y(t_n)].
\]

A generalization of this result to smoothing splines that are not piecewise polynomials was also developed in Wahba (1978).

1.2 Cubic Smoothing Splines.

In the literature the most important case of Smoothing Splines used in practice is the Cubic Smoothing Spline (see, for example, de Boor (1978)). From the discussion in Section 1.1 a Cubic Smoothing
Spline is obtained by minimizing

\[ F = \sum_{i=1}^{n} (f(t_i) - y_i)^2 + \mu \int_{t_1}^{t_n} (f''(u))^2 \, du \]  

(1.2.1)

over \( f(t) \) for given \( \mu \). It will be shown that the function \( f(t) \) which minimizes \( F \) is a piecewise cubic polynomial with 2 continuous derivatives at the points \( t_2, \ldots, t_{n-1} \) and zero second and third derivatives at the end points.

We will find

\[ \min_{f(t)} F \]

by a standard variational argument. This is done by replacing \( f(t) \) by \( f(t) + \varepsilon \eta(t) \) in (1.2.1) where \( \eta \) is assumed to be continuously differentiable but otherwise arbitrary, differentiating the resulting expression with respect to \( \varepsilon \), then equating it to zero, and taking the limit as \( \varepsilon \to 0 \). It will be seen that to uniquely define the function \( f(t) \) some constraints have to be imposed.

Replacing \( f(t) \) by \( f(t) + \varepsilon \eta(t) \) in (1.2.1) yields

\[ F = \sum_{i=1}^{n} (f(t_i) + \varepsilon \eta(t_i) - y_i)^2 + \mu \int_{t_1}^{t_n} (f'(u) + \varepsilon \eta'(u))^2 \, du. \]

Differentiating this expression with respect to \( \varepsilon \) gives

\[ \frac{1}{2} \frac{dF}{d\varepsilon} = \mu \int_{t_1}^{t_n} \eta''(t)(f''(t) + \varepsilon \eta''(t)) \, dt + \sum_{i=1}^{n} \eta(t_i)(f(t_i) + \varepsilon \eta(t_i) - y_i). \]

Now

\[ \frac{1}{2} \lim_{\varepsilon \to 0} \frac{dF}{d\varepsilon} = \mu \int_{t_1}^{t_n} \eta''(t)f''(t) \, dt + \sum_{i=1}^{n} \eta(t_i)(f(t_i) - y_i). \]

and equating this expression to zero produces

\[ \mu \sum_{i=2}^{n} \int_{t_i-1}^{t_i} \eta''(t)f''(t) \, dt = \sum_{i=1}^{n} \eta(t_i)(y_i - f(t_i)). \]
which can be evaluated by integrating by parts to give
\[
\mu \sum_{i=2}^{n} (\eta'(t_i)f''(t_i^-) - \eta'(t_{i-1})f''(t_{i-1}^+)) - \mu \sum_{i=2}^{n} (\eta(t_i)f''(t_i^+))
\]
\[
- \eta(t_{i-1})f''(t_{i-1}^+)) + \mu \sum_{i=2}^{n} \int_{t_{i-1}}^{t_i} \eta(t) f^{(4)}(t) dt =
\]
\[
\sum_{i=1}^{n} \eta(t_i)(y_i - f(t_i)).
\]
(1.2.2)

We can't solve this uniquely without imposing some constraints. It is desirable to make the integral vanish, since \(\eta(t)\) is assumed to be arbitrary this can be achieved by making \(f(t)\) a cubic polynomial. Now (1.2.2) can be rewritten as
\[
\mu \eta'(t_n)f''(t_n^-) - \mu \eta'(t_1)f''(t_1^+) + \mu \sum_{i=2}^{n} \eta'(t_i)(f''(t_i^-) - f''(t_i^+))
\]
\[
- \mu \eta(t_n)f''(t_n^-) + \mu \eta(t_1)f''(t_1^+) - \mu \sum_{i=2}^{n} \eta(t_i)(f''(t_i^-) - f''(t_i^+))
\]
\[
= \eta(t_1)(y_1 - f(t_1)) + \sum_{i=2}^{n} \eta(t_i)(y_i - f(t_i)) + \eta(t_n)(y_n - f(t_n)).
\]

This defines a unique cubic polynomial if the following constraints are imposed
\[
f(t_i^+) = f(t_i^-)
\]
\[
f'(t_i^+) = f'(t_i^-)
\]
\[
f''(t_i^+) = f''(t_i^-)
\]
\[
f''(t_1) = f''(t_n) = f''(t_1) = f''(t_n) = 0
\]
\[
f''(t_i^+) - f''(t_i^-) = \mu^{-1}(y_i - f(t_i))
\]
(1.2.3.a) (1.2.3.b) (1.2.3.c) (1.2.3.d) (1.2.3.e)

which implies that a convenient form of writing \(f(t)\) is
\[
f(t) = a_t^3 + b_t(t-t_1) + c_t(t-t_1)^2 + d_t(t-t_1)^3.
\]
\[
t \in [t_1, t_{i-1}]
\]
(1.2.4)

All that is left to do is determine the Cubic Smoothing Spline coefficients. This is achieved by substituting (1.2.4) into equations (1.2.3) and manipulating them. Substituting (1.2.4) into (1.2.3c) and
(1.2.3d) gives after simplification
\[ c_1 = c_n = 0, \quad d_i = \frac{c_{i+1} - c_i}{3h_i}, \quad i = 1, \ldots, n-1. \quad (1.2.5) \]
where
\[ h_i = t_{i+1} - t_i. \]
Inserting (1.2.4) into (1.2.3b) using (1.2.5) and rearranging it as an expression for \( b_i \) yields
\[ b_i = \frac{a_{i+1} - a_i}{h_i} - \left( \frac{2}{3} c_i + \frac{1}{3} c_{i+1} \right) h_i, \quad i = 1, \ldots, n-1. \quad (1.2.6) \]
Putting (1.2.4) into (1.2.3b) and the utilizing (1.2.5) and (1.2.6) after simplification produces
\[ \frac{1}{h_i} a_{i+1} - \left( \frac{1}{h_i} + \frac{1}{h_{i-1}} \right) a_i + \frac{1}{h_{i-1}} a_{i-1} = \frac{1}{3} h_i c_{i+1} + \frac{2}{3} (h_{i-1} + h_i) c_i \]
\[ + \frac{1}{3} h_{i-1} c_{i-1}, \quad i = 1, \ldots, n-1. \quad (1.2.7) \]
Substituting (1.2.4) into (1.2.3e) using (1.2.5) after simplification yields
\[ \frac{1}{h_i} c_{i+1} - \left( \frac{1}{h_i} + \frac{1}{h_{i-1}} \right) c_i + \frac{1}{h_{i-1}} c_{i-1} = \frac{1}{2\mu} (y_i - a_i) \quad (1.2.8) \]
\[ i = 1, \ldots, n \]
Equations (1.2.7) and (1.2.8) can be rewritten in matrix form. They become
\[ Tc = Q^T a \quad (1.2.9) \]
and
\[ Qc = \frac{1}{2\mu} (y - a) \quad (1.2.10) \]
where \( T \in \mathbb{R}^{n-2 \times n-2} \) and \( Q \in \mathbb{R}^{n \times n-2} \) are tridiagonal matrices defined by
\[ t_{ii} = \frac{2}{3} (h_i + h_{i-1}), \quad t_{ii+1} = t_{i+1i} = \frac{1}{3} h_i, \]
\[ q_{i-1i} = \frac{1}{h_{i-1}}, \quad q_{ii} = \left( \frac{1}{h_i}, -\frac{1}{h_{i-1}} \right), \quad q_{i+1i} = \frac{1}{h_i}, \]
and
\[ c^T = (c_2, \ldots, c_{n-1}). \]
\[ y^T = (y_1, \ldots, y_n), \]
\[ a^T = (a_1, \ldots, a_n). \]

The matrix \( T \) is positive definite. Premultiplying (1.2.10) by \( Q^T \) and using (1.2.9) gives
\[ (Q^T Q + \frac{1}{2} \mu^{-1} T)c = \frac{1}{2} \mu^{-1} Q^T y \]
which can be solved for \( c \). Then from (1.2.10)
\[ a = y - 2\mu Qc. \]

Coefficients \( b_i \) and \( d_i \) can then be obtained from (1.2.5) and (1.2.6).

The term
\[ Q^T Q + \frac{1}{2} \mu^{-1} T \]
is a pentadiagonal matrix which is positive definite since the smoothing parameter \( \mu \) is non-negative. This derivation of Cubic Smoothing Splines is included here to contrast with a different derivation given in CHAPTER 7 where the Smoothing Spline is posed as the function of the elements of a solution to a stochastic differential equation conditioned on all of the available data; that is, Wahba's (1978) equivalence is exploited.

An implementation of Reinsch's (1967) algorithm is given in de Boor (1978) for cubic smoothing splines.

The arguments in this section can be generalized to derive a \((2p-1)\) degree Smoothing Spline.

1.3 Choosing the smoothing parameter.

Reinsch (1967) proposed determining the optimal smoothing parameter \( \mu \) by finding the smallest \( \mu \) for which
\[ S(f) = \sum_{i=1}^{n} (y_i - f(t_i))^2 \leq S \]
(1.3.1)
where he assumed that the \( S \) was given. If \( S \) is not given Reinsch
(1967) suggests that it should be chosen to lie in the interval 
\[ \sigma^2(n-1(2n)), \sigma^2(n+1(2n)) \] which is a confidence interval for S.

One of the most popular methods for choosing the smoothing parameter is Cross Validation (see, for example, Wahba (1975)) which is also similar to Allen's PRESS (see Allen (1974)) which is employed within the context of ridge regression. This is a leave one out at a time estimator. Optimal \( \mu \) is determined by minimizing

\[
CV(\mu) = \sum_{i=1}^{n} (y_i - f^i(\mu, t_i))^2
\]

over \( \mu \) where \( f^i(\mu, t_i) \) is the Smoothing Spline based on all of the data except \((t_i, y_i)\) for given \( \mu \) evaluated at the missing data point. The rationale behind this criterion is to see how closely \( f^i(\mu, t_i) \) predicts the missing data \( y_i \).

A related criterion to cross validation is Generalized Cross Validation which was pioneered by Craven and Wahba (1979). The idea here is to choose \( \mu \) to minimize

\[
GCV(\mu) = \frac{\| (I - \delta(\mu))y \|^2}{\| \text{tr}(I - \delta(\mu)) \|^2}
\]

where \( \delta(\mu) \in \mathbb{R}^{nxn} \) satisfies

\[
\begin{bmatrix}
f(\mu, t_1) \\
\vdots \\
f(\mu, t_n)
\end{bmatrix} = \delta(\mu)y, \quad y^T=(y_1, \ldots, y_n),
\]

and the function \( f(\mu, t) \) is the Smoothing Spline based on all of the available data for given \( \mu \). The matrix \( \delta(\mu) \) is known as the influence matrix. It is worthwhile noting that the numerator in (1.3.3) is the residual sum of squares (RSS) evaluated for a particular value of \( \mu \). This method now has strong theoretical support (see Speckman (1980)). Golub, Heath and Wahba (1979) show that the GCV estimate is a rotation invariant version of Allen's PRESS and that it can be thought of as a
weighted version of $CV(\mu)$. They show that it is superior to PRESS because in certain situations PRESS does not possess a unique minimizer (for example when $s(\mu)$ is diagonal). Craven and Wahba (1979) showed that GCV should asymptotically choose the best $\mu$ in the sense of minimizing the predictive mean square

$$R(\mu) = \sum_{i=1}^{n} (f(t_i) - f(\mu, t_i))^2$$

and this is borne out by results published by Golub, Heath and Wahba (1979). Until recently, finding GCV($\mu$) was a computationally expensive procedure. The major cost was in evaluating the trace of $s(\mu)$. Numerical approximations of $O(n)$ were employed to overcome this, for example see Silverman (1984) and Utreras (1980). Hutchinson and de Hoog (1985) developed the first $O(n)$ method to evaluate GCV($\mu$) exactly. Their method was given for cubic polynomial smoothing splines. Ansley and Kohn (1987) also developed an $O(n)$ method to evaluate GCV($\mu$) which was within a stochastic framework. Their method will be discussed in CHAPTer 9 after the appropriate theoretical development has been given. The sequel to Wahba (1978), Wahba (1983) provides Bayesian "confidence intervals" for the General Cross Validation Smoothing Spline.

For the Cubic Smoothing Spline developed in section 1.2, using (1.2.11) and (1.2.12) the function (1.3.3) becomes

$$GCV(\mu) = \frac{\| Q(Q^T Q + \frac{1}{2} \mu^{-1}T)^{-1} Q^T y \|^2}{(\text{tr}(Q(Q^T Q + \frac{1}{2} \mu^{-1}T)^{-1} Q))^2}$$

The matrix $(Q^T Q + \frac{1}{2} \mu^{-1}T)$ is pentadiagonal. Hutchinson and de Hoog (1985) exploited this structure in their algorithm.
1.4 Material covered in this thesis.

Wecker and Ansley (1983) initiated a signal extraction approach to polynomial Smoothing Splines in which they exploit Wahba's (1978) equivalence between Smoothing Splines and Bayesian estimation. This will be developed for Lg Smoothing Splines in CHAPTER 6 and extended to Generalized Smoothing Splines in CHAPTER 9. Within this setting the actual Smoothing Spline is a function of the elements of the solution to a stochastic differential equation conditioned on all of the available data. This stochastic differential equation can be solved explicitly and a recursion can be developed for the past values. The observations are assumed to be decomposable as a function of the elements of the solution to this stochastic differential equation at the appropriate data point plus a noise component. This facilitates the state space formulation which can be written as an overdetermined system of equations plus a noise component which is normally distributed. A likelihood can be attached to this system of equations and this is used to obtain $\lambda$ by MLE. Weinert, Byrd and Sidhu (1980) were the first people to use a state space approach to Smoothing Splines. It will be seen in CHAPTER 6 that the Kalman Filter can be used as a computational tool in evaluating the likelihood for given $\lambda$. The Kalman Filter, Fixed interval, Discrete time Smoother and Interpolation Smoother are crucial in obtaining the Smoothing Spline within this stochastic setting. Also this setting allows formal confidence intervals to be attached to the Smoothing Spline.

In CHAPTER 2 several estimation problems are investigated. Choosing an appropriate Hilbert Space and using the Projection Theorem furnishes the solutions. This is a fairly conventional approach: it can be found in, for example, Luenberger (1969). The most important
problem considered is the recursive estimation problem solved by the Kalman Filter. Here advantage is taken of a successive updating property by exploiting the Projection Theorem. Recursive estimation is an important component of many of the algorithms investigated in this thesis. CHAPTER 3 investigates Smoothers, encompassing the two mentioned above, and gives proofs along the lines of those found in Anderson and Moore (1979) but made more general. Ansley and Kohn's (1982) elegant proof for the Discrete time, Fixed interval Smoother is extended here to derive the Interpolation Smoother. Stable implementations of the Kalman Filter, and the Fixed interval Discrete time Smoother are the main thrust of CHAPTER 4 and CHAPTER 5.

In CHAPTER 7 other stochastic approaches by Ansley and Kohn (1985), (1986) and Kohn and Ansley (1987), are investigated along with those of Weinert, Byrd and Sidhu (1980) and an adaption of some Time Series ideas summarized in Harvey and Phillips (1979). These methods differ from Ansley and Wecker's approach by a different initial assumption being placed upon the stochastic differential equation and they differ amongst themselves by the handling of this assumption. Wecker and Ansley consider that the initial conditions are a constant but unknown vector and suggest estimating this by MLE. The rest consider a diffuse prior on the vector of initial conditions; that is, the vector of initial conditions is assumed to be normally distributed with zero mean and very large covariance which in the limit tends to infinity. This is consistent with having no prior information about the vector of initial conditions. Harvey and Phillips (1979) work explicitly with $\gamma^2I$ covariance setting $\gamma^2 = 10000$. The others note that the influence of the diffuse prior vanishes after a while and adjust their algorithms accordingly thus avoiding working explicitly
with the diffuse prior. Osborne and Prvan (1987a) derive Weinert, Byrd and Sidhu's (1980) formulation of Smoothing Splines within a framework which avoids using the reproducing kernel Hilbert space that they favoured. Also included in this chapter is a stochastic derivation of Reinsch's (1967) algorithm extended to Lg Smoothing Splines. This is based on work done by Ansley and Kohn (1985) who gave a very direct derivation for cubic splines by making an application of the Projection Theorem.

In CHAPTER 8 a generalization of the stochastic formulation of Smoothing Splines is developed. This framework produces smoothed curves which can possess less than the usual order of continuity at the data points.

CHAPTER 9 gives an analysis of two algorithms for computing the methods discussed. Estimates of the condition numbers for these two algorithms are derived. They are used to differentiate between the two algorithms on the basis of numerical stability. The numerical results support this differentiation. Also included is a brief description of the approach to GCV pioneered by Ansley and Kohn (1987) mentioned in the preceding section.

Even though determining the optimal smoothing parameter is important it is not considered in detail here. The problem of determining the order of p is not addressed.
2.1 Introduction.

In this chapter some estimation problems will be considered. One standard approach in solving these estimation problems, adopted by, for example, Luenberger (1969), is to work within a Hilbert Space of random variables so that the Projection Theorem can be exploited. In a three dimensional Euclidean space it is the theorem that states that the shortest line from a point to a plane is that line from the point which is perpendicular to the plane. The Projection Theorem in its full generality is given below.

**Theorem 2.1.1** Let $H$ be a Hilbert Space and $M$ a closed subspace of $H$. Corresponding to any vector $x \in H$, there is a unique vector $m_0 \in M$ such that

$$
\| x - m_0 \| \leq \| x - m \| \quad \forall m \in M.
$$

Furthermore, a necessary and sufficient condition that $m_0 \in M$ be the unique minimizing vector is that $x - m_0$ be orthogonal to $M$.

A brief outline of the estimation problems to be considered in this chapter will be given below.

The first estimation problem to be considered is that of Least Squares. Here it is hypothesized that the data

$$
y^T = [y_1, \ldots, y_n]
$$

is of the form

$$
y = W\beta + r \quad (2.1.1)
$$

where $W \in \mathbb{R}^{nxp}$ is known, $\beta \in \mathbb{R}^p$ is an unknown parameter, $r \in \mathbb{R}^n$ is
required to make the system consistent and \( n > p \). The problem is to estimate \( \beta \) by making a measure of \( r \) small. The system (2.1.1) is overdetermined and usually doesn't have a solution. Instead of finding the exact solution the problem becomes to minimize

\[
\| y - W\beta \|_2^2
\]

over \( \beta \) in order to make the residual \( r \) small. This is the Least Squares problem. An application of the Projection Theorem provides the solution.

For the rest of the problems considered in this chapter the following Hilbert Space will be employed. It is the Hilbert Space \( \mathcal{H} \) of \( n \) dimensional random vectors where an elementary element in this space can be expressed as

\[
x = \sum_{i=1}^{n} K_i y_i, \quad y_1, \ldots, y_n, x \in \mathcal{H},
\]

where \( K_i \in \mathbb{R}^{nxn} \), along with the inner product

\[
\langle x, z \rangle = E(x^T z) = \text{trace} E(z x^T)
\]

where the norm of an element \( x \) can be written as

\[
\| x \| = \left\{ \text{trace} E(xx^T) \right\}^{1/2}.
\]

The next problem to be considered is that of Minimum-variance Unbiased Estimation. Here the data are assumed to be random variables of the form

\[
y = W\beta + \epsilon
\]

(2.1.2)

where \( W \in \mathbb{R}^{nxp} \) is known, \( \beta \in \mathbb{R}^p \) is unknown parameter and \( \epsilon \in \mathbb{R}^p \) is a random vector with zero mean and covariance \( Q \) which is assumed to be positive definite. Here a linear estimate \( \hat{\beta} \) of the form

\[
\hat{\beta} = Ky
\]

where \( K \in \mathbb{R}^{pxn} \) is a constant matrix is wanted. From (2.1.2) we have that \( y \) is a random vector thus the estimate \( \hat{\beta} \) and the error \( \hat{\beta} - \beta \) are
random vectors whose mean and covariance are determined by $e$ and the choice of $K$. An appropriate optimality criterion is to minimize the variance
\[ \| \hat{\beta} - \beta \|^2 \] over $\hat{\beta}$. It will be shown that to make the above criterion independent of $\beta$ the constraint that $\hat{\beta}$ is an unbiased estimate of $\beta$ has to be imposed. Minimizing (2.1.3) over $\hat{\beta}$ subject to $\hat{\beta}$ being an unbiased estimate of $\beta$ is the Minimum-variance Unbiased Estimation problem whose solution can be obtained by exploiting the Projection Theorem.

The third problem to be considered is that of Minimum-variance estimation. In the preceding two problems the vector $\beta$ was assumed to be an unknown parameter; that is, it was assumed that there is no prior information concerning their values. However if there is prior information, it makes sense to consider the unknown $\beta$ to be a random variable with known mean and covariance. The a priori information can be used to produce an estimate possessing lower error variance than the minimum-variance unbiased estimate. Again the data is assumed to be of the form
\[ y = W\beta + \epsilon \]
but now both $\beta$ and $\epsilon$ are random vectors. Here the problem is to minimize
\[ \| \hat{\beta} - \beta \|^2 \]
over $\beta$. Using the Projection Theorem produces the answer.

Some additional properties of Minimum-variance estimates will be considered such as, what does the minimum-variance estimate of a linear function of $\beta$ plus an error term look like and most importantly, how is the estimate updated when additional information becomes available assuming that the optimal estimate $\hat{\beta}$ of $\beta$ has been
formed on the basis of past measurements with error covariance \( R \)? If additional information of the form

\[
y = W \beta + \epsilon
\]

is given where \( \epsilon \) is a zero mean random vector which is uncorrelated with \( \beta \) and the past data we want the updated estimate \( \hat{\beta} \) of \( \beta \) using \( \hat{\beta} \). This updating must be based on the part of the new data which is orthogonal to the old data. Again the Projection Theorem is being employed. This updating property is a crucial ingredient of this thesis along with recursive updating which uses successive projections. The algorithms used here all depend on this recursive updating property.

Most importantly the following recursive estimation problem is considered. The model consists of an observation equation

\[
y_i = h_i^T x_i + \epsilon_i \quad i=1, \ldots, n,
\]

and a state transition equation

\[
x_i = T_i x_{i-1} + u_i \quad i=2, \ldots, n,
\]

starting with \( x_1 \) having a known mean and covariance. Here \( h_i \in \mathbb{R}^p \) is a constant vector, \( \epsilon_i \in \mathbb{R} \) is a measurement error possessing zero mean and satisfying

\[
E(\epsilon_i \epsilon_j) = \sigma_\epsilon^2 \delta_{ij},
\]

\( y_i \in \mathbb{R} \) is the observation, \( x_i \in \mathbb{R}^p \) is the state vector which is a random variable, \( T_i \in \mathbb{R}^{p \times p} \) is known, and \( u_i \in \mathbb{R}^p \) is a random vector with zero mean satisfying

\[
E(u_i^T u_k) = \delta_{ik} \Sigma_u.
\]

The random variables \( x_1, u_j \) and \( \epsilon_k \) are all assumed to be uncorrelated for \( j \geq 1, k \geq 1 \). The estimation problem is to obtain the minimum variance estimate of the state \( x_k \) from the data \( y_1, \ldots, y_j \). Let \( x(k | j) \equiv x_k | j \) denote the optimal estimate of \( x_k \) based on the observations \( y_1, \ldots, y_n \).
It will be shown that $x(k|j)$ is the appropriate projection of $x_k$ onto the space generated by $y_1, \ldots, y_j$. The Kalman Filter provides the recursion for $x(k|k)$ and $x(k+1|k)$ along with their respective covariances $S(k|k)$ and $S(k+1|k)$, $k=1, \ldots, n$. This involves a succession of projections.

2.2 Some estimation problems.

The Least Squares problem involves finding a parameter vector $\beta \in \mathbb{R}^p$ such that
\[
y = W\beta + r
\]
(2.2.1)

where $W \in \mathbb{R}^{n \times p}$ and $y \in \mathbb{R}^n$ are given with $n \geq p$. The system (2.2.1) is overdetermined, thus it usually does not have an exact solution. Instead of finding the exact solution the problem becomes to minimize
\[
\|y - W\beta\|_2^2.
\]
(2.2.2)

over $\beta$, in order to make the residual $r$ small. This is the Least Squares Problem. The following theorem states the solution to this problem.

**THEOREM 2.2.1** Let (2.2.1) be given where $W$ is assumed to have linearly independent columns. Then there exists a unique $\hat{\beta} \in \mathbb{R}^p$ which minimizes (2.2.2), over all $\beta$. The Projection theorem gives $r^TW = 0$ from which it follows that
\[
\hat{\beta} = (W^TW)^{-1}W^Ty.
\]
(2.2.3)

Least Squares problems are studied in more detail by Lawson and Hanson (1974) in their book.

The Least Squares problem formulated above is not a statistical problem but the technique is widely used in Statistics. It can be
converted into a statistical problem by considering that the deviations from the model have a random origin so that
\[ y = W\beta + \epsilon \]  
(2.2.4)
where \( \epsilon \) is a gaussian random vector for definiteness with zero mean and identity covariance. Then the Least Squares estimate corresponds to the Maximum Likelihood Estimate (see for example Rao (1973)).

Now if
\[ y = W\beta + r \]  
(2.2.5)
where \( r \sim N(0, Q) \) with \( Q \) being positive definite we can't obtain a least squares estimate since identity covariance is required. This can be achieved by premultiplying (2.2.5) by \( (Q^{-1/2})^T \), then the problem is to minimize
\[ \| (Q^{-1/2})^T(y - W\beta) \|_2^2 \]
over \( \beta \). This problem is known as the Weighted or Generalized Least Squares Problem. When \( W \) is assumed to have linearly independent columns its solution is
\[ \hat{\beta} = (W^TQ^{-1}W)^{-1}W^TQ^{-1}y. \]

The second problem considered is that \( y \) is assumed not to be known exactly and is of the form
\[ y = W\beta + \epsilon \]  
(2.2.6)
where \( \epsilon \sim N(0, Q) \) is the error, \( Q \) is a \( n \times n \) positive definite matrix and \( W \in \mathbb{R}^{p \times p} \) is known. A linear estimate of \( \beta \) of the form
\[ \hat{\beta} = Ky \]
where \( K \in \mathbb{R}^{p \times n} \), is wanted. To determine \( K \) the criterion
\[ \min \| \hat{\beta} - \beta \|_2^2 \]  
(2.2.7)
is used where
\[ \| x \|_2^2 = E(x^Tx) = tr(E(xx^T)). \]
Now
\[
\min \| \hat{\beta} - \beta \|^2 = \min \| K_y - \beta \|^2 = \min \| K\omega \beta - \beta + K\epsilon \|^2 = \min \| (K\omega - I)\beta + K\epsilon \|^2.
\]
It follows from this expansion by inspection that the constraint
\[
K\omega = I
\]
has to be imposed to make (2.2.7) independent of \( \beta \). This suggests an alternative problem which is finding the estimate \( \hat{\beta} = K_y \) which minimizes (2.2.7) subject to the constraint (2.2.8). The constraint (2.2.8) is equivalent to making \( \hat{\beta} \) an unbiased estimate of \( \beta \). The standard result for this linear minimum variance unbiased estimate of \( \beta \) is given in the theorem below. It is known as the Gauss Markov Theorem.

**THEOREM 2.2.2** Assume that (2.2.6) is given, then the linear minimum variance unbiased estimate of \( \beta \) is
\[
\hat{\beta} = (W^TQ^{-1}W)^{-1}W^TQ^{-1}y
\]
with corresponding error covariance
\[
E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T] = (W^TQ^{-1}W)^{-1}.
\]

The next estimation problem to be considered in this section asks what happens when \( \beta \) is assumed to be a random variable with known mean and covariance in (2.2.3). The same optimality criterion as in the preceding paragraph is employed. This estimate of \( \beta \) is known as the minimum variance estimate and is given explicitly in the following theorem.
THEOREM 2.2.3 Let \( y \in \mathbb{R}^n \) and \( \beta \in \mathbb{R}^p \) be random vectors. Assume that \( [E(yy^T)]^{-1} \) exists. Then the linear estimate \( \hat{\beta} \) of \( \beta \), based on \( y \), minimizing (2.2.7) is

\[
\hat{\beta} = E(\beta y^T)[E(yy^T)]^{-1}y
\]

(2.2.6)

with corresponding error covariance

\[
E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T] = E(\beta\beta^T) - E(\hat{\beta}\hat{\beta}^T) = E(\beta\beta^T) - E(\beta y^T)[E(yy^T)]^{-1}E(y\beta^T).
\]

(2.2.7)

PROOF. Let \( K = Ky \), the criterion (2.2.4) is used to determine \( \hat{\beta} \).

Assume \( K \) is optimal and consider the variation

\[
K + \eta A
\]

where \( A \) is arbitrary, and define

\[
f = \|K + \eta A\|y - \beta\|2
\]

\[
= E[tr\{Kyy^TK + \eta Ayy^TA^T + 2\eta Kyy^TA - 2\beta y^TK - 2\eta \beta y^TA\}].
\]

Stationarity at \( K \) implies that the \( \eta \) term must vanish as \( A \) is arbitrary (so \( A \) and \( -A \) are permitted),

\[
E[tr\{-2\beta y^TA^T + 2Kyy^TA^T + 2\eta Ayy^TA^T\}] = 0.
\]

Letting \( \eta \to 0 \)

\[
[\Rightarrow E[tr\{(Kyy^T - \beta y^TA^T)A^T\}] = 0.
\]

Using Lemma 1 in Appendix 1 gives

\[
E(-\beta y^T + Kyy^T) = 0
\]

\[
[\Rightarrow KE(yy^T) = E(\beta y^T)
\]

\[
\therefore K = E(\beta y^T)[E(yy^T)]^{-1}.
\]

while its corresponding estimate of \( \beta \) is

\[
\hat{\beta} = Ky
\]

\[
= E(\beta y^T)[E(yy^T)]^{-1}y.
\]

Consider the error covariance

\[
E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T] = E(\beta\beta^T) - E(\hat{\beta}\hat{\beta}^T) - E(\beta\beta^T) + E(\hat{\beta}\hat{\beta}^T).
\]
Using (2.2.6) gives
\[
E(\hat{\beta}\hat{\beta}^T) = E(\beta y^T)[E(yy^T)]^{-1}E(y\beta^T) = E(\beta\beta^T)
\]
and
\[
E(\hat{\beta}\hat{\beta}^T) = E(\beta y^T)[E(yy^T)]^{-1}E(yy^T)[E(yy^T)]^{-1}E(y\beta^T)
\]
\[
= E(\beta y^T)[E(yy^T)]^{-1}E(y\beta^T)
\]
therefore
\[
E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T] = E(\beta\beta^T) - E(\hat{\beta}\hat{\beta}^T)
\]
\[
= E(\beta\beta^T) - E(\beta y^T)[E(yy^T)]^{-1}E(y\beta^T).
\]
QED.

If ε and β are independent and have known covariances the above theorem becomes

COROLLARY 2.2.1 Suppose that the data y is represented by (2.2.6)

where β is an unknown random variable. Let ε and β satisfy
\[
E(\varepsilon\varepsilon^T) = Q, \quad Q \in \mathbb{R}^{n \times n},
\]
\[
E(\beta\beta^T) = R, \quad R \in \mathbb{R}^{p \times p},
\]
and
\[
E(\varepsilon\beta^T) = 0
\]

where R and Q are assumed to be positive definite matrices. Also assume that (WRW^T + Q) is non singular. Then the linear estimate \( \hat{\beta} \) of \( \beta \) minimizing (2.2.4) is
\[
\hat{\beta} = RW^T(WRW^T + Q)^{-1}y
\]
with error covariance
\[
E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T] = R - RW^T(WRW^T + Q)^{-1}WR.
\]
PROOF. Using
\[
K = E(\beta y^T)[E(yy^T)]^{-1}
\]
\[
= E(\beta(\beta + \varepsilon)^T)[E((\beta + \varepsilon)(\beta + \varepsilon)^T)]
\]
gives
\[
\hat{\beta} = RW^T(WRW^T + Q)^{-1} y.
\]

Using
\[
E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T] = E(\beta\beta^T) - E(\beta y^T)[E(yy^T)]^{-1}E(y\beta^T)
\]
gives
\[
E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T] = R - RW^T(WRW^T + Q)^{-1}WR.
\]
QED.

Interesting points arising from this result are no longer requiring \( n \geq p \) and that the estimate exists even if there is no measurement error.

Using the matrix inversion lemma in Appendix 1 the above corollary can be stated differently and. this observation is the gist of the following result.

COROLLARY 2.2.2 The estimate given in COROLLORY 2.2.1 can be expressed alternatively as
\[
\hat{\beta} = (W^TQ^{-1}W + R^{-1})W^TQ^{-1}
\]
with corresponding error covariance
\[
E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T] = (W^TQ^{-1}W + R^{-1})^{-1}.\]
PROOF. The Matrix inversion lemma proved in Appendix 1 states that
\[
RW^T(WRW^T + Q)^{-1} = (W^TQ^{-1}W + R^{-1})^{-1}W^TQ^{-1}.
\]
Substituting this into
\[
\hat{\beta} = RW^T(WRW^T + Q)^{-1} y
\]
gives
\[
\hat{\beta} = (W^T Q^{-1} W + R^{-1})^{-1} W^T Q^{-1} y
\]
and into
\[
E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T] = R - RW^T (WRW^T + Q)^{-1} WR
\]
gives
\[
E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T] = R - (W^T Q^{-1} W + R^{-1})^{-1} W^T Q^{-1} W
= (W^T Q^{-1} W + R^{-1})^{-1} (W^T Q^{-1} W + R^{-1} - W^T Q W) R
= (W^T Q^{-1} W + R^{-1})^{-1}.
\]
QED.

If $R^{-1} = 0$, which corresponds to having no prior information about $\beta$, this result is identical with the Gauss Markov estimate (refer to THEOREM 2.2.2). This also illustrates that $\hat{\beta}$ can be represented as the solution to a weighted least squares problem. This is the track that leads to Duncan and Horn's (1972) weighted linear least squares approach to recursive estimation.

2.3 The Kalman Filter.

Some additional problems associated with the minimum variance estimates discussed in section 2.2 are considered here. The solutions to these problems will be crucial in deriving the Kalman Filter.

The first problem being considered is what does the optimal estimate of a linear function of $\beta$ look like? The following theorem furnishes the answer.

THEOREM 2.3.1 Given an arbitrary matrix $T \in \mathbb{R}^{N \times P}$ and $\epsilon \in \mathbb{R}^n$ uncorrelated with $y \in \mathbb{R}^n$ then the best estimate of $(T\beta + \epsilon)$ is $T\hat{\beta}$ where $\hat{\beta}$ is the minimum variance estimate of $\beta$ given in
**THEOREM 2.2.2.**

**PROOF.** Suppose that the optimal estimate of \((T\beta + \epsilon)\) is \(Ty\). By the Projection Theorem (THEOREM 2.1.1)

\[
\langle T\beta + \epsilon - Ty, Wy \rangle = 0
\]

\[
\Longleftrightarrow E[\text{tr}((T\beta - Ty)y^T W^T)] + E[\text{tr}(\epsilon y^T W^T)] = 0
\]

but \(\epsilon\) is uncorrelated with \(y\) thus

\[
E[\text{tr}(\epsilon y^T W)] = \text{tr}(E(\epsilon)E(y^T W^T)) = 0.
\]

Using this result gives

\[
E[\text{tr}((T\beta - Ty)y^T W^T)] = 0
\]

but by Lemma 1 in Appendix 1

\[
E[(T\beta - Ty)y^T] = 0
\]

\[
\Longleftrightarrow E(T\beta y^T) = E(Tyy^T)
\]

\[
\Longleftrightarrow T(E(\beta y^T)) = T(E(yy^T))
\]

\[
\Longleftrightarrow \Gamma = T(E(\beta y^T))[E(yy^T)]^{-1}
\]

\[
\Longleftrightarrow \Gamma y = T\beta.
\]

QED.

The next problem to investigate is whether optimal \(\beta\) changes when the optimality criterion is stated in terms of a more general quadratic form. Surprisingly it does not change.

**THEOREM 2.3.2** If \(\hat{\beta} = Ky\) is the minimum variance estimate of \(\beta\) then \(\hat{\beta}\) is also the linear estimate minimizing

\[
E[(\hat{\beta} - \beta)^T C(\hat{\beta} - \beta)^T]
\]

where \(C\) is any \(p \times p\) positive definite matrix.

**PROOF.** Let \(C^{1/2}\) be the square root of \(C\). According to THEOREM 2.3.1 \(C^{1/2}\hat{\beta}\) is the minimum variance estimate of \(C^{1/2}\beta\) and hence minimizes
The last problem considered in this section is that of determining how an estimate of $\beta$ is changed if additional data becomes available. Thus an updating property is being considered and orthogonality properties of projections in the Hilbert Space will be exploited. This is the basic problem of recursive least squares estimation. Let $Y_1$ and $Y_2$ be subspaces of a Hilbert Space then the projection of $\beta$ onto the subspace $Y_1 + Y_2$ is equal to the projection onto $Y_1$ plus the projection onto $Y_2$ where $Y_2$ is orthogonal to $Y_1$. In addition, if $Y_2$ is generated by a finite set of vectors, the differences between these vectors and their projections onto $Y_1$ generate $Y_1'$. The following theorem is pivotal to our derivation of the Kalman Filter equations.

**THEOREM 2.3.3** Let $\mathcal{H}$ denote a Hilbert Space of random variables. Let $\beta \in \mathcal{H}$ and let $\hat{\beta}_1$ be the orthogonal projection on a closed subspace $Y_1$ of $\mathcal{H}$, this means that $\hat{\beta}_1$ is the best estimate of $\beta$ in $Y_1$. Let $y_2$ be an $n$ vector of random variables generating a subspace $Y_2$ of $\mathcal{H}$ and let $\hat{y}_2$ be the $n$ dimensional vector of the projection of the components $y_2$ onto $Y_1$, this means that $\hat{y}_2$ is the vector of best estimates of $y_2$ in $Y_1$. Let $y_2' = y_2 - \hat{y}_2$; that is, $y_2'$ is orthogonal to $Y_1$, then the projection of $\beta$ onto the subspace $Y_1 + Y_2'$, denoted by $\hat{\beta}$ is

$$\hat{\beta} = \hat{\beta}_1 + E(\beta y_2^T)[E(y_2' y_2'^T)]^{-1}y_2'. $$

The following problem is crucial in developing the Kalman Filter
PROBLEM 2.3.1 Suppose that an optimal estimate \( \hat{\beta} \) of a random \( p \) dimensional vector \( \beta \) has been formed on the basis of past data with error covariance

\[
E[(\beta - \beta)(\beta - \beta)^T) = R.
\]

Additional data of the form

\[
y = W\beta + \epsilon
\]

is given where \( \epsilon \) is a random vector with zero mean and covariance \( Q \) and is orthogonal to both \( \beta \) and the past. What is the optimal estimate \( \hat{\beta} \) of \( \beta \)?

ANSWER. The best estimate of \( W\beta + \epsilon \) and therefore \( y \) on the past data is

\[
\hat{y} = \hat{W}\beta
\]

by THEOREM 2.3.1. Introducing the innovation

\[
y' = y - \hat{W}\beta
\]

which is orthogonal to the past data we have that the updated estimate of \( \hat{\beta} \) is

\[
\hat{\beta} = \hat{\beta} + E(\beta y'y^T)[E(y'y^T)]^{-1} y'.
\]

by THEOREM 2.3.3. Now

\[
E(y'y'^T) = E[(y - \hat{W}\beta)(y - \hat{W}\beta)^T] = E[(W\beta + \epsilon - \hat{W}\beta)(W\beta + \epsilon - \hat{W}\beta)^T] = WRW^T + Q
\]

and

\[
E(\beta y'^T) = E(\beta y'^T - \hat{\beta} \hat{W}^T)
\]

\[
= E(\beta(\beta - \hat{\beta}) W^T).
\]
But by the Projection Theorem

\[ \text{tr}\{E\{MB(\beta - \hat{\beta})^T\}\} = 0 \quad \forall \ M. \]

By Lemma 1 in Appendix 1

\[ E[\hat{\beta}(\beta - \hat{\beta})^T] = 0 \]

therefore

\[ E[\hat{\beta}(\beta - \hat{\beta})^TW^T] = 0. \]

Thus

\[ E(\beta y^T) = E[(\beta - \hat{\beta})(\beta - \hat{\beta})^TW^T] \]

\[ = RW^T \]

and

\[ \hat{\beta} = \beta + RW^T(WRW^T + Q)^{-1}y'. \]

Let

\[ S = RW^T(WRW^T + Q)^{-1}y'. \]

then

\[ E[(\beta - \hat{\beta})(\beta - \hat{\beta})^T] = E[(\beta - \hat{\beta})(\beta - \hat{\beta})^T] \]

\[ = E[SS^T] - E[S(\beta - \hat{\beta})^T] - E[(\beta - \hat{\beta})S^T]. \]

Now

\[ E(SS^T) = RW^T(WRW^T + Q)^{-1}[E(y - WB)(y - WB)^T]^{-1}(WRW^T + Q)^{-1}WR \]

\[ = RW^T(WRW^T + Q)^{-1}WR \]

\[ E[S(\beta - \hat{\beta})^T] = RW^T(WRW^T + Q)^{-1}E[(y - WB)(y - WB)^T] \]

\[ = RW^T(WRW^T + Q)^{-1}E[(y - WB)(y - WB)^T] \]

\[ = RW^T(WRW^T + Q)^{-1}WR \]

and

\[ E[(\beta - \hat{\beta})S^T] = E[(\beta - \hat{\beta})(y - \hat{y})^T](WRW^T + Q)^{-1}WR \]

\[ = RW^T(WRW^T + Q)^{-1}WR. \]
Thus

\[ E[(\beta - \hat{\beta})(\beta - \hat{\beta})^T] = R - RW^T(WRW^T + Q)^{-1}WR. \] (2.3.2)

REMARK 2.3.1 Duncan and Horn (1972) showed using the matrix inversion lemma that (2.3.1) and (2.3.2) are the weighted least squares estimates for

\[ y = W\beta + \epsilon \]
\[ \beta = \mu + u \]

where \( u \sim (0, Q) \). They condense the model by writing

\[ \tilde{y} = X\tilde{\beta} + \tilde{\epsilon} \]

for

\[ \begin{bmatrix} \mu \\ y \end{bmatrix} = \begin{bmatrix} I \\ W \end{bmatrix} \beta + \begin{bmatrix} -u \\ \epsilon \end{bmatrix}. \]

This can be written as

\[ \tilde{y} - X\tilde{\beta} \sim (0, \tilde{\Sigma}) \]

where

\[ \tilde{\Sigma} = \begin{bmatrix} R & 0 \\ 0 & Q \end{bmatrix}. \]

The Kalman Filter is the solution to the following recursive estimation problem. Kalman (1960) furnished the solution to this problem, a derivation in a similar vein to his is given below. Before stating the recursive estimation problem the model used is outlined in the following definition.

DEFN 2.3.1 The following state space formulation is considered. It consists of a state equation and an observation equation.

The state equation is

\[ x_{i+1} = T_{i+1}x_i + u_{i+1}, \quad i=2,\ldots,n-1 \] (2.3.3)

where \( x_i \in \mathbb{R}^p \) is a state vector which is a random variable.
\( T_{i+1} \in \mathbb{R}^{P \times P} \) is the known state transition matrix and \( u_i \in \mathbb{R}^P \) is a random vector with zero mean satisfying
\[
E(u_i u_i^T) = \delta_{jk} \Omega_i.
\]

Data is represented by the observation equation which is of the form
\[
y_i = h_i^T x_i + \epsilon_i, \quad i=1, \ldots, n \tag{2.3.4}
\]
where \( h_i \in \mathbb{R}^P \) is known, \( y_i \in \mathbb{R} \) is known and \( \epsilon_i \) is a random measurement error with zero mean and \( \sigma_i^2 \) variance. Sometimes the \( \epsilon_i \) are referred to as 'noise'. Attention is restricted to scalar \( y_i \) but the theory is the same for vector \( y_i \).

In addition it is assumed that the random vectors \( x_i, u_j \) and \( \epsilon_k \) are all uncorrelated for \( j \geq 1, k \geq 1 \).

The recursive estimation problem is that of finding the linear minimum variance estimate of the state vector \( x \) from the measurements \( y \). In particular we want \( x_{k+1|k} \) and \( x_{k+1|k+1} \) which are the linear minimum variance estimates of the state vector \( x_{k+1} \) based on the data \( y_1, \ldots, y_k \) and \( y_1, \ldots, y_{k+1} \) respectively. Error covariances associated with \( x_{k+1|k} \) and \( x_{k+1|k+1} \) are denoted by \( S_{k+1|k} \) and \( S_{k+1|k+1} \). The term \( x_{k+1|k} \) is known as the predicted estimate and \( x_{k+1|k+1} \) as the filtered estimate. As stated earlier the Kalman Filter provides the solution to this particular recursive estimation problem. The following theorem furnishes the recursion.
THEOREM 2.3.4 For the given state space formulation (2.3.3) and 
(2.3.4) in DEFN 2.3.1 the following equations, known as the 
Kalman Filter, generate the linear minimum variance estimates of 
the state vector $x_{k+1}$ based on the data $y_1, \ldots, y_k$ and $y_1, \ldots, y_{k+1}$ respectively. The Kalman Filter is

\[ x_{k+1|k} = T_{k+1}x_k | k \]  
\[ (2.3.5) \]

(predicted estimate)

\[ S_{k+1|k} = T_{k+1}S_k | k T_{k+1}^T + \Omega_{k+1} \]  
\[ (2.3.6) \]

(predicted estimate error covariance)

\[ \epsilon_{k+1} = y_{k+1} - h_k^T x_{k+1|k} \]  
\[ (2.3.7) \]

(innovations residual)

\[ d_{k+1} = h_k^T S_{k+1|k} h_k + \sigma_{k+1}^2 \]  
\[ (2.3.8) \]

(innovations residual variance)

\[ x_{k+1|k+1} = x_{k+1|k} + S_{k+1|k} h_k^T \epsilon_{k+1} / d_{k+1} \]  
\[ (2.3.9) \]

(filtered estimate)

and

\[ S_{k+1|k+1} = S_{k+1|k} - S_{k+1|k} h_k h_k^T S_{k+1|k} / d_{k+1} \]  
\[ (2.3.10) \]

(filtered estimate error covariance)

\[ k=1, \ldots, n-1. \]

PROOF. Let $Y_k$ be the subspace which contains the measurements 
$y_1, \ldots, y_k$.

Suppose that $y_1, \ldots, y_k$ have been measured and that the 
estimate $x_k | k$ together with its covariance $S_k | k$ have been 
computed. Using $x_k | k$ we can compute the optimal estimate $x_{k+1|k}$
of

\[ x_{k+1} = T_{k+1} x_k + u_{k+1} \]
given the observations $y_1, \ldots, y_k$. THEOREM 2.3.1 implies that the 
optimal estimate of $T_{k+1} x_k$ is $T_{k+1} x_k | k$ because $u_{k+1}$ is
uncorrelated to $e_k$ and $x_k$, this gives equation (2.3.5). The error covariance of $x_{k+1|k}$ is

$$S_{k+1|k} = E[(x_{k+1|k} - x_{k+1})(x_{k+1|k} - x_{k+1})^T]$$

which furnishes

$$E[(T_{k+1}x_{k|k} - T_{k+1}x_k - u_{k+1}) (T_{k+1}x_{k|k} - T_{k+1}x_k - u_{k+1})^T] = T_{k+1}S_kT_{k+1}^T + \Omega_{k+1}$$

We now have $x_{k+1|k}$, which is the projection of $x_{k+1}$ onto the subspace $Y_k$, and its associated covariance $S_{k+1|k}$. Suppose at time $t_{k+1}$ the following measurement is obtained

$$y_{k+1} = h_{k+1}x_{k+1} + e_{k+1}$$

which gives additional information about $x_{k+1}$. Here an update of $x_{k+1|k}$ based on the additional information $y_{k+1}$ is wanted. This is essentially the same as PROBLEM 2.3.1 which was solved earlier, making suitable identifications we get

$$x_{k+1|k+1} = x_{k+1|k} + S_{k+1|k}h_{k+1}^T(h_{k+1}^TS_{k+1|k}h_{k+1} + \sigma_{k+1}^2)^{-1}y_{k+1}$$

and

$$S_{k+1|k+1} = S_{k+1|k} - S_{k+1|k}h_{k+1}^T(h_{k+1}^TS_{k+1|k}h_{k+1} + \sigma_{k+1}^2)^{-1}h_{k+1}S_{k+1|k}$$

where

$$y_{k+1}' = y_{k+1} - h_{k+1}x_{k+1|k}$$

which are the same as (2.3.9) and (2.3.10).

QED.

REMARK 2.3.2 Kalman and Bucy (1961) established the stability of the Kalman Filter for continuous time systems. Sorenson (1985) asserts that a similar result can be proven for discrete time systems. This result of stability is very important because it is
this which differentiates the Kalman Filter from other similar recursive solutions.

REMARK 2.3.3 Duncan and Horn (1972) rewrote the state space formulation (2.1.4) and (2.1.5) in regression form as

\[
\begin{bmatrix}
    y_1 \\
    y_2 \\
    \vdots
    \\
    y_i
\end{bmatrix}
= \begin{bmatrix}
    I \\
    -T_2 I \\
    \vdots
    \\
    -T_i I
\end{bmatrix}
\begin{bmatrix}
    x_1 \\
    x_2 \\
    \vdots
    \\
    x_i
\end{bmatrix}
+ \begin{bmatrix}
    u_1 \\
    u_2 \\
    \vdots
    \\
    u_i
\end{bmatrix}
\]

(2.1.6)

\[
y_i^* = A_i^* x + \varepsilon_i^*, \quad i = 1, \ldots, n.
\]

where the rows involving \( T_j, \ j=1, \ldots, i \), come from writing (2.1.5) as

\[
0 = x_j - T_j x_{j-1} - u_i.
\]

Let

\[
\Omega_i^* = \text{cov}(\varepsilon_i^*)
\]

then the right hand side of the normal equations for the regression above is

\[
g_i^* = (A_i^*)^T(\Omega_i^*)^{-1}y_i^*
\]

which can be rewritten as

\[
g_i^* = [ (A_{i-1}^*)^T(\Omega_{i-1}^*)^{-1}y_{i-1}^* ]_{h_1 \sigma_i^{-2} y_i}
\]

which demonstrates that the estimate of \( x_i \) is a linear function of the right hand side of the normal equations for (2.1.6). They show that the estimate of \( x_i \) is indeed identical to \( x_i|_{i} \), the linear minimum variance estimate based on the measurements \( y_1, \ldots, y_i \), that the Kalman Filter furnishes. The argument is a
direct extension of the equivalent weighted least squares formulation in REMARK 3.3.1. This equivalence facilitates least squares algorithms for obtaining the Kalman Filter estimates.

3.1 Introduction

In this chapter we are interested in estimating the state \( x(t) \) based on the observations \( y_1, \ldots, y_j \) where \( t > t \).

This is the estimation of past values of the state vector which is known as the Smoothing Problem. A Smoother is any estimator producing a smoothed estimate. We are interested in the Discrete time, Fixed point Smoothing problem, which is that of determining the estimates of \( x(t) \) and its associated covariance based on the observations \( y_1, \ldots, y_j \) where \( k > j \), because it is needed for the development of the Fixed Interval, Discrete time Smoother and the Interpolation Smoother. The Fixed Interval, Discrete time Smoother provides the estimate of \( x(t) \) and its associated covariance based on all of the data \( y_1, \ldots, y_n \). Brown, Robert and Strebel (1963) developed the sequential algorithm which gives these estimates. The Fixed Interval, Discrete time Smoother will be referred to as the EIS Smoother. The Interpolation Problem is that of finding the estimate of \( x(t) \) and its associated covariance based on all of the available data \( y_1, \ldots, y_n \) where \( t > t_n \). Once \( t_n \) is the Interpolation Smoother reduces to the EIS Smoother. The Interpolation problem is largely ignored in the standard literature, I have seen it quoted only in Weather and Amosley (1963). It is derived here using an extension of the usual arguments for deriving the EIS Smoother found in Anderson and Moore (1979). An alternative derivation of the EIS Smoother is given by Amosley and Luhn (1962). It will be generalized to handle the Interpolation Smoother after the Fixed Smoother has been given.
3.1 Introduction.

In this chapter we are interested in estimating the state $x(t)$ based on the observations $y_1, \ldots, y_j$ where $t_j > t$. This is the estimation of past values of the state vector which is known as the Smoothing Problem. A Smoother is any estimator producing a smoothed estimate. We are interested in the Discrete time, Fixed point Smoothing problem, which is that of determining the estimate of $x(t_j)$ and its associated covariance based on the observations $y_1, \ldots, y_k$ where $k > j$, because it is needed for the development of the Fixed interval, Discrete time Smoother and the Interpolation Smoother. The Fixed interval, Discrete time Smoother provides the estimate of $x(t_j)$ and of its associated covariance based on all of the data $y_1, \ldots, y_n$. Rauch, Tung and Streibel (1965) developed the sequential algorithm which gives these estimates. The Fixed interval, Discrete time Smoother will be referred to as the RTS Smoother. The Interpolation Problem is that of finding the estimate of $x(t)$ and of its associated covariance based on all of the available data $y_1, \ldots, y_n$ where $t_{i-1} \leq t \leq t_i$. When $t = t_i$ the Interpolation Smoother reduces to the RTS Smoother. The interpolation problem is largely ignored in the standard literature. I have seen it quoted only in Wecker and Ansley (1983). It is derived here using an extension of the usual arguments for deriving the RTS Smoother found in Anderson and Moore (1979). An alternative derivation of the RTS Smoother is given by Ansley and Kohn (1982). It will be generalized to handle the Interpolation Smoother after the other derivation has been given.
Again the following state space formulation is being considered

\[
x(t_i) = T(t_i, t_{i-1})x(t_{i-1}) + u(t_i, t_{i-1}) \\
y(t_i) = h(t_i)^T x(t_i) + \epsilon(t_i)
\]

where \( x(t_i), h(t_i), u(t_{i+1}, t_i) \in \mathbb{R}^p \), \( T(t_{i+1}, t_i) \in \mathbb{R}^{p \times p} \) and \( y(t_i), \epsilon(t_i) \in \mathbb{R} \). The quantities \( x(t_i), u(t_{i+1}, t_i) \) and \( \epsilon(t_j) \) are all uncorrelated for \( i \geq 1, j \geq 1 \). Also

\[
E(\epsilon(t_i)\epsilon(t_j)) = \delta_{ij}\sigma_j^2
\]

and

\[
E(u(t_i, t_{i-1})u(t_j, t_{j-1})^T) = \delta_{ij}\Omega(t_i, t_{i-1}).
\]

By definition \( x(t|j) \) is the estimate of \( x(t) \) based on the observations \( y_1, \ldots, y_j \) and likewise \( S(t|j) \) is the covariance of \( x(t|j) \). If the two random variables \( u_k \) and \( e_k \), \( k=1, \ldots, n \), in the state space model (3.1.1) and (3.1.2) are normally distributed then

\[
x(t_k|j) = E[x(t_k)|y(t_1), \ldots, y(t_j)]
\]

and

\[
S(t_k|j) = E[(x(t_k) - x(t_k|j))(x(t_k) - x(t_k|j))^T].
\]

3.2 Discrete time, Fixed point Smoother.

As mentioned in the introduction this type of Smoother is of interest because it is required for the development of the RTS Smoother and Interpolation Smoother. We will derive the recursion using a line of argument similar to that of Anderson and Moore (1979).

Here the estimate \( x_j|k \) and its associated covariance \( S_j|k \) based on the observations \( y_1, \ldots, y_k \) for some fixed \( j \) and for all \( k \geq j \) are wanted. These estimates can be obtained by considering an augmenting state vector \( z(t_k) \) for the model (3.1.1) and (3.1.2) satisfying the
equation for \( k \geq j \)

\[
Z_{k+1} = Z_k
\]  
(3.2.1)

and initialized at time \( j \) by \( Z_j = z(t) \). From (3.2.1) we have that

\[
Z_{k+1} = z(t).
\]

If the augmenting state vector \( z(t) \) is correlated with \( x_j \) then it will be related to the observations. Having this correlation is pivotal to the argument. Assuming normality for the random variables \( u_k \) and \( e_k \) in the state space model (3.1.1) and (3.1.2), we have that

\[
Z_{k+1}|_k = z(t|k)
\]

and

\[
S_{k+1}\text{zz}|_k = E[(z(t) - z(t|k))(z(t) - z(t|k))^T].
\]

The Kalman Filter equations ((2.3.5) - (2.3.10)) are applied to the augmented model

\[
\begin{bmatrix}
  x_{k+1} \\
  z_{k+1}
\end{bmatrix} =
\begin{bmatrix}
  T_{k+1} & 0 \\
  0 & I
\end{bmatrix}
\begin{bmatrix}
  x_k \\
  z_k
\end{bmatrix} +
\begin{bmatrix}
  u_{k+1} \\
  0
\end{bmatrix}
\]  
(3.2.2)

\[
y_k =
\begin{bmatrix}
  h_k^T & 0^T
\end{bmatrix}
\begin{bmatrix}
  x_k \\
  z_k
\end{bmatrix} + e_k.
\]  
(3.2.3)

Applying the equations to the augmented model (3.2.2) and (3.2.3) gives

\[
\begin{bmatrix}
  x_{k+1}|_k \\
  z_{k+1}|_k
\end{bmatrix} =
\begin{bmatrix}
  T_{k+1} & 0 \\
  0 & I
\end{bmatrix}
\begin{bmatrix}
  x_k|_k \\
  z_k|_k
\end{bmatrix} +
\begin{bmatrix}
  u_{k+1} \\
  0
\end{bmatrix}
\]

\[
\begin{align*}
  \begin{bmatrix}
    S_{k|k-1} & S_{xz}\text{zz}
  \end{bmatrix}
  \begin{bmatrix}
    h_k \\
    0
  \end{bmatrix}
  \frac{(y_k - h_k^T x_{k|k-1})/d_k}{d_k}
  \\
  S_{k|k-1}\text{sxzz}
\end{align*}
\]

\[
\begin{bmatrix}
  T_{k+1} x_{k|k} + T_{k+1} S_{k|k-1} h_k (y_k - h_k^T x_{k|k-1})/d_k \\
  z_{k|k-1} + S_{k|k-1}\text{szx} h_k (y_k - h_k^T x_{k|k-1})/d_k
\end{bmatrix}
\]
and

\[
\begin{bmatrix}
S_{k+1|k}^{xz} \\
S_{k+1|k}^{zz}
\end{bmatrix} = \begin{bmatrix}
T_{k+1} & 0 \\
0 & I
\end{bmatrix}\begin{bmatrix}
S_{k|k}^{xz} \\
S_{k|k}^{zz}
\end{bmatrix}
\times \begin{bmatrix}
T_{k+1} & 0 \\
0 & I
\end{bmatrix} + \begin{bmatrix}
\Omega_{k+1} \\
0
\end{bmatrix}
\]

\[
\begin{bmatrix}
S_{k|k-1}^{xz} \\
S_{k|k-1}^{zz}
\end{bmatrix} = \begin{bmatrix}
S_{k|k-1}^{xz} \\
S_{k|k-1}^{zz}
\end{bmatrix} - \frac{1}{d_k} \begin{bmatrix}
S_{k|k-1}^{xz} \\
S_{k|k-1}^{zz}
\end{bmatrix}
\]

\[
\begin{bmatrix}
h_k \\
h_k^T
\end{bmatrix} \begin{bmatrix}
S_{k|k-1}^{xz} \\
S_{k|k-1}^{zz}
\end{bmatrix} = \begin{bmatrix}
S_{k|k-1}^{xz} \\
S_{k|k-1}^{zz}
\end{bmatrix} + \begin{bmatrix}
\Omega_{k+1} \\
0
\end{bmatrix}
\]

\[
\begin{bmatrix}
S_{k|k-1}^{xz} \\
S_{k|k-1}^{zz}
\end{bmatrix} = \begin{bmatrix}
T_{k+1}S_{k|k-1}^{xz} \\
S_{k|k-1}^{xz}
\end{bmatrix} + \begin{bmatrix}
\Omega_{k+1} \\
0
\end{bmatrix}
\]

\[
\begin{bmatrix}
S_{k|k-1}^{xz} \\
S_{k|k-1}^{zz}
\end{bmatrix} = \begin{bmatrix}
S_{k|k-1}^{xz} \\
S_{k|k-1}^{zz}
\end{bmatrix} - \frac{1}{d_k} \begin{bmatrix}
T_{k+1}S_{k|k-1}^{xz} \\
S_{k|k-1}^{xz}
\end{bmatrix}
\]

From the above equations we extract

\[z_{k+1|k} = z_{k|k-1} + S_{k|k-1}^{xz}(y_k - h_k^Tx_{k|k-1})/d_k\] (3.2.4)

and

\[S_{k+1|k}^{xz} = S_{k|k-1}^{xz}(T_{k+1}^T - h_k^T h_k^{xz} S_{k|k-1}^{xz} T_{k+1}^T (d_k)^{-1}).\] (3.2.5)

The Fixed point smoothing equations are obtained by initializing (3.2.4) and (3.2.5) with

\[\begin{bmatrix}
x_j \\
z_j
\end{bmatrix}^T = \begin{bmatrix}
x_j \\
x_j
\end{bmatrix}
\]

which gives

\[S_{j|j-1}^{xz} = S_{j|j-1}.\]

The equations are

\[x_{j+k+1|k} = x_{j+k+1|k} + S_{j+k+1|k}^{xz}(y_{k} - h_{k}^T x_{k|k-1})/d_k\] (3.2.6)

and

\[S_{j+k+1|k}^{xz} = S_{j+k+1|k}^{xz}(T_{k+1}^T - h_k^T h_k^{xz} S_{k+k-1}^{xz} T_{k+1}^T (d_k)^{-1})\] (3.2.7)

where \(k_{2j}\) with the initial state \(x_{j|j-1}\) for (3.2.6) and initial
covariance $S_{j|j-1}^{zx} = S_j|j-1$ for (3.2.7). This is a forward filter. Appropriate quantities are extracted from the Kalman Filter equations.

Let

$$
\phi(t_{\ell}, t_{\ell-1}) = T(t_{\ell+1}, t_{\ell}) - T(t_{\ell+1}, t_{\ell})S(t_{\ell}|l-1)h(t_{\ell})h(t_{\ell})^T(d_{\ell})^{-1} = \phi_{\ell}
$$

and

$$
\phi(t_{\ell}, t_{j}) = \phi(t_{\ell-1}, t_{\ell-2})\phi(t_{\ell-2}, t_{\ell-3})\ldots\phi(t_{j}, t_{j-1}).
$$

Repetitive applications of (3.2.4) using (3.2.5) gives

$$
z_j|k = z_j|j + \sum_{l=j+1}^{k} S_{j|j-1}^{zx} \phi(t_{\ell}, t_{j})^T h_{\ell}(y_{\ell} - h_{\ell}^T x_{\ell}|l-1)/d_{\ell}
$$

$$
= z_j|j-1 + \sum_{l=j}^{k} S_{j|j-1}^{zx} \phi(t_{\ell}, t_{j})^T h_{\ell}(y_{\ell} - h_{\ell}^T x_{\ell}|l-1)/d_{\ell}.
$$

For the Fixed point smoothing equations these become

$$
x_j|k = x_j|j + \sum_{l=j+1}^{k} S_{j|j-1}^{zx} \phi(t_{\ell}, t_{j})^T h_{\ell}(y_{\ell} - h_{\ell}^T x_{\ell}|l-1)/d_{\ell} \quad (3.2.8)
$$

$$
= x_j|j-1 + \sum_{l=j}^{k} S_{j|j-1}^{zx} \phi(t_{\ell}, t_{j})^T h_{\ell}(y_{\ell} - h_{\ell}^T x_{\ell}|l-1)/d_{\ell}. \quad (3.2.9)
$$

The development of the Fixed point Smoother follows closely the argument given by Anderson and Moore (1979). It is more general in the sense that $z(t)$ can be any quantity correlated with $x_j$. This generality is important because it is required for the development of the Interpolation Smoother, a type of Smoother not considered by them.

3.3 RTS Smoother.

This type of Smoother provides the recursion for $x(t_{\ell}|n)$ and its associated error covariance $S(t_{\ell}|n)$ for all $1 \leq \ell \leq n$ and $n$ is fixed.

The RTS Smoother is derived from the Fixed point Smoother using
results (3.2.8) and (3.2.9). From (3.2.8) we have that

\[ x_{j-1|k} = x_{j-1|j-1} + \sum_{\ell=j}^{k} S_{j-1|j-2} \Phi(t_{\ell}, t_{j-1})^T h_{\ell}(y_{\ell} - h_{\ell}^T x_{\ell|\ell-1})/d_{\ell} \]

\[ = x_{j-1|j-1} + S_{j-1|j-2} \Phi^T h_{j-1}(y_{j-1} - h_{j-1}^T x_{j-1|j-1}) \]

Rearranging (3.2.9) gives

\[ x_{j|k} - x_{j|j-1} = \sum_{\ell=j}^{k} S_{j|j-1} \Phi(t_{\ell}, t_{j})^T h_{\ell}(y_{\ell} - h_{\ell}^T x_{\ell|\ell-1})/d_{\ell}. \] (3.3.2)

Using (3.3.2) simplifies (3.3.1) to

\[ x_{j|k} = x_{j|j-1} + S_{j-1|j-2} \Phi^T (S_{j|j-1})^{-1}(x_{j|k} - x_{j|j-1}). \] (3.3.3)

Using the definition of \( \phi_{j-1} \) and equation (2.3.10) we obtain the following relationship

\[ S_{j-1|j-2} \Phi^T = S_{j-1|j-2}(T_{j} - T_{j} S_{j-1|j-2} h h^T d_{j-1})^T \]

\[ = (S_{j-1|j-2} - S_{j-1|j-2} h h^T S_{j-1|j-2} d_{j-1})^T \]

\[ = S_{j-1|j-1}^T. \]

Using this result (3.3.3) becomes

\[ x_{j|k} = x_{j|j-1} + S_{j-1|j-1}^T S_{j|j-1}(x_{j|k} - x_{j|j-1}). \] (3.3.4)

The covariance of \( x_{j-1|k} \) is

\[ S_{j-1|k} = E[(x_{j-1|k} - x_{j-1})(x_{j-1|k} - x_{j-1})^T] \]

\[ = E[(x_{j-1|j-1} + A_{j-1}(x_{j|k} - x_{j|j-1}) - x_{j-1})^T (x_{j-1|j-1} + A_{j-1}(x_{j|k} - x_{j|j-1}) - x_{j-1})^T] \]

where

\[ A_{j-1} = S_{j-1|j-1}^T (S_{j|j-1})^{-1}. \] (3.3.5)

This simplifies to

\[ S_{j-1|k} = S_{j-1|j-1} + E[A_{j-1}(x_{j|k} - x_{j} + x_{j} - x_{j|j-1})^T (x_{j|k} - x_{j} + x_{j} - x_{j|j-1})^T A_{j-1}^T] \]

because

\[ E[(x_{j-1|j-1} - x_{j-1})(x_{j|k} - x_{j|j-1})^T A_{j-1}^T] = 0. \]
Expanding the second expression above gives

\[ A_{j-1} E[(x_j|k - x_j)(x_j|k - x_j)^T] \]

\[ - A_{j-1} E[(x_j|k - x_j)(x_j|j-1 - x_j)^T] \]

\[ + A_{j-1} E[(x_j|j-1 - x_j)(x_j|k - x_j)^T] \]

\[ = A_{j-1} S_j[k A_j^{-1} + A_{j-1} S_j A_j^{-1}] \]

\[ - A_{j-1} E[(x_j|k - x_j|j-1 + x_j|j-1 - x_j)(x_j|j-1 - x_j)^T] \]

\[ = A_{j-1} S_j A_j^{-1} - A_{j-1} S_j A_j^{-1} \]

thus

\[ S_{j-1} = S_j A_j^{-1} + A_{j-1} S_j[k A_j^{-1} + A_{j-1} S_j A_j^{-1}] \]  \hspace{1cm} (3.3.6)

By letting \( k=n \) in (3.3.4) and (3.3.5) the RTS Smoother is obtained. This is a backward recursion which is initialized with \( x_n|n \) and its associated error covariance \( S_n|n \) which are obtained from a forward pass of the Kalman Filter. Other quantities like \( x_{j-1}|j-1, x_{j-1}|j-1, S_{j-1}|j-1 \) and \( S_{j-1}|j-1 \) are also obtained from this forward pass.

3.4 Interpolation Smoother.

Here an estimate of \( x(t) \), where \( t_{j-1} \leq t \leq t_j \), based on all the observations \( y_1, \ldots, y_n \) is wanted. The Interpolation Smoother reduces to the RTS Smoother when \( t=t_j \).

To obtain the Interpolation Smoother we initialize the Fixed point Smoother derivation in Section 3.3 at \( z(t) = x(t) \) and use (3.3.2). Now

\[ S_{j-1} = E[(x(t) - x(t|j-1))(x_j - x_j|j-1)^T] \]

\[ = E[(x(t) - x(t|j-1))(T(t_j,t)x(t) - T(t_j,t)x(t|j-1) + u(t_j,t))^T] \]

\[ = S(t|j)T(t_j,t)^T. \]
The expansion for $S(t|j)$ is

$$S(t|j) = E[(x(t) - x(t|j-1))(x(t) - x(t|j-1))^T]$$

$$= E[(T(t, t_{j-1})x_{j-1} - T(t, t_{j-1})x_{j-1}|j-1 + u(t, t_{j-1}))$$

$$+ (T(t, t_{j-1})x_{j-1} - T(t, t_{j-1})x_{j-1}|j-1 + u(t, t_{j-1}))^T]$$

$$= T(t, t_{j-1})S_{j-1}|j-1 T(t, t_{j-1})^T + \Omega(t, t_{j-1})$$

hence

$$S(t|j-1)T(t_{j}, t)^T = T(t, t_{j-1})S_{j-1}|j-1 T(t_{j}, t)^T + \Omega(t, t_{j-1})T(t_{j}, t)^T.$$

Using \(3.2.8\) and initializing at \(z(t) = x(t)\) gives

$$x(t|k) = x(t|j-1) + \sum_{\ell=j}^{k} S(t|j-1)T(t_{\ell}, t)^T \Phi(t_{\ell}, t) T_{\ell} (y_{\ell} - h_{\ell}^T x_{\ell}|\ell-1)/d_{\ell}.$$

Utilizing \(3.3.2\) in the above expression furnishes

$$x(t|k) = x(t|j-1) + S(t|j-1)T(t_{j}, t)^T S_{j-1}^{-1} (x_j|k - x_j|j-1).$$

Set

$$A(t_{j}, t) = S(t|j-1)T(t_{j}, t)^T S_{j}^{-1}|j-1.$$

The error covariance of \(x(t|k)\) is

$$S(t|k) = E[(x(t) - x(t|k))(x(t) - x(t|k))^T]$$

$$= E[(x(t) - x(t|j-1) - A(t_{j}, t)(x_j|k - x_j|j-1))$$

$$+ (x(t) - x(t|j-1) - A(t_{j}, t)(x_j|k - x_j|j-1))^T]$$

$$= S(t|j-1) + A(t_{j}, t)E[(x_j|k - x_j + x_j - x_j|j-1)$$

$$+ (x_j|k - x_j + x_j - x_j|j-1)^T]A(t_{j}, t)^T$$

The expectation term expansion is

$$S_j|k = E[(x_j|k - x_j)(x_j|j-1 - x_j)^T]$$

$$- E[(x_j|j-1 - x_j)(x_j|k - x_j)^T] + S_j|j-1.$$
Hence
\[ S(t|k) = S(t|j-1) + A(t_j,t)(S_j|k - S_j|j-1)A(t_j,t)^T. \]
Letting \( k = n \) provides the desired equations.

Summarizing, the Interpolation Smoother is
\[ x(t|n) = T(t,t_{j-1})x_{j-1|j-1} + A(t_j,t)(x_{j|n} - x_{j|j-1}) \]
where
\[ A(t_j,t) = [T(t,t_{j-1})S_{j-1|j-1}T_j + \Omega(t,t_{j-1})T(t_j,t)^T](S_j|j-1)^{-1} \]
and
\[ S(t|n) = \Omega(t,t_{j-1}) + T(t,t_{j-1})S_{j-1|j-1}T(t,t_{j-1})^T - \]
\[ A(t_j,t)(S_j|j-1 - S_j|n)A(t_j,t)^T \]
for \( t_{j-1} \leq t \leq t_j \).

The Interpolation Smoother can be derived geometrically by exploiting an argument used by Ansley and Kohn (1982) to obtain the RTS Smoother. This elegant derivation is included in this section to show an alternative approach.

THEOREM 3.4.1 The state space formulation (3.1.1) and (3.1.2) is assumed. Let \( \mathbf{w}_i = \{e_i, u_{i+1}^T\} \) which is serially uncorrelated in time. We have that \( \mathbf{w}_i \) is uncorrelated with \( x_j \) for \( j \leq i \). All random variables are normally distributed and have zero mean.

For \( 1 \leq j \leq n \) the normality assumptions give
\[ x(t|j) = E(x(t)|y_1,\ldots,y_j) \]
and
\[ S(t|j) = E[(x(t) - x(t|j))(x(t) - x(t|j))^T] \]
hence \( E(x(t)|y_1,\ldots,y_j) \) is the appropriate projection of \( x(t) \) onto the space generated by \( y_1,\ldots,y_j \).
Suppose that \( x_i | i-1, x_i | i, S_i | i-1 \) and \( S_i | i \) are given by the usual Kalman Filter equations.

Under the above assumptions the Interpolation Smoother is given by (3.4.1), (3.4.2) and (3.4.3).

PROOF. The assumption of normality allows the use of conditional expectations as the projections. Define

\[
q(t) = E(x(t) | y_1, \ldots, y_i, x_{i+1} - x_{i+1|i}, w_{i+1}, \ldots, w_n) \quad (3.4.4)
\]

then

\[
q(t) = E(x(t) | y_1, \ldots, y_i) + E(x(t) | x_{i+1} - x_{i+1|i})
\]

\[
+ E(x(t) | w_{i+1}, \ldots, w_i)
\]

because the three sets of random variables \( \{y_1, \ldots, y_i\}, \{x_{i+1} - x_{i+1|i}\} \) and \( \{w_{i+1}, \ldots, w_n\} \) are mutually uncorrelated. The quantity \( q(t) \) becomes

\[
q(t) = x(t|i) + E(x(t) | x_{i+1} - x_{i+1|i}) + 0 \quad (3.4.5)
\]

where the last term is zero because \( x(t) \) is uncorrelated with \( w_{i+1}, \ldots, w_n \).

Assuming that the inverse of the error covariance of \( x_{i+1|i} \) exists, the linear estimate of \( x(t) \), say \( \hat{x}(t) \), is given by the Projection Theorem

\[
\hat{x}(t) = E(x(t) (x_{i+1} - x_{i+1|i+1})^T)
\]

\[
= E((x_{i+1} - x_{i+1|i}) (x_{i+1} - x_{i+1|i+1})^T)^{-1}
\]

\[
= E((x(t) - x(t|i)) (x(t) - x(t|i))^T)^{-1}
\]

\[
= E((x(t) - x(t|i)) (x(t) - x(t|i))^T)^{-1}
\]

\[
= S(t|i)T(t_{i+1}^T, t)^T(S_{i+1|i}^i)^{-1}(x_{i+1} - x_{i+1|i}). \quad (3.4.6)
\]

The term \( E((x(t) - x(t|i)) (x(t) - x(t|i))^T) \) can be rewritten as

\[
E((x(t) - x(t|i)) (x(t) - x(t|i))^T)
\]
44.

because 

\[ E(x(t|i)(x(t) - x(t|i))^T) \]

equals zero by the Projection Theorem. Using (3.4.6) expression (3.4.5) becomes

\[ q(t) = T(t,t_i)x_i|_i + S(t|i)T(t_{i+1},t)^T(S_{i+1|i})^{-1}(x_{i+1} - x_{i+1|i}). \]

The conditioning variables on the right hand side of (3.4.4) generate \( y_1, \ldots, y_n \). This gives us that

\[ E(x(t)|y_1, \ldots, y_n) = E(q(t)|y_1, \ldots, y_n) \]

\[ = T(t,t_i)x_i|i + S(t|i)T(t_{i+1},t)^T(S_{i+1|i})^{-1}(x_{i+1}|n - x_{i+1|i}). \]

It is straightforward to obtain the error covariance associated with \( x(t|i) \).

QED.
4.1 Introduction.

In this chapter numerically stable ways of implementing the Kalman Filter using recursive Linear Least Squares and recursive Generalized Linear Least Squares will be discussed. In the last two decades considerable interest has been shown in implementing the Kalman Filter in a numerically stable manner. Bierman (1977) gave a comprehensive state of the art exposition for its time. Sorenson (1985) covers the important literature appearing in the IEEE journals. Anderson and Moore (1979) have a chapter on computational aspects of the Kalman Filter.

Square root algorithms along with the concepts of covariance and information filters will be introduced. A modified version of Paige and Saunders (1977) square root information filter implementation of the Kalman Filter will be given. Their approach also encompasses smoothed estimates. They formulate the Kalman Filter as a solution to a recursive weighted Linear Least Squares Problem using the connection established by Duncan and Horn (1972) between the estimation of the state of a discrete linear dynamic system subject to noise and regression analysis. After this Osborne and Prvan's (1987b) square root covariance implementation will be given. Here the Kalman Filter is formulated as a solution to a recursive Generalized Least Squares Problem. The smoothed estimates can be obtained by solving a Generalized Least Squares Problem.

Paige and Saunders' (1977) method is not numerically competitive when state covariances are given explicitly instead of their inverses.
but is of interest in giving information about the attainable accuracy (it will be shown to possess slight advantages) in this situation. It also provides a suggestion for a recursive solution which is competitive in this situation and appears to possess good stability properties. This will be shown in CHAPTER 9 for a specific problem. The Linear Least Squares implementation has the advantage that the predicted, filtered and smoothed estimates can be obtained from one system of equations in $O(n)$ operations instead of running a forward pass of the Kalman Filter followed by a backward pass of the RTS Smoother. However, if the state error covariances $\Omega_j, j=1,\ldots,n$, are ill conditioned then the predicted, filtered and smoothed estimates will be poorly determined.

The Generalized Linear Least Squares implementation proposed by Osborne and Prvan (1987b) does not possess the advantage of solving the predicted, filtered and smoothed estimates from one system of equations in $O(n)$ operations. Obtaining the filtered estimates requires $O(n)$ operations if we formulate the Kalman Filter as the solution to a recursive Generalized Linear Least Squares problem but obtaining the smoothed estimates in Paige's (1979) Generalized Linear Least Squares setting requires $O(n^3)$ operations since we have been unable to derive a recursive formulation within this setting. However our approach permits the state error covariances, $\Omega_j, j=1,\ldots,n$, and observation error covariances $\varepsilon_j, j=1,\ldots,n$, to be ill conditioned provided the original problem has a well determined solution. Our method allows some elements of the state vectors to be known precisely and also permits the possibility that the observations may be very accurate which in the Linear Least Squares framework for our state space formulation causes the approach using Paige and Saunders (1977)
4.2 Covariance and information filters.

Various people like Schlee et al (1967) have noted that the Kalman Filter equations are sensitive to round off errors and that sometimes the accuracy degenerates so much that the results become meaningless. Implementing the Kalman Filter equations ((2.3.5) - (2.3.10)) as they stand can generate filtered and predicted error covariances which are no longer positive semi definite and hence meaningless. We want a stable implementation of the recursions which avoids obtaining such meaningless results.

One way to do this is to replace the Kalman Filter algorithm ((2.3.5) - (2.3.10)) with a numerically better conditioned algorithm like a square root filter. A square root filter updates the square roots of the covariances or inverses of the covariances. A filter which updates the covariances is known as a covariance filter while one which updates the inverses of the covariances is known as an information filter. The square root covariance filter ensures that the covariances are positive semi definite while the square root information filter ensures that the inverse of the covariances are positive semi definite. Bierman (1977) attributes the improved behaviour of square root filters to the reduction in the numerical ranges of the variables. According to him square root filter accuracy in single precision is comparable with the usual formulation of the Kalman Filter equations ((2.3.5) - (2.3.10)) performed in double precision. Potter (see Battin (1964)) is credited with developing the first square root factorization in the area of recursive least squares estimation. People like Kaminski, Bryson and Schmidt (1971) and
Bierman (1974) extended Potter's result to include the presence of input noise and vector measurements. Golub (1965) was the first person to popularize the use of Householder matrices in developing a square root algorithm for the Kalman Filter. Householder (1958) used orthogonal transformations to solve linear least squares problems. Golub (1965) and Businger and Golub (1965) were responsible for working out the practical details.

4.3 Information filter.

Paige and Saunders (1977) developed a square root information filter implementation of the Kalman Filter which uses the alternative formulation of the estimation problem given by Duncan and Horn (1972) (see REMARK 2.3.1 and REMARK 2.3.3). As has been seen the problem becomes a Linear Least Squares Problem with the noise being normally distributed with zero mean and identity covariance. Orthogonal transformations can be used to convert the data matrix into an upper triangular matrix, this is the standard stable approach to solving the Linear Least Squares Problem. Paige and Saunders show how the factorization can be arranged to exploit the problem structure and show the connection between this approach and the Kalman Filter. The smoothed estimates are obtained by back substitution of the transformed data matrix for the linear least squares problem. In this section their method is outlined for our state space formulation (2.3.3) and (2.3.4). No control vector exists in the state transition equation considered here unlike the state space formulation that they considered.

Premultiplying the state equation (2.3.3) by the transpose of the inverse of the square root of the covariance $\Omega_{j+1}$ of $u_{j+1}$ transforms
the equation into one having identity covariance. The mean of the transformed state is still zero. A convenient square root factorization to choose is the Cholesky factorization. Let $L_{j+1}^{-1}$ be the Cholesky factor of $\Omega_{j+1}^{-1}$, in which case $\Omega_{j+1}$ must be positive definite. Premultiplying the observation equation (2.5.4) by $\sigma_j^{-1}$ transforms the equation into one having unit variance. The state space formulation can be rewritten as

$$0 = L_{j+1}^T x_{j+1} + F_j x_j + u_{j+1}$$

$$(4.3.1)$$

$$y'_{j+1} = h_{j+1}^T x_{j+1} + \epsilon_{j+1}$$

$$(4.3.2)$$

where

$$F_{j+1} = -L_{j+1}^T L_{j+1}, \quad u_{j+1}' = L_{j+1}^T u_{j+1}, \quad y'_{j+1} = y_{j+1}/\sigma_{j+1}, \quad h_{j+1}' = h_{j+1}\sigma_{j+1}^{-1}$$

and

$$\epsilon_{j+1}' = \epsilon_{j+1}/\sigma_{j+1}.$$ 

Note that

$$E(u_{j+1}') = 0, \quad E(u_{j+1}'u_{j+1}'^T) = \delta_{ji} I, \quad E(\epsilon_{j+1}') = 0$$

and

$$E(\epsilon_{j+1}'\epsilon_{j+1}') = \delta_{ji}.$$ 

Assuming that the initial state $x_1$ has known covariance $\Omega_1$ the transformed state space formulation (4.3.1) and (4.3.2) can be rewritten as the following system of equations for the first $k$ observations

$$y' = Ax + \epsilon$$

$$(4.3.3)$$

where

$$y' = [y_1', y_2', \ldots, y_k'], \quad 0 \in \mathbb{R}^p,$$

$$x = [x_1, x_2, x_3, \ldots, x_k],$$

$$\epsilon = [\epsilon_1, \epsilon_2, \ldots, \epsilon_k].$$
and

\[
A^T = \begin{bmatrix}
L_1 h_1 & F_1^T \\
L_2 h_2 & F_2^T \\
\vdots & \vdots \\
L_{k-1} h_{k-1} & F_{k-1}^T \\
L_k h_k & F_k^T
\end{bmatrix}.
\]

The vector \( \epsilon \) has zero mean and identity covariance. The least squares estimate of \( x \), say \( \hat{x} \), can be found by

\[
\hat{x} = \arg \min_x \| y' - Ax \|_2^2.
\]

This estimate will include \( x_{k|k} \) since the estimate of \( x_k \) is based on the observations \( y_1, \ldots, y_k \). By appending

\[
0 = L_{k+1}^T x_{k+1} + F_{k+1} x_k + u_{k+1} \tag{4.3.6}
\]

to (4.3.3), the least squares estimate of \( x \) will include \( x_{k+1|k} \) since the estimate of \( x_{k+1} \) is based on the observations \( y_1, \ldots, y_k \). This is the recursive problem which the Kalman Filter solves.

The numerical solution to the linear least squares problem can be obtained by premultiplying (4.3.5) by an orthogonal matrix \( Q^T = [Q_1, Q_2]^T \) which is chosen to transform \( A \) into an upper triangular matrix \( R \).

Since the 2-norm is unaffected by orthogonal transformations

\[
\hat{x} = \arg \min_x \left\| \begin{bmatrix} Q_1^T y' \\ Q_2^T y' \end{bmatrix} - \begin{bmatrix} R \\ 0 \end{bmatrix} x \right\|_2^2
\]

which gives

\[
R \hat{x} = Q_1^T y'.
\]

The matrix \( A \) in (4.3.3) is easily reduced to upper block bidiagonal form by a series of elementary reflectors. Assume that the matrix \( A \) excluding the last block \( \begin{bmatrix} R_k^T \\ h_k^T \end{bmatrix} \) has been reduced to upper
triangular form, ignoring the zero rows we have that
\[
\begin{bmatrix}
    b^{(k-1)} & & \\
    \hat{b}_k & & \\
    y_k & & \\
\end{bmatrix} = 
\begin{bmatrix}
    R^{(k-1)} & & \\
    0_{2k \times (k-1)p} & & \\
    \hat{h}_k & & \\
\end{bmatrix} \begin{bmatrix}
    x^{(k-2)} & & \\
    x_{k-1} & & \\
    x_k & & \\
\end{bmatrix} + \text{noise}
\]  
(4.3.7)

where

\[
b^{(k-1)^T} = [b_1^T, \ldots, b_{k-1}^T]^T, \\
x^{(k-2)} = [x_1, \ldots, x_{k-2}].
\]

and

\[
R^{(k-1)} = 
\begin{bmatrix}
    R_1 & R_{1,2} & & \\
    & R_2 & R_{2,3} & \\
    & & \ddots & \ddots \\
    & & & R_{k-1} & R_{k-1,k}
\end{bmatrix}.
\]

The quantities \( \hat{R}_k \) and \( h_k \) are directly below \( R_{k-1,k} \). The \( R_i \) and \( R_{i,i+1} \) are \( p \times p \) matrices and \( b_1, b_k \in \mathbb{R}^p \). The matrix in (4.3.7) can be reduced to upper block bidiagonal by reducing

\[
\begin{bmatrix}
    \hat{R}_k \\
    h_k^T 
\end{bmatrix} \text{ to } 
\begin{bmatrix}
    \bar{R}_k \\
    0^T 
\end{bmatrix},
\]

where \( \bar{R}_k \) is an upper triangular, by a series of elementary reflectors which zero one element of \( h_k^T \) at a time by using the diagonal element immediately above the element being zeroed working from left to right.

Denoting this series of elementary reflectors by \( Q_k \), then

\[
Q_k^T \begin{bmatrix}
    \hat{R}_k \\
    \hat{h}_k^T 
\end{bmatrix} = \begin{bmatrix}
    \bar{R}_k \\
    0^T 
\end{bmatrix}.
\]

(4.3.8)

This gives

\[
\bar{R}_k x_k|_k = \bar{b}
\]

which can be solved for \( x_k|_k \) by back substitution. Appending (4.3.6)
to this system gives

$$\begin{bmatrix}
    b_{k+1} \\
    b_k \\
    r_k \\
    0
\end{bmatrix} =
\begin{bmatrix}
    R_k \\
    0
\end{bmatrix}
\begin{bmatrix}
    x_{k-1} \\
    x_{k+1}
\end{bmatrix}
+ \text{noise.}
$$

(4.3.9)

The matrix in (4.3.9) can be reduced to an upper block bidiagonal form by reducing

$$\begin{bmatrix}
    \bar{R}_k & 0 \\
    0 & 0 \\
    F_{k+1} & L_{k+1}^T
\end{bmatrix}
\begin{bmatrix}
    b_k \\
    r_k \\
    0
\end{bmatrix}
= \begin{bmatrix}
    R_k & R_{k+1} \\
    0 & 0 \\
    0 & \hat{R}_{k+1}
\end{bmatrix}
\begin{bmatrix}
    b_{k+1} \\
    r_{k+1} \\
    \hat{b}_{k+1}
\end{bmatrix}
$$

where $\hat{R}_{k+1}$ is an upper triangular matrix, by a series of elementary reflectors. This can be done efficiently by starting by zeroing the bottom left hand element using the diagonal element of $\bar{R}_k$ above it, then working from left to right, bottom to top zeroing elements in a similar fashion. Denoting this series of elementary reflectors by $Q^T$ then

$$Q^T \begin{bmatrix}
    \bar{R}_k & 0 \\
    0 & 0 \\
    F_{k+1} & L_{k+1}^T
\end{bmatrix}
\begin{bmatrix}
    b_k \\
    r_k \\
    0
\end{bmatrix}
= \begin{bmatrix}
    R_k & R_{k+1} \\
    0 & 0 \\
    0 & \hat{R}_{k+1}
\end{bmatrix}
\begin{bmatrix}
    b_{k+1} \\
    r_{k+1} \\
    \hat{b}_{k+1}
\end{bmatrix}
$$

(4.3.10)

Deleting the row corresponding to zero in the matrix this gives

$$\hat{R}_{k+1}x_{k+1|k} = \hat{b}_{k+1}
$$

which can be solved by back substitution.

In summary, to obtain the filtered and predicted states we reduce

$$\begin{bmatrix}
    \hat{R}_k & 0 \\
    h_k & 0 \\
    F_{k+1} & L_{k+1}^T
\end{bmatrix}
\begin{bmatrix}
    b_k \\
    y_k \\
    0
\end{bmatrix}
= \begin{bmatrix}
    R_k & R_{k+1} \\
    0 & 0 \\
    0 & \hat{R}_{k+1}
\end{bmatrix}
\begin{bmatrix}
    b_{k+1} \\
    r_{k+1} \\
    \hat{b}_{k+1}
\end{bmatrix}
$$

(4.3.11)

in the two steps (4.3.8) and (4.3.10) initializing at $\hat{R}_1 = L_1^T$ and $\hat{b}_1 = 0$. Step (4.3.8) and step (4.3.10) are applied $n$ and $n-1$ times respectively to obtain all the filtered states and predicted states.
The rows containing zero in the transformed matrix $A$ are deleted from the final system of equations. This system solved by back substitution yields the smoothed estimates since the estimates are based on all of the available data. We need only store $R_k$, $R_{k,k+1}$ and $b_k$ after executing (4.3.11), the quantities $\hat{R}_{k+1}$ and $\hat{b}_k$ are required for the next update.

So far the filtered, predicted and smoothed states have been obtained. Their covariances are also wanted. The error covariance of $\hat{x}$, after simplification, is

$$E[(x - \hat{x})(x - \hat{x})^T] = (R^T R)^{-1}$$

(4.3.12)

where $R$, the upper bidiagonal matrix, is the Cholesky factor of $H^{-1}$. Consequently the error covariance of $x_k|k'$, $S_k|k'$ can be obtained from the bottom right hand corner block of (4.3.8) which from (4.3.12) is

$$(R^T R_k)^{-1}$$

and the error covariance of $x_{k+1}|k$, $S_{k+1}|k'$ can be obtained from the bottom right hand corner block of (4.3.10) which from (4.3.12) is

$$(\hat{R}_T \hat{R}_k)^{-1}.$$}

The smoothed covariances are obtained by reducing $R$ in the system of equations below

$$[\begin{bmatrix} b(n-1) \\ \bar{b}_n \end{bmatrix}] = [\begin{bmatrix} R(n-1) \\ 0_{p \times (n-1)p} \bar{R}_n \end{bmatrix}] x(n|n)$$

to block lower diagonal form by premultiplying by orthogonal transformations in $n-1$ steps beginning at the bottom right hand corner of $R$. After $n-1$ steps we obtain

$$[\begin{bmatrix} R_1 \\ \hat{R}_1 \hat{R}_1 \hat{R}_1 \hat{R}_1 \end{bmatrix}] = R(1)$$

(4.3.13)

where the entries with two primes are in final form and the entry with
a single prime will be altered when zeroing the block above it. From (4.3.12) the error covariance of \(x_{i|n}, S_{i|n}\) is obtained from the \(i\)'th \(p \times p\) block on the diagonal of (4.3.13) and is 

\[
S_{i|n} = (R_i^T R_i)^{-1}.
\]

Paige and Saunders algorithm is a square root information filter because the square roots of the inverse of the covariances are extracted easily. The algorithm requires \(O(n)\) storage. Their algorithm is convenient in the sense that the predicted, filtered and smoothed estimates can be obtained by solving one system of equations. They have converted a Generalized Linear Least Squares problem into a Linear Least Squares Problem by premultiplying the state space formulation (2.3.3) and (2.3.4), rewritten as

\[
\begin{bmatrix}
0 \\
y_1 \\
o \\
y_2 \\
0 \\
\vdots \\
y_n
\end{bmatrix}
= 
\begin{bmatrix}
I \\
(h_1)^T \\
-(T_2)^T & I \\
(h_2)^T \\
\vdots \\
-(T_n)^T & I
\end{bmatrix}
\begin{bmatrix}
x_1 \\
x_2 \\
\vdots \\
x_n
\end{bmatrix}
+ 
\begin{bmatrix}
-u_1 \\
e_1 \\
-u_2 \\
e_2 \\
\vdots \\
e_n
\end{bmatrix}.
\]

This can be rewritten as

\[
\hat{x} = \arg \min_x \| B^{-1}(Cx - f) \|_2^2.
\]

This can be rewritten as

\[
\hat{x} = \arg \min_x \| \bar{C}x - \bar{f} \|_2^2
\]

where \(\bar{C} = B^{-1}C\) and \(\bar{f} = B^{-1}f\). Golub and Van Loan (1983) note that if \(B\) is ill conditioned then \(x\) will be poorly determined. Paige's (1979) alternative formulation of the Generalized Least Squares problem seeks to avoid this problem.
4.4 Covariance filter.

Osborne and Prvan (1987b) developed a square root covariance filter implementation of the Kalman Filter incorporating the RTS Smoother by realizing that Paige and Saunders (1977) in deriving their algorithm had converted a Generalized Linear Least Squares problem into a Linear Least Squares problem. We used Paige's (1979) device for solving the Generalized Linear Least Squares problem. This gave an $O(n^3)$ operations algorithm. Converting Paige's device into a recursive Generalized Linear Least Squares problem produced an $O(n)$ operations algorithm for the filtered estimates but we are still left with using the standard RTS Smoother. Prvan and Osborne (1987) developed an $O(n)$ operations algorithm for the RTS Smoother by converting Bierman's (1983) embedded recursion which is essentially a covariance filter into a square root filter. This method will be outlined in the next chapter. In this section an outline of the Generalized Linear Least Squares covariance filter implementation of the Kalman Filter equations and the RTS Smoother for the state space formulation (2.3.3) and (2.3.4) will be given.

The state space formulation (2.3.3) and (2.3.4) can be rewritten as

$$0 = x_{i+1} - T_{i+1}x_i - u_{i+1} \quad (4.4.1)$$

$$y_i = h_i^T x_i + e_i \quad (4.4.2)$$

where $x_{i|0}$ and its associated error covariance $S_{i|0}$ are assumed to be given. The above equations (4.4.1) and (4.4.2) can be written as the
following system of equations

\[
\begin{bmatrix}
  x_1 \\
  y_1 \\
  0 \\
  y_2 \\
  \vdots \\
  y_n
\end{bmatrix} =
\begin{bmatrix}
  I & (h_1)^T \\
  -T_2 & I \\
  (h_2)^T \\
  -T_n & I
\end{bmatrix}
\begin{bmatrix}
  x_1 \\
  x_2 \\
  \vdots \\
  x_n
\end{bmatrix} +
\begin{bmatrix}
  -u_1 \\
  -u_2 \\
  \vdots \\
  -u_n
\end{bmatrix}
\]

which is written more conveniently as

\[ f = Cx + r. \]  (4.4.3)

Now the covariance of the vector \( r \) is \( J \) where

\[ J = \text{diag}(J_1, i=1, \ldots, n) \]

with the \( J_i \in \mathbb{R}^{(p+1) \times (p+1)} \) which is defined by

\[ J_1 = \begin{bmatrix} S_1 & 0 \\ 0 & \sigma_1^2 \end{bmatrix}, \quad J_i = \begin{bmatrix} \Omega_i & 0 \\ 0 & \sigma_i^2 \end{bmatrix}, \quad i=2, \ldots, n. \]

The Paige and Saunders' Linear Least Squares formulation of the Kalman Filter and RTS Smoother can be obtained by premultiplying (4.4.3) by \((J^{1/2})^{-T}\). Thus the Linear Least Squares problem becomes

\[ \min_r r^T J^{-1} r \]  (4.4.4)

which is the Duncan and Horn (1972) formulation of the Generalized Linear Least Squares problem (see REMARK 2.3.3).

Paige (1979) poses the Generalized Linear Least Squares problem as

\[ \min_{x,s} s^T s : (J^{1/2} s = Cx - f. \]  (4.4.5)

This is an equality constrained sum of squares and avoids the inversion of \((J^{1/2})\) which may be ill conditioned. In this form \(J^{1/2}\) is even allowed to be singular although \(J\) is still positive semi definite. It would be desirable in (4.4.5) for the equality constraint to involve only terms in \(s\). This is achieved by choosing a matrix \(Z^T\)
of maximum rank such that
\[ Z^T C = 0. \]

Then (4.4.5) is equivalent to the problem
\[
\min_{s} \frac{T}{T} s : Z^T (J)^{1/2} s = -Z^T f. \tag{4.4.6}
\]

This has a well defined solution for every right hand side \( f \) iff \( Z^T J^{1/2} \) has full rank (Osborne (1985)). Once \( s \) is known \( x \) can be found from (4.4.5).

To apply Paige's procedure to (4.4.4) \( Z^T \) has to be found. It is easy to verify that for our problem one possibility is
\[
Z^T = \begin{bmatrix}
  h_1^T & -1 \\
  h_1^T & 0 \\
  h_1^T & -1 \\
  h_2^T & 0 \\
  h_2^T & -1 \\
  \vdots & \vdots \\
  h_n^T & 0 \\
  h_n^T & -1 \\
\end{bmatrix}.
\]

This can be written more conveniently as
\[ Z = [Z_1, Z_2, \ldots, Z_n] \]

where
\[
Z_i = Z_{i1} \times Z_{i2} \times \ldots \times Z_{in}
\]
\[
Z_{ii} = \begin{bmatrix} h_i \\ -1 \end{bmatrix}, \quad Z_{ij} = \begin{bmatrix} T_{j+1}^T h_j \\ 0 \end{bmatrix} \quad j \neq i, \quad Z_{ij} = 0 \quad j = i.
\]

The quantity \( Z^T (J)^{1/2} \) contains the submatrix \( \sigma I \) where \( I \in R^{n \times n} \) so is of full rank. To solve (4.4.6) for \( s \), an orthogonal matrix \( Q = [Q_1, Q_2] \) can be chosen so that
\[
(J^{1/2})^T Z = [Q_1, Q_2][U \ 0]
\]

where \( U \) is an upper triangular matrix. Using (4.4.6) this gives
\[ s = -Q_1 U^{-T} Z^T f. \]
This can be seen by setting

\[
\begin{bmatrix}
Q_1^T \\
Q_2^T
\end{bmatrix}
\begin{bmatrix}
v_1 \\
v_2
\end{bmatrix}
\]

so that problem (4.4.6) becomes

\[
\min \{v_1^T v_1 + v_2^T v_2\}
\]

subject to the constraint

\[
U^T v_1 = -Z^T f
\]

which gives

\[
v_1 = -U^{-T} Z^T f
\]

but

\[
v_1 = Q_1^T s
\]

so we end up with the desired result by premultiplying \(v_1\) by \(Q_1\). The quantity \((J^{1/2})^T Z\) is block upper Hessenberg which lends itself to some storage savings but apparently still requires \(O(n^2)\) storage unlike Paige and Saunders' algorithm and has greater computational cost \(O(n^3)\).

This storage problem can be overcome by applying Paige's (1979) approach recursively. This is achieved by initializing with \(x_i|\bar{i}\) and computing \(x_{i+1}|\bar{i+1}\) by using the \((i+1)'\)st observation equation to set up the corresponding Generalized Linear Least Squares problem. Doing this generates the filtered estimates in \(O(n)\) operations. The predicted estimates can be obtained by premultiplying the filtered estimates by the appropriate state transition matrix. This recursive approach needs a smoothing step since we cannot improve upon \(O(n^3)\) operations to obtain the smoothed estimates within a Generalized Linear Least Squares framework. The recursive Generalized Linear Least Squares
problem is
\[
\min_{s_i} s_i^T s_i : (J_i)^{1/2} s_i = C_i x_i^i - f_i
\]  \tag{4.4.7}

where

\[
J_i = \text{diag}(S_i, \Omega_{i+1}, \sigma_{i+1}^2),
\]

\[
\Omega_{i+1} = L_{i+1}L_{i+1}^T.
\]  \tag{4.4.8a}

\[
C_i = \begin{bmatrix}
I & 
-T_{i+1} & 
-I \\
\text{T} & 
-h_{i+1}^T & 
\end{bmatrix}.
\]

\[
x_i^i = \begin{bmatrix} x_i^i_{i+1} \\
x_i_{i+1}^i_{i+1} \\
\end{bmatrix}.
\]  \tag{4.4.8c}

\[
f_i^i = \begin{bmatrix} x_i^i_{i} \\
0 \\
y_{i+1} \\
\end{bmatrix}.
\]  \tag{4.4.8d}

Here the matrix \( Z_i \) required to eliminate \( x_i^i \) in (4.4.7) is a subcase of \( Z \) and is the vector

\[
Z_i^T = [h_i^T L_{i+1}^T h_{i+1}^T -1]
\]  \tag{4.4.9}

which converts constraint (4.4.7) into

\[
Z_i^T (J_i)^{1/2} s_i = -Z_i^T f_i.
\]

Problem (4.4.7) is now equivalent to

\[
\min_{s_i} s_i^T s_i
\]

subject to the constraint

\[
Z_i^T J_i^{1/2} s_i = -Z_i^T f_i
\]

which by the same minimum norm argument given in the preceding paragraph has the solution

\[
s_i = -Z_i^T f_i (Z_i^T J_i Z_i)^{-1} (J_i^{1/2})^T Z_i = \begin{bmatrix}
1 \\
S_i^1 \\
2 \\
S_i^2 \\
3 \\
S_i^3 \\
\end{bmatrix}.
\]
From the constraint (4.4.7) we have that
\[ x_{i|i+1} = s_{i|i}^{-1} x_i + x_{i|i} \]
and
\[ x_{i+1|i+1} = T_{i+1} x_{i|i+1} + L_{i+1}s_{i}^{-1} \]
which are essentially the Kalman Filter equations. We still require
\((s_{i+1|i+1})^{1/2}\) for the next step. This quantity is kept conveniently as
an upper triangular matrix.

One standard approach for obtaining this quantity is by using a
series of elementary reflectors, denoted by \(Q^T\), to reduce
\[
\begin{bmatrix}
(s_{i+1|i+1})^{1/2} \\
L_{i+1}
\end{bmatrix}
\]
to
\[
\begin{bmatrix}
(s_{i+1|i+1})^{1/2} \\
0
\end{bmatrix}.
\]
This is possible because the square root of a matrix is determined up
to an orthogonal matrix, and this can be chosen to achieve the result.

This can be seen by evaluating the following expression
\[
[T_{i+1} s_{i+1|i}^{1/2} L_{i+1}] Q Q^T \begin{bmatrix}
(s_{i+1|i+1})^{1/2} \\
L_{i+1}
\end{bmatrix}
= T_{i+1} s_{i+1|i}^{1/2} + \Omega_{i+1}
= s_{i+1|i}.
\]

Then another series of elementary reflectors denoted by \(P^T\) can be
employed to reduce
\[
\begin{bmatrix}
(s_{i+1|i+1})^{1/2} \\
0
\end{bmatrix}
\]
to
\[
\begin{bmatrix}
(s_{i+1|i+1})^{1/2} \\
\sigma_{i+1}
\end{bmatrix}.
\]
This can be verified by comparing
\[
[T_{i+1/2} s_{i+1|i}^{1/2} \sigma_{i+1}] P P^T \begin{bmatrix}
(s_{i+1|i+1})^{1/2} \\
0
\end{bmatrix}
= \begin{bmatrix}
s_{i+1|i} \\
\sigma_{i+1}
\end{bmatrix}.
\]
with

\[
\begin{bmatrix}
S_{i+1|i+1}^{1/2} & u \\
0^T & \begin{bmatrix}
(S_{i+1|i+1}^{1/2})^T & 0 \\
0 & (\sigma_{i+1|i+1}^2 + h_i^T S_{i+1|i+1} h_i)^{1/2}
\end{bmatrix}
\end{bmatrix} \begin{bmatrix}
(S_{i+1|i+1}^{1/2})^T & 0 \\
0 & (\sigma_{i+1|i+1}^2 + h_i^T S_{i+1|i+1} h_i)^{1/2}
\end{bmatrix}^{-1/2}
\]

\[
= \begin{bmatrix}
S_{i+1|i+1} & uu^T \\
u^T \sigma_{i+1|i+1}^2 + h_i^T S_{i+1|i+1} h_i & (\sigma_{i+1|i+1}^2 + h_i^T S_{i+1|i+1} h_i)^{1/2}
\end{bmatrix}^{1/2}
\]

The required relationship

\[
S_{i+1|i} = S_{i+1|i+1} + uu^T
\]

where

\[
u = \frac{S_{i+1|i+1} h_i (\sigma_{i+1|i+1}^2 + h_i^T S_{i+1|i+1} h_i)^{-1/2}}{}
\]

follows directly.

In CHAPTER 5 an \(O(n)\) square root algorithm for the smoothed estimates and their respective covariances will be given.
CHAPTER 5

A SQUARE ROOT FIXED - INTERVAL, DISCRETE TIME SMOOTHER.

5.1 Introduction.

Reports of computational experience with the RTS smoother (see section 3.3) suggest that on occasions numerical problems can occur. Bierman (1983) identifies these problems as being caused by computations involving the differences of positive definite matrices which can cause loss of significance due to cancellation, and to a lesser extent by the inversion of covariance matrices which can be ill conditioned. In addition these explicit inversions can be computationally expensive. He suggested an alternative formulation which avoided these problems.

In this chapter we plan to give a more comprehensive development of Bierman’s (1983) embedded smoother approach and to show how the equations can be rearranged to exploit the inherent structure. It will be shown that his algorithm can be implemented in a numerically stable way as a square root smoother. This is desirable because the positive definiteness of the smoothed covariances can be preserved. Numerical results are given for the usual form of the RTS Smoother, Bierman’s algorithm and our square root formulation of his algorithm. These confirm the expected advantages. The material in this chapter can also be found in Prvan and Osborne (1987).

5.2 Preliminaries.

The following signal model is considered

\[ x_j = T_j x_{j-1} + w'_j \]  
\[ y_j = h'_j x_j + v_j \]  

(5.2.1)  
\[ j=1, \ldots, n. \]  

(5.2.2)
where $x_j$, $w_j$, $h_j$, $y_j$, $v_j$ $\in \mathbb{R}$ and $T_j \in \mathbb{R}^{p \times p}$. An initial estimate of the first state $x_1$, say $\hat{x}_1$, is assumed to be known and likewise an estimate of its covariance $S_1$. The following assumptions are made that

$$
E(w_j^T) = 0, \quad E(v_j^T) = 0, \quad E((x_1 - \hat{x}_1)v_j) = 0
$$

where $\{v_k\}$, $\{w_k\}$ and $x_1$ are uncorrelated.

The Kalman Filter recursion for the signal model (5.2.1) and (5.2.2) under assumptions (5.2.3), (5.2.4) and (5.2.5) is

$$
x_{j|j-1} = T_j x_{j-1|j-1}
$$

(one step prediction estimate)

$$
S_{j|j-1} = T_j S_{j-1|j-1} T_j^T + \Omega_j
$$

(estimate prediction error covariance)

$$
e_j = y_j - h_j x_{j|j-1}
$$

(innovations residual)

$$
d_j = h_j^T S_j | j-1 h_j + \sigma^2
$$

(innovations residual covariance)

$$
x_{j|j} = x_{j|j-1} + S_{j|j-1} h_j e_j / d_j
$$

(estimate update)

and

$$
S_{j|j} = S_{j|j-1} - S_{j|j-1} h_j h_j^T S_{j|j-1} / d_j
$$

j=1,...,n.

The RTS Smoother is

$$
x_{j|n} = x_{j|j} + A_j (x_{j+1|n} - x_{j+1|j})
$$

(smoothed estimate)

and

$$
S_{j|n} = S_{j|j} + A_j (S_{j+1|n} - S_{j+1|j}) A_j^T
$$

(smoothed estimate error covariance)
where
\[ A_j = S_j j^{T} j^{+1} (S_{j+1} j^{+1})^{-1}. \]  
(smoothen gain)  

\[ j=n, \ldots, 1. \]

The quantities \( x_n \) and \( S_n \) which initiate this backward recursion are obtained from the Kalman Filter. It is in equation (5.2.7b) that the possibility of losing the property of positive definiteness for the smoothed covariance may arise because of the differencing of the positive definite matrices. Our algorithm exploits Bierman’s (1983) equations for the RTS Smoother by arranging it as a square root smoother which ensures that the smoothed covariances remain positive definite. In some contexts it is desirable to maintain the positive definiteness of the smoothed covariances. Another distinct advantage of using a square root implementation of the RTS Smoother is that it is numerically stable for the covariance calculations. This was supported experimentally by numerical results obtained from running the usual form of the RTS Smoother, Bierman's formulation of the RTS Smoother, and our square root implementation of Bierman's formulation of the RTS Smoother in both single and double precision to obtain estimates of the accuracy of the smoothed estimates and covariances for the three algorithms. The results are given in section 5.5 for two examples.

5.3 Bierman’s algorithm.

In this section a modified version of Bierman’s (1983) approach for obtaining smoothed estimates and their respective covariances will be given.
As a preliminary the following signal model will be considered

\[ x_j = x_{j-1} + b w_j \quad j=1, \ldots, n. \]  

which has no associated observation equation. The vectors \( x_j \) and \( b \) are of dimension \( p \) and \( w_j \) is a scalar with zero mean and unit variance. In this special case the Kalman Filter collapses to the following equations

\begin{align}
    x_{j+1|j} &= x_j | j \\
    x_{j+1|j+1} &= x_{j+1|j} \\
    S_{j+1|j} &= S_j | j + b b^T \\
    S_{j+1|j+1} &= S_{j+1|j}. 
\end{align}

\( j=1, \ldots, n-1. \)

The RTS Smoother becomes

\begin{align}
    x_j | n &= x_j | j + A_j (x_{j+1|n} - x_{j+1|j+1}) \\
    S_j | n &= S_j | j + A_j (S_{j+1|n} - S_{j+1|j+1}) A_j^T \quad j=n-1, \ldots, 1.
\end{align}

where

\[ A_j = S_j | j (S_{j+1|j+1})^{-1}. \]

Rearranging \( 5.3.2c \) as an expression for \( S_j | j \) and substituting this into \( 5.3.3c \) we have that \( A_j \) is given by the rank 1 matrix

\[ A_j = I - b v^T \]

where

\[ v = (S_{j+1|j+1})^{-1} b. \]

Substituting this expression for \( A_j \) into \( 5.3.3a \) and \( 5.3.3b \), using \( 5.3.2c \) and simplifying gives

\begin{align}
    x_j | n &= x_{j+1|n} + b v^T (x_{j+1|j+1} - x_{j+1|j}) \\
    S_j | n &= (I - b v^T) S_{j+1|n} (I - b v^T)^T + (1-v^T b) b b^T. \quad (5.3.4b)
\end{align}

From equations \( 5.3.2a \) and \( 5.3.2b \) the term \( x_{j+1|j+1} \) is equivalent
to $x_{1|1}$. The formula for the smoothed covariance is a rank II matrix plus a correction term. This formulation is desirable because it lends itself to numerically stable implementation. The model (5.3.1), and recursion (5.3.2) and (5.3.4) are pivotal to Bierman’s development of his algorithm.

To make use of (5.3.4) it is necessary to organize (5.2.1) into a series of steps each of which has the form (5.3.1) with uncorrelated noise terms. Here this is done by rewriting the state equation (5.2.1) as

$$x_j = T_j x_{j-1} + L_j w_j, \quad w_j = (L_j)^{-1} w_j$$  \hspace{1cm} (5.3.5)

where $L_j$ is the Cholesky factor of the covariance $\Omega_j$ which for the present is assumed to be positive definite. The quantity $L_j^{-1}$ is never evaluated. It is worthwhile noting here that any square root factorization of the covariance $\Omega_j$ will suffice to make $w_j$ have identity covariance which is equivalent to the elements of $w_j$ being uncorrelated. We choose the Cholesky decomposition because it is readily calculated. Let $w_j(k)$ denote the k'th element of the vector $w_j$. The state equation (5.2.1) can be expressed in terms of $w_j(k)$ by considering operating on the columns of

$$L_j = [\mathbf{a}_{j1}, \ldots, \mathbf{a}_{jp}]$$

where

$$\mathbf{a}_{jk} \in \mathbb{R}^p, \ k=1, \ldots, p.$$  

This gives

$$x_j = \mathbf{T} x_{j-1} + \mathbf{a}_{j1} w_j(1) + \ldots + \mathbf{a}_{jp} w_j(p)$$  \hspace{1cm} (5.3.6)

which can be written recursively as

$$\mathbf{x}_j^{(D)} = T_j x_{j-1}$$  \hspace{1cm} (5.3.7a)

$$\mathbf{x}_{ji+1} = \mathbf{x}_{ji} + \mathbf{a}_{ji} w_j(1) \quad \text{if} \quad i=1, \ldots, p.$$  \hspace{1cm} (5.3.7b)

where
\[ \mathbf{x}_{j1} = \mathbf{x}_j^{(D)} \]  

(5.3.7c)

This gives a deterministic update (5.3.7a) and a noise process update (5.3.7b). The state equation (5.2.1) can now be replaced with (5.3.7a), (5.3.7b) and (5.3.7c). The recursion (5.3.7) considered as a state space formulation does not contain an observation equation.

To obtain the solution for one step prediction the special case Kalman Filter (5.3.2) is applied to (5.3.7) which gives

\[ \hat{x}_j = \hat{T}_j x_{j-1} | j-1 \]  

(5.3.8a) \hspace{1cm} \[ \hat{y}_j = \hat{T}_j s_{j-1} | j-1 \hat{T}_j^T \]  

(5.3.8d)

\[ \hat{x}_{j+1} = \hat{x}_j \]  

(5.3.8b) \hspace{1cm} \[ \hat{y}_{j+1} = \hat{y}_j + \hat{g}_j \hat{x}_j \]  

(5.3.8e)

\[ i=1,...,p. \]

where

\[ \hat{x}_{j1} = \hat{x}_j^{(D)} \]  

(5.3.8c)

and

\[ \hat{y}_{j1} = \hat{y}_j^{(D)} \]  

(5.3.8f)

Note that \( \hat{x}_{ji} \) is being used to represent \( x_{ij} | j \) and \( \hat{y}_{ji} \) to represent \( \hat{y}_{ij} | j \). The recursion (5.3.8) gives

\[ \hat{x}_{jp+1} = x_{j} | j-1 \]  

(5.3.8g)

and

\[ \hat{y}_{jp+1} = s_{j-1} | j-1. \]  

(5.3.8h)

Upper case bold script letters are used to denote embedded recursions. The subscript \( j \) indicates that the \( j \)'th state equation is of interest while the subscript \( i \) indicates that the \( i \)'th embedded recursion is of interest.
The special case RTS Smoother (5.3.3a) is applied to (5.3.7) to obtain

\[ \mathbf{X}_{ji+1|n} = \mathbf{X}_{ji+1|n} + \mathbf{S}_{ji} \mathbf{j}_i^{(x_j|j-1 - \mathbf{X}_{ji+1|n})} \]

\[ i=p, \ldots, 1. \]

\[ \mathbf{x}_{j-1|n} = \mathbf{T}_j^T \mathbf{X}_{j1|n} \]

\[ j=n, \ldots, 1. \]

as the recursion for the smoothed estimate where

\[ \mathbf{y}_{ji} = (\mathbf{y}_{ji+1})^{-1} \mathbf{s}_{ji}. \]

To obtain \( \mathbf{y}_{ji} \), \( i=1, \ldots, p \), a forward pass of the recursion (5.3.8) for the covariances is required. If a Cholesky decomposition is applied to \( \mathbf{y}_{ji+1} \) the following equation

\[ \mathbf{y}_{ji} = \mathbf{y}_{ji+1} (\mathbf{y}_{ji+1})^{-1} \mathbf{y}_{ji} \]

can be solved for \( \mathbf{y}_{ji} \) by a forward and backward substitution thus avoiding inverting \( \mathbf{y}_{ji+1} \). The special case RTS Smoother (5.3.3b) is applied to (5.7) to obtain

\[ \mathbf{y}_{ji+1|n} = \mathbf{S}_j \mathbf{n} \]

\[ \mathbf{y}_{ji|n} = (I - \mathbf{y}_{ji} \mathbf{y}_{ji}^T) \mathbf{y}_{ji+1|n} (I - \mathbf{y}_{ji} \mathbf{y}_{ji}^T) + \eta_{ji} \mathbf{s}_{ji} \mathbf{s}_{ji}^T \]

\[ i=p, \ldots, 1. \]

\[ \mathbf{s}_{j-1|n} = \mathbf{T}_j^T \mathbf{y}_{j1|n} \mathbf{y}_{j1|n}^T \]

\[ j=n, \ldots, 1. \]

as the backward recursion for the smoothed covariance where

\[ \eta_{ji} = (I - \mathbf{y}_{ji} \mathbf{y}_{ji}^T). \]

To identify (5.3.9) and (5.3.10) with the recursions obtained by Bierman (1983) we make a LDL^T decomposition on the state covariances \( \Omega_j \) where the lower triangular matrix \( L \) has unit entries along its principal diagonal and \( D \) is a diagonal matrix with elements
(q_{j1}, \ldots, q_{jp})$ and make the replacement
$$\bar{\gamma}_{ji} \leftarrow \lambda_{ji} \gamma_{ji},$$
where
$$\lambda_{ji} = (1 + \gamma_{ji}^T \Sigma_{ji} q_{ji}^{-1}).$$

As mentioned earlier, $L_j^{-1}$ is never evaluated. Bierman's algorithm will now be modified to handle $\Omega_j$ being positive semi definite, as follows. Recollect that since by definition $\Omega_j$ is a covariance matrix it must be symmetric. Thus there exists a non singular matrix $G_j$ such that
$$G_j \Omega_j G_j^T = \Lambda_j,$$
where $\Lambda_j$ is a diagonal matrix, furthermore $\Omega_j$ being positive semi definite restricts the entries of $\Lambda_j$ to being non negative which means that $\Omega_j$ possesses the square root $G_j^{-1} \Lambda_j^{1/2}$. Since square roots of matrices are determined up to an orthogonal matrix we can convert $G_j^{-1} \Lambda_j^{1/2}$ to a lower triangular matrix $L_j$ by appropriate multiplications by orthogonal transformations. If $\Omega_j$ is not of full rank neither will $L_j$ be of full rank. Let
$$L_j = [\Sigma_{j1}, \ldots, \Sigma_{jp}],$$
we are now in the position to use the formulation of the Kalman Filter and the RTS Smoother given in this section. The crucial step is that $w_j$ given in (5.3.5), even when $\Omega_j$ is positive semi definite, still has a covariance which is diagonal, albeit one which may have some entries that have to be interpreted as approaching infinity (although we never need evaluate it), so the elements of $w_j$ are still equivalent to being uncorrelated. Bierman (1983) didn't consider using his algorithm for the state transition covariance $\Omega_j$ being positive semi definite, he argued that if this occurred then it served as a warning that the state space formulation needed reformulation. This will not always be
the case in the following chapters.

5.4 A square root algorithm.

As noted in the preceding section, the $\mathbf{F}_j (i=1,\ldots,p)$ are required to apply the recursions (5.3.9) and (5.3.10) for given $j$ which in turn requires the quantities $\mathbf{G}_{ji}^{-1}$ given $j (i=1,\ldots,p)$. First of all a square root algorithm for updating $\mathbf{G}_{ji}$ will be developed. To do this the recursion (5.3.8d), (5.3.8c) and (5.3.8f) will be exploited.

A square root Kalman Filter is used initially to obtain the square roots of the covariances $(S^1_j|j)^T$. The matrix $(S^1_j|j)^T$ is kept in upper triangular form. The quantity $(S^1_j|j)^T$ initiates the forward recursion for $(\mathbf{G}^1_j|j)^T$ (c.f. (5.3.8f)). The result

$$(\mathbf{G}^1_j|j)^T = (S^1_j|j)^T \mathbf{T}_j.$$

(5.4.1)

when computed is not necessarily an upper triangular matrix but it is determined up to an orthogonal matrix which can be chosen to restore upper triangular form. Thus (5.4.1) is premultiplied by orthogonal transformations to restore it to the required form. Using a similar approach orthogonal transformations are used to obtain $(\mathbf{G}^1_j|j+1)^T$ by reducing

$$
\begin{bmatrix}
(\mathbf{G}^1_j|j)^T \\
(\mathbf{G}^1_j|j)^T
\end{bmatrix}
$$

to

$$
\begin{bmatrix}
(\mathbf{G}^1_j|j+1)^T \\
0
\end{bmatrix},
$$

(5.4.2)

for $i=1,\ldots,p$, where $(\mathbf{G}^1_j|j)^T$ and $(\mathbf{G}^1_j|j+1)^T$ are constrained to be upper triangular matrices. This can be done efficiently by a series of elementary reflectors where only one element of the last row is zeroed at a time starting with the the (1,1)'th element to zero the (p+1,1)'th element and finishing with using the (p,p)'th element to
zero the \((p+1,p)\)’th element. More details about this can be found in Golub and Van Loan (1983). Then the following equation

\[
\mathbf{s}_{ji} = \mathbf{g}_{ji}^{1/2} (\mathbf{g}_{ji+1}^{1/2})^\top \mathbf{g}_{ji}^{1/2}
\]

is solved for \(\mathbf{g}_{ji}^{1/2}\) by a backward and forward substitution thus avoiding inverting \(\mathbf{g}_{ji}^{1/2}\).

To downdate \((\mathbf{y}_{ji+1|n})^\top\); which is kept as an upper triangular matrix; in the embedded recursion for the square root of the smoothed covariance the square root of the leading term in (5.3.10b) is first formed. The quantity

\[
(\mathbf{y}_{ji+1|n})^\top = (\mathbf{S}_{ji+n})^\top
\]

initializes this backward recursion where \((\mathbf{S}_{ji+n})^\top\) is kept in upper triangular form. By inspection of (5.3.10b) the square root of the first term is

\[
(\mathbf{y}_{ji+1|n})^\top - (\mathbf{y}_{ji+1|n})^\top \mathbf{g}_{ji}^{1/2} \mathbf{g}_{ji}^{1/2}.
\]

The term \((\mathbf{y}_{ji+1|n})^\top \mathbf{g}_{ji}^{1/2}\) can be transformed to \(\sigma^2 \mathbf{e}_1\) by a series of orthogonal transformations, \(Q^\top\), in such a way that \(Q^\top (\mathbf{y}_{ji+1|n})^\top\) is upper Hessenberg. This is achieved by introducing one zero at a time into \((\mathbf{y}_{ji+1|n})^\top \mathbf{g}_{ji}^{1/2}\) starting with its last element and ascending to its second element using the element immediately above it to send it to zero. Subtracting \(Q^\top (\mathbf{y}_{ji+1|n})^\top \mathbf{g}_{ji}^{1/2}\) from \(Q^\top (\mathbf{y}_{ji+1|n})^\top\) only alters the first row of \(Q^\top (\mathbf{y}_{ji+1|n})^\top\) so the quantity

\[
Q^\top (\mathbf{y}_{ji+1|n})^\top - Q^\top (\mathbf{y}_{ji+1|n})^\top \mathbf{g}_{ji}^{1/2} \mathbf{g}_{ji}^{1/2} = \hat{\mathbf{R}}
\]

is also upper Hessenberg. It only takes \(O(p^2)\) operations to reduce an upper Hessenberg matrix \(\hat{\mathbf{R}}\) to an upper triangular matrix \(\mathbf{R}\) by a series of elementary reflectors \(P^\top\); that is,

\[
P^\top \hat{\mathbf{R}} = \mathbf{R}.
\]

For more details see Golub and Van Loan (1983). Once the square root of the leading term in (5.3.10b) is evaluated the square root of the
smoothed covariance can be formed by reducing

\[
\begin{bmatrix}
R \\
\sqrt{\eta}_{ji}^{T}
\end{bmatrix}
\text{ to }
\begin{bmatrix}
(y_{j1}^{1/2}|n)^{T} \\
0
\end{bmatrix}
\] (5.4.7)

by a series of elementary reflectors as outlined for (5.4.2). After going through (5.4.5) to (5.4.7), which implements (5.3.10b) \( p \) times, the quantity \((y_{j1}^{1/2}|n)^{T}\) is obtained. The final step computes

\[
(S_{j-1|n}^{1/2})^{T} = (y_{j1}^{1/2}|n)^{T}T_{j}^{-T}
\] (5.4.8)

which is not necessarily in upper triangular form. This must be restored by orthogonal transformation. Note that both this reduction and that required following (5.4.1) are relatively expensive (of the same order as the internal recursion).

5.5 Examples.

The RTS Smoother, Bierman's (1983) implementation of the RTS Smoother and the square root formulation of the RTS Smoother in section 5.4 were run in single and double precision to obtain estimates of the precision of the smoothed estimates and covariances for two data sets. The data sets being Gallant (1975) (here \( n=72 \)) and the Sunspot data in Pandit and Wu (1983) p487 (here \( n=176 \)).

The following covariance was used for the transition equation (5.2.1)

\[
\Omega_{lk} = \lambda((2p-l-k+1)(p-1)!(p-k)!)^{-1}
\]

\( l=1,\ldots,p, \ k=1,\ldots,p, \)

for \( j=2,\ldots,n \) where the algorithms were used for \( p=1,\ldots,5 \). This covariance structure was chosen because we were interested in fitting to the data a polynomial Smoothing Spline employing a method which made use of the Kalman Filter and RTS Smoother. For more details refer to Wecker and Ansley (1983), Osborne and Prvan (1987a), (1987b).
chapters 6 to 9 of this thesis and references contained therein. The results in Table 5.1 and Table 5.2 are for values of \( \lambda \) which straddle the optimal \( \lambda \) for the Smoothing Spline. The smoothed estimates do not vary with the mode of calculation and we believe they reflect the attainable accuracy. Thus they provide a standard by which to judge the covariance calculations. Looking at the results in Table 5.1 and Table 5.2 it is apparent that Bierman's algorithm is more accurate than the standard implementation of the RTS Smoother and that the square root formulation of the RTS Smoother is more accurate than either algorithm for \( p \geq 3 \). Thus our formulation of the RTS Smoother has two distinct advantages over its competitors by being more accurate for large \( p \) and by ensuring that the smoothed covariances are positive definite.
Table 5.1. Error estimates for Smoothed Estimates and Smoothed Covariances.

<table>
<thead>
<tr>
<th>p</th>
<th>λ</th>
<th>RTS Smoother</th>
<th>Bierman (1983)</th>
<th>Square root Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.1</td>
<td>est $0.36 \times 10^{-9}$</td>
<td>$0.36 \times 10^{-9}$</td>
<td>$0.36 \times 10^{-9}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cov $0.24 \times 10^{-8}$</td>
<td>$0.36 \times 10^{-8}$</td>
<td>$0.38 \times 10^{-8}$</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>est $0.88 \times 10^{-9}$</td>
<td>$0.88 \times 10^{-9}$</td>
<td>$0.88 \times 10^{-9}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cov $0.18 \times 10^{-8}$</td>
<td>$0.29 \times 10^{-8}$</td>
<td>$0.43 \times 10^{-8}$</td>
</tr>
<tr>
<td></td>
<td>0.001</td>
<td>est $0.19 \times 10^{-8}$</td>
<td>$0.19 \times 10^{-8}$</td>
<td>$0.19 \times 10^{-9}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cov $0.93 \times 10^{-9}$</td>
<td>$0.28 \times 10^{-9}$</td>
<td>$0.39 \times 10^{-9}$</td>
</tr>
<tr>
<td></td>
<td>0.0001</td>
<td>est $0.70 \times 10^{-9}$</td>
<td>$0.70 \times 10^{-9}$</td>
<td>$0.70 \times 10^{-9}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cov $0.11 \times 10^{-9}$</td>
<td>$0.12 \times 10^{-9}$</td>
<td>$0.27 \times 10^{-10}$</td>
</tr>
<tr>
<td>2</td>
<td>0.1</td>
<td>est $0.16 \times 10^{-10}$</td>
<td>$0.16 \times 10^{-10}$</td>
<td>$0.16 \times 10^{-10}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cov $0.99 \times 10^{-9}$</td>
<td>$0.10 \times 10^{-8}$</td>
<td>$0.16 \times 10^{-8}$</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>est $0.17 \times 10^{-10}$</td>
<td>$0.17 \times 10^{-10}$</td>
<td>$0.17 \times 10^{-10}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cov $0.93 \times 10^{-9}$</td>
<td>$0.30 \times 10^{-9}$</td>
<td>$0.15 \times 10^{-8}$</td>
</tr>
<tr>
<td></td>
<td>0.001</td>
<td>est $0.18 \times 10^{-10}$</td>
<td>$0.18 \times 10^{-10}$</td>
<td>$0.18 \times 10^{-10}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cov $0.41 \times 10^{-9}$</td>
<td>$0.40 \times 10^{-9}$</td>
<td>$0.19 \times 10^{-9}$</td>
</tr>
<tr>
<td></td>
<td>0.0001</td>
<td>est $0.25 \times 10^{-10}$</td>
<td>$0.25 \times 10^{-10}$</td>
<td>$0.25 \times 10^{-10}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cov $0.63 \times 10^{-9}$</td>
<td>$0.42 \times 10^{-9}$</td>
<td>$0.16 \times 10^{-10}$</td>
</tr>
<tr>
<td>3</td>
<td>0.1</td>
<td>est $0.13 \times 10^{-10}$</td>
<td>$0.13 \times 10^{-10}$</td>
<td>$0.13 \times 10^{-10}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cov $0.41 \times 10^{-8}$</td>
<td>$0.30 \times 10^{-8}$</td>
<td>$0.15 \times 10^{-8}$</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>est $0.91 \times 10^{-11}$</td>
<td>$0.91 \times 10^{-11}$</td>
<td>$0.91 \times 10^{-11}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cov $0.17 \times 10^{-8}$</td>
<td>$0.16 \times 10^{-8}$</td>
<td>$0.74 \times 10^{-9}$</td>
</tr>
<tr>
<td></td>
<td>0.001</td>
<td>est $0.60 \times 10^{-11}$</td>
<td>$0.60 \times 10^{-11}$</td>
<td>$0.60 \times 10^{-11}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cov $0.16 \times 10^{-8}$</td>
<td>$0.20 \times 10^{-8}$</td>
<td>$0.77 \times 10^{-9}$</td>
</tr>
<tr>
<td></td>
<td>0.0001</td>
<td>est $0.26 \times 10^{-12}$</td>
<td>$0.26 \times 10^{-12}$</td>
<td>$0.26 \times 10^{-12}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cov $0.16 \times 10^{-8}$</td>
<td>$0.58 \times 10^{-9}$</td>
<td>$0.42 \times 10^{-9}$</td>
</tr>
<tr>
<td>4</td>
<td>0.1</td>
<td>est $0.24 \times 10^{-10}$</td>
<td>$0.24 \times 10^{-10}$</td>
<td>$0.24 \times 10^{-10}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cov $0.51 \times 10^{-7}$</td>
<td>$0.70 \times 10^{-8}$</td>
<td>$0.26 \times 10^{-8}$</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>est $0.42 \times 10^{-11}$</td>
<td>$0.42 \times 10^{-11}$</td>
<td>$0.42 \times 10^{-11}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cov $0.25 \times 10^{-7}$</td>
<td>$0.67 \times 10^{-8}$</td>
<td>$0.25 \times 10^{-8}$</td>
</tr>
<tr>
<td></td>
<td>0.001</td>
<td>est $0.30 \times 10^{-11}$</td>
<td>$0.30 \times 10^{-11}$</td>
<td>$0.30 \times 10^{-11}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cov $0.12 \times 10^{-8}$</td>
<td>$0.46 \times 10^{-8}$</td>
<td>$0.11 \times 10^{-8}$</td>
</tr>
<tr>
<td></td>
<td>0.0001</td>
<td>est $0.56 \times 10^{-12}$</td>
<td>$0.56 \times 10^{-12}$</td>
<td>$0.56 \times 10^{-12}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cov $0.74 \times 10^{-8}$</td>
<td>$0.24 \times 10^{-8}$</td>
<td>$0.13 \times 10^{-8}$</td>
</tr>
<tr>
<td></td>
<td>est</td>
<td>cov</td>
<td>est</td>
<td>cov</td>
</tr>
<tr>
<td>------</td>
<td>-----</td>
<td>------</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>5</td>
<td>0.1</td>
<td>0.21×10^{-10}</td>
<td>0.45×10^{-7}</td>
<td>0.21×10^{-10}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.78×10^{-6}</td>
<td></td>
<td>0.57×10^{-8}</td>
</tr>
<tr>
<td>0.01</td>
<td>est 0.75×10^{-11}</td>
<td>0.75×10^{-7}</td>
<td>0.75×10^{-8}</td>
<td></td>
</tr>
<tr>
<td></td>
<td>cov 0.45×10^{-6}</td>
<td>0.51×10^{-7}</td>
<td>0.35×10^{-8}</td>
<td></td>
</tr>
<tr>
<td>0.001</td>
<td>est 0.94×10^{-12}</td>
<td>0.94×10^{-7}</td>
<td>0.94×10^{-8}</td>
<td></td>
</tr>
<tr>
<td></td>
<td>cov 0.32×10^{-6}</td>
<td>0.30×10^{-7}</td>
<td>0.31×10^{-8}</td>
<td></td>
</tr>
<tr>
<td>0.0001</td>
<td>est 0.56×10^{-12}</td>
<td>0.56×10^{-7}</td>
<td>0.56×10^{-8}</td>
<td></td>
</tr>
<tr>
<td></td>
<td>cov 0.18×10^{-6}</td>
<td>0.13×10^{-7}</td>
<td>0.32×10^{-8}</td>
<td></td>
</tr>
</tbody>
</table>
Table 5.2 Error Estimates for Smoothed Estimates and Smoothed Covariances.

<table>
<thead>
<tr>
<th>p</th>
<th>λ</th>
<th>RTS Smoother</th>
<th>Bierman (1983)</th>
<th>Square root algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
<td>est 0.89x10^{-9}</td>
<td>0.89x10^{-9}</td>
<td>0.89x10^{-9}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cov 0.16x10^{-7}</td>
<td>0.15x10^{-7}</td>
<td>0.79x10^{-7}</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>est 0.17x10^{-9}</td>
<td>0.17x10^{-9}</td>
<td>0.17x10^{-9}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cov 0.71x10^{-8}</td>
<td>0.71x10^{-8}</td>
<td>0.12x10^{-7}</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>est 0.29x10^{-9}</td>
<td>0.29x10^{-9}</td>
<td>0.29x10^{-9}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cov 0.56x10^{-8}</td>
<td>0.56x10^{-8}</td>
<td>0.63x10^{-8}</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>est 0.73x10^{-9}</td>
<td>0.73x10^{-9}</td>
<td>0.73x10^{-9}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cov 0.18x10^{-9}</td>
<td>0.18x10^{-9}</td>
<td>0.19x10^{-9}</td>
</tr>
<tr>
<td>2</td>
<td>100</td>
<td>est 0.32x10^{-8}</td>
<td>0.32x10^{-8}</td>
<td>0.32x10^{-8}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cov 0.31x10^{-7}</td>
<td>0.31x10^{-7}</td>
<td>0.92x10^{-7}</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>est 0.14x10^{-8}</td>
<td>0.14x10^{-8}</td>
<td>0.14x10^{-8}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cov 0.56x10^{-8}</td>
<td>0.56x10^{-8}</td>
<td>0.63x10^{-8}</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>est 0.27x10^{-9}</td>
<td>0.27x10^{-9}</td>
<td>0.27x10^{-9}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cov 0.87x10^{-9}</td>
<td>0.87x10^{-9}</td>
<td>0.13x10^{-8}</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>est 0.95x10^{-9}</td>
<td>0.95x10^{-9}</td>
<td>0.95x10^{-9}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cov 0.65x10^{-9}</td>
<td>0.65x10^{-9}</td>
<td>0.13x10^{-8}</td>
</tr>
<tr>
<td>3</td>
<td>100</td>
<td>est 0.18x10^{-8}</td>
<td>0.18x10^{-8}</td>
<td>0.18x10^{-8}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cov 0.45x10^{-6}</td>
<td>0.24x10^{-6}</td>
<td>0.43x10^{-7}</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>est 0.17x10^{-8}</td>
<td>0.17x10^{-8}</td>
<td>0.17x10^{-8}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cov 0.15x10^{-7}</td>
<td>0.15x10^{-7}</td>
<td>0.74x10^{-8}</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>est 0.14x10^{-8}</td>
<td>0.14x10^{-8}</td>
<td>0.14x10^{-8}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cov 0.47x10^{-8}</td>
<td>0.47x10^{-8}</td>
<td>0.22x10^{-8}</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>est 0.54x10^{-9}</td>
<td>0.54x10^{-9}</td>
<td>0.54x10^{-9}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cov 0.19x10^{-8}</td>
<td>0.19x10^{-8}</td>
<td>0.85x10^{-9}</td>
</tr>
<tr>
<td>4</td>
<td>100</td>
<td>est 0.27x10^{-9}</td>
<td>0.27x10^{-9}</td>
<td>0.27x10^{-9}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cov 0.97x10^{-6}</td>
<td>0.97x10^{-6}</td>
<td>0.80x10^{-7}</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>est 0.22x10^{-9}</td>
<td>0.22x10^{-9}</td>
<td>0.22x10^{-9}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cov 0.11x10^{-6}</td>
<td>0.11x10^{-6}</td>
<td>0.16x10^{-7}</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>est 0.20x10^{-9}</td>
<td>0.20x10^{-9}</td>
<td>0.20x10^{-9}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cov 0.18x10^{-7}</td>
<td>0.18x10^{-7}</td>
<td>0.48x10^{-8}</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>est 0.10x10^{-9}</td>
<td>0.10x10^{-9}</td>
<td>0.10x10^{-8}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cov 0.36x10^{-8}</td>
<td>0.36x10^{-8}</td>
<td>0.16x10^{-8}</td>
</tr>
</tbody>
</table>
Wecker and Amiel (1992) use a stochastic setting to obtain an algorithm for polynomial spline smoothing. The advantage of this setting are that it facilitates the use of coded and efficient algorithms based on the Kalman Filter and BIB design. That means likelihood estimation can be used to estimate the smoothing parameter \( \lambda \), and thus forecast confidence intervals can be calculated for the smoothing spline. In this chapter their setting will be derived within a more general framework which encompasses smoothing splines.

Let \( L \) be the following differential operator
\[
L = -a_0(x) y'' - a_1(x) y' - a_2(x) y - b(x) \frac{d}{dx} y + c(x)
\]
where \( y(x) \) is \( y(x_0) \). Defining \( L \) to be the differential operator of Eq. (4.1) with \( \lambda = 0 \), one considers \( p \times 2 \) derivatives which are smooth functions. When \( \lambda = 0 \), spline is defined as follows.

**DEFN 4.1.** An \( L \) Smoothing spline provides a smooth curve through the data \( y_i, \{x_i\}_{i=1}^{N} \), at \( x_i \), \( i = 1, \ldots, N \), respectively by minimizing
\[
\| y - f \|^2 = \int (y - f)^2 dx = \int \left( \sum_{i=1}^{N} (y_i - f(x_i))^2 \right) dx
\]
where \( f \in H \). The assumptions that \( f \) is made.

Instead of the assumption that the data can be decomposed as
\[
y_i = f(x_i) + \varepsilon_i, \quad i = 1, \ldots, N.
\]
6.1 Introduction.

Wecker and Ansley (1983) use a stochastic setting to obtain an algorithm for polynomial spline smoothing. The advantages of this setting are that it facilitates the use of stable and efficient algorithms based on the Kalman Filter and RTS Smoother, that maximum likelihood estimation can be used to estimate the smoothing parameter \( \lambda \), and that formal confidence intervals can be attached to the Smoothing Spline. In this chapter their method will be derived within a more general framework which encompasses Smoothing Splines.

Let \( L \) be the following differential operator

\[
L = D^p - a_{p-1}(t)D^{p-1} - \ldots - a_1(t)D - a_0(t)
\]

where \( D^p \equiv d^p/dt^p \). Define \( H \) to be the Hilbert space of functions \( f(t) \), \( t_1 \leq t \leq t_n \), possessing \( p' \)th derivatives which are square integrable then an Lg spline is defined as follows.

DEFN 6.1.1 An Lg Smoothing Spline provides a smooth curve through the data \( y_i, i=1,\ldots,n \), at \( t_i, i=1,\ldots,n \), respectively by minimizing

\[
S(f) = \mu \int (Lf)^2 + \sum_{i=1}^{n} (y_i - f(t_i))^2
\]

where \( f \in H \). The assumption that \( n \geq p \) is made.

Explicit is the assumption that the data can be decomposed as

\[
y_i = f(t_i) + \varepsilon_i, i=1,\ldots,n.
\]
where $\epsilon_i$ is normally distributed with zero mean and $\sigma^2$ variance. The $f(t_i)$ can be thought of as an unknown signal and the $\epsilon_i$ as a noise term. When $L = D^p$ the definition of an $Lg$ Smoothing Spline reduces to that of a polynomial Smoothing Spline. Some authors refer to $Lg$ Smoothing Splines as Generalized Smoothing Splines. In this thesis a different meaning is reserved for this term.

Wahba (1978) obtains a similar result for $Lg$ Smoothing Splines by modelling the signal by the stochastic differential equation

$$Lx = \sigma \lambda (d\omega(t)/dt)$$

(6.1.4)

where $\omega(t)$ is a Wiener process with unit dispersion parameter (see for example Billingsley (1979)) and $\lambda$ is a scale parameter corresponding to the reciprocal of the smoothing parameter $\mu$ which is to be determined. Let

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

be the vector of initial conditions on (6.1.4). She shows that if $x(t_1)$ is allowed to have a diffuse prior distribution; that is, setting $x(t_1) \sim N(0, \gamma^2)$ and letting $\gamma^2 \to \infty$, then

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

(6.1.5)
6.2 State Space Formulation.

Equation (6.1.4) can be formulated in matrix companion form by setting

\[ \begin{align*}
\dot{x}_1(t) &= x_2(t) \\
\dot{x}_2(t) &= x_3(t) \\
&\vdots \\
\dot{x}_{p-1}(t) &= x_p(t) \\
\dot{x}_p(t) &= a_0(t)x_1(t) + \ldots + a_{p-1}(t)x_{p-1}(t) + \sigma \sqrt{2} \frac{d\omega(t)}{dt}
\end{align*} \]

where

\[ x_1(t) = x(t) \]

giving

\[ \dot{x}(t) = Mx(t) + g(t)e_p \tag{6.2.1} \]

where

\[ M = \begin{bmatrix}
0 & & \\
& \ddots & \\
& & I_{p-1}
\end{bmatrix}, \tag{6.2.2}
\]

\[ x(t)^T = [x(t), x(1)(t), \ldots, x(p-1)(t)]^T. \tag{6.2.3} \]

and

\[ g(t) = \sigma \sqrt{2} \frac{d\omega(t)}{dt}. \tag{6.2.4} \]

Assuming that the vector of initial conditions \( x(t_1) \) is a constant vector to be determined then the solution to equation (6.2.1) is

\[ x(t) = T(t, t_1)x(t_1) + \int_{t_1}^{t} T(t, s)e_p g(s)d\omega(s) \tag{6.2.5} \]

where \( T(t, t_1) \) is the fundamental matrix solution for

\[ \dot{x}(t) = Mx(t); \tag{6.2.6} \]
that is, $T(t,t_1)$ satisfies the initial value problem
\[ \frac{d}{dt}T(t,t_1) = MT(t,t_1), \quad T(t_1,t_1) = I. \quad (6.2.7) \]

Two properties of $T$ which will be used are

(i) $T(t,s)T(s,t_1) = T(t,t_1)$ \quad (6.2.8a)

and

(ii) $\frac{d}{dt}T(t,s) = -T(t,s)M$. \quad (6.2.8b)

Equation (6.2.5) can be written recursively as

\[ x(t_{i+1}) = T(t_i,t_{i+1})x(t_i) + u(t_i,t_{i+1}) \quad (6.2.9) \]

where

\[ u(t_{i+1},t_1) = \sigma \sqrt{\int_{t_i}^{t_{i+1}} T(t_{i+1},s)e_p^T(d\omega(s)/ds)ds}. \quad (6.2.10) \]

By definition $u_{i+1}$ is independent of $x(t)$ for $t < t_i$ and is distributed normally with zero mean and covariance given by

\[ \Omega(t_1,t_{i+1}) = \sigma \sqrt{\int_{t_i}^{t_{i+1}} T(t_{i+1},s)e_p e_p^T T(t_{i+1},s)^T ds}. \quad (6.2.11) \]

Utilizing property (6.2.8a) equation (6.2.10) can be written as

\[ u(t_{i+1},t_1) = T(t_{i+1},t_i)u(t_i,t_1) + u(t_{i+1},t) \quad (6.2.12) \]

Wahba's (1978) result suggests choosing the signal $f(t)$ to equal $x(t)$, the first entry of the vector $x(t)$. Then the observation equation (6.1.3) becomes

\[ y_i = e_1^T x_i + \epsilon_i \quad (6.2.13). \]

Substituting for $x_i$ as defined by (6.2.5) gives

\[ y_i = e_1^T T(t_i,t_1)x_1 + e_1^T \int_{t_i}^{t_1} T(t_i,s)e_p g(s)ds + \epsilon_i \quad (6.2.14) \]

which simplifies to the following expression with $x_i = \alpha$ and the integral term replaced by using (6.2.10)

\[ y_i = e_1^T T(t_i,t_1)\alpha + e_1^T u(t_1,t_1) + \epsilon_i. \quad (6.2.15) \]
Setting
\[ z_i = z(t_i) = u(t_i, t_1) \] \hspace{1cm} (6.2.16)
equation (6.2.15) can be rewritten as
\[ y_i = e_{i1}^T T(t_i, t_1) a + e_{i1}^T z_i + \epsilon_i. \] \hspace{1cm} (6.2.17a)
Using (6.2.12) the following recursion is obtained for \( z_{i+1} \)
\[ z_{i+1} = T_{i+1} z_i + u_{i+1} \] \hspace{1cm} (6.2.17b)
where by definition \( z_i \) is independent of \( u_{i+1} \) which has covariance
\( \Omega_{i+1} \) given by (6.2.11).

Equations (6.2.17a) and (6.2.17b) furnish the state space
formulation where the former is the "observation equation" and the
latter is the "state transition equation" which is usually
abbreviated to "state equation". The Kalman Filter can be applied to
this formulation.

6.3 Maximum Likelihood Estimation.

The state space formulation (6.2.17) can be written as a system
of equations in terms of \( a \) plus a noise term. Since the noise term has
a known distribution a likelihood can be attached to the state space
formulation. This likelihood can be used to determine the constant \( a \)
for a given \( \lambda \) after determining \( \sigma^2 \), the residual sum of squares
(RSS), also by maximum likelihood. Once they have been determined the
likelihood or log likelihood can be evaluated for the given value of
\( \lambda \). This approach can be used to determine the optimal \( \lambda \) by a nonlinear
optimization algorithm. For each candidate the RSS and \( a \) have to be
determined in that order before evaluating the function at this \( \lambda \). The
reason that the RSS has to be evaluated is that it appears in the
likelihood calculation.
Letting
\[ y^T = [y_1, \ldots, y_n]^T, \]
\[ X = [r_1, \ldots, r_n]^T \]
where
\[ r_i^T = e_1^T(t_i, t_1), \]
and
\[ v^T = [v_1, \ldots, v_n]^T \]
where
\[ v_i = e_1^T z_i + \epsilon_i, \]
permits (6.2.17) to be written in matrix form as
\[ y = X\alpha + v. \] (6.3.1)

Here \( \alpha \), the vector of initial conditions on (6.1.4), is being treated as an unknown constant. The maximum likelihood estimate (MLE) of \( \lambda \) and \( \sigma^2 \) is obtained by treating (6.3.1) as a generalized linear least squares problem. Let \( V \) denote the covariance of \( v \) in (6.3.1). The \((i,j)\)'th element of \( V \) for \( i \neq j \) is
\[ E(v_i v_j) = E[(e_1^T z_i + \epsilon_i)(e_1^T z_j + \epsilon_j)] \]
\[ = E(\epsilon_i \epsilon_j) + E[e_1^T z_i e_1] \]
\[ = \delta_{ij} \sigma^2 + e_1^T E[(\sigma \Lambda \int T(t_i, s) e_p d\omega(s) + \sigma \Lambda \int T(t_j, s) e_p d\omega(s))] \]
\[ t_i \]
\[ t_j \]
\[ t_1 \]
\[ t_1 \]
\[ t_1 \]
\[ = \delta_{ij} \sigma^2 + \lambda^2 e_1^T T(t_i, t_j) E[\int T(t_j, s) e_p e_p^T T(t_j, s) d\omega] e_1 \]
\[ t_1 \]
\[ t_1 \]
\[ t_1 \]
\[ = \delta_{ij} \sigma^2 + e_1^T T(t_i, t_j) \Omega(t_j, t_1) e_1. \] (6.3.2)

Thus the covariance matrix of \( v \) has the form \( \sigma^2 \Lambda \) where the matrix \( \Lambda \) is a function of the scale parameter \( \lambda \) only.
The likelihood of (6.3.1) is
\[ l = l(\sigma^2, \lambda, \alpha | y) = \left[ (2\pi \sigma^2)^{n/2} |\lambda|^{1/2} \right]^{-1} e^{-\frac{1}{2}(y - X\alpha)^T \lambda^{-1} (y - X\alpha)/(2\sigma^2)}. \]
Taking the log likelihood of this gives
\[ \ell = \ell(\sigma^2, \lambda, \alpha | y) = \text{constant} - \frac{n}{2} \log(\sigma^2) - \frac{1}{2} \log(|\lambda|) - \frac{1}{2} \left( \frac{y - X\alpha}{\sigma^2} \right)^T \lambda^{-1} \left( \frac{y - X\alpha}{\sigma^2} \right). \]
(6.3.3)

where
\[ Q = (y - X\alpha)^T \lambda^{-1} (y - X\alpha). \]

Minimizing over \( \sigma^2 \) furnishes \( \sigma^2 = n^{-1}Q \). Substituting this result into (6.3.3) yields, after absorbing all constants,
\[ \ell^*(\lambda, \alpha | y) = \text{constant} - \frac{n}{2} \log(Q) - \frac{1}{2} \log(|\lambda|). \]
(6.3.4)

Let \( L \) be the Cholesky factor of \( \Lambda \) then (6.3.4) can be rewritten as
\[ \ell^*(\lambda, \alpha | y) = \text{constant} - \frac{n}{2} \log(\left( y - X\alpha \right)^T L^{-1} L^{-1} (y - X\alpha)) - \frac{1}{2} \log(||L^T||) \]
which is maximized over \( \alpha \) if \( Q \) is minimized over \( \alpha \) since the last term is independent of \( \alpha \). Doing this implies that \( \hat{\alpha} \), the minimizer of \( Q \), is obtained by regressing \( L^{-1}y \) on \( L^{-1}X \) which gives
\[ \hat{\alpha} = ((L^{-1}X)^T (L^{-1}X))^{-1} (L^{-1}X)^T L^{-1} y \]
and residuals
\[ \alpha = L^{-1}(y - X\alpha). \]

Thus the log likelihood concentrated over \( \sigma^2 \) and \( \alpha \) is
\[ \ell^**(\lambda | y, X) = \text{constant} - \frac{n}{2} \log(\alpha^T \alpha) - \frac{1}{2} \log(|\lambda|). \]
(6.3.5)

Then (6.3.5) can be used to maximize over \( \lambda \) by any standard non linear estimation method. For each \( \lambda \) used the values of \( \sigma^2 \) and \( \alpha \) have to be determined to evaluate \( \ell^**(\lambda | y, X) \). Since the log likelihood (6.3.5) only depends on one variable \( \lambda \), a line search method which uses only function values will suffice. To evaluate this log likelihood the quantities \( L^{-1}y \), \( L^{-1}X \) and \( |\lambda| \) are required. This is where the Kalman Filter enters as a computational tool.
6.4 The Kalman Filter as a computational tool.

The Kalman Filter equations for the state space formulation

\[ z_i = T_i z_{i-1} + u_i \] (6.4.1)

\[ y_i - e_i^T(t_i) \alpha = e_i^T z_i + \epsilon_i \] (6.4.2)

where \( u_i \) is given by (6.2.10), \( \epsilon_i \) is normally distributed with zero mean and variance \( \sigma^2 \) and \( z_{i-1} \). \( u_i \), \( \epsilon_j \) are mutually independent for \( 1 \leq i \leq n \), where here \( 1 \leq i \leq n \), from (2.3.5) - (2.3.10) are

\[ z_i | i-1 = T_i z_{i-1} | i-1 \] (6.4.3a)

\[ S_i | i-1 = T_i S_{i-1} | i-1 T_i^T + \bar{\Omega}_i \] (6.4.3b)

\[ d_i = e_i^T S_i | i-1 e_i + 1 \] (6.4.3c)

\[ x_i^* = (y_i - \sum_{k=1}^{p} \alpha_k X_{ik} - e_i^T z_{i-1} t_{i-1})/(d_i^{1/2}) \] (6.4.3d)

\[ z_i | i = z_i | i-1 + S_i | i-1 e_i(y_i - \sum_{k=1}^{p} \alpha_k X_{ik} - e_i^T z_{i-1} t_{i-1})/d_i \] (6.4.3e)

and

\[ S_i | i = S_i | i-1 - S_i | i-1 e_i^T S_i | i-1 (d_i)^{-1}. \] (6.4.3f)

where here

\[ S_i | i = \sigma^{-2} \text{cov}(z_i | i), \]

\[ S_i | i-1 = \sigma^{-2} \text{cov}(z_i | i-1) \]

and

\[ d_i = \sigma^{-2} \text{cov}(y_i - h^T z_{i-1}). \]

The quantity \( \bar{\Omega}_i \) is the covariance of \( \sigma^{-1} u_i \). The sequence \( \{X_i^*\} \) is the Gram-Schmidt orthonormalization of the sequence \( \{y_i - \sum_{k=1}^{p} \alpha_k X_{ik}\} \) by construction of the innovations (residual) sequence by the Projection Theorem (refer to PROBLEM 3.3.1 and THEOREM 3.3.3) divided by the respective standard deviations. Thus

(i) \( \{X_i^*\} \) is an orthogonal sequence
(ii) \( \{X^*_i\} \) has constant variance \( \sigma^2 \) and

\[
(iii) X^*_i = \sum_{j=1}^{i} \beta_{ij} (y_j - \sum_{k=1}^{p} \alpha_{jk} X_{ik}) \quad \text{where } \beta_{ij} \text{ form the entries of a lower triangular matrix.}
\]

Let

\[
v_i = e_i^T z_i + \epsilon_i
\]

where

\[
v^T = [v_1, \ldots, v_n], \quad \text{E}(v) = 0
\]

and

\[
\text{E}(vv^T) = \Lambda \sigma^2
\]

where \( \Lambda \) is a positive definite matrix. Applying the Cholesky decomposition to \( \Lambda \) gives

\[
\Lambda = LL^T
\]

where \( L \) is lower triangular, note that \( L^{-1} \) is also of the same form.

Defining

\[
L^{-1}(y - X\alpha) = X^***
\]

(6.4.4)

yields

\[
\text{E}[L^{-1}(y - X\alpha)(L^{-1}(y - X\alpha))^T] = \sigma^2 I
\]

which means that \( X^***_i \), the \( i \)th element of \( L^{-1}(y - X\alpha) \), is a member of an orthogonal sequence with constant variance \( \sigma^2 \). From (6.4.4)

\[
X^***_i = \sum_{j=1}^{i} \hat{\beta}_{ij} (y_j - \sum_{k=1}^{p} \alpha_{jk} X_{ik}) \quad \text{where } \hat{\beta}_{ij} , j=1, \ldots, i, \text{ are constants which can be determined.}
\]

Since Gram-Schmidt orthonormalization is essentially unique

\[
X^***_i = X^*_i \quad \text{i=1, \ldots, n.}
\]

Thus equation (6.4.3d) can be used to determine \( L^{-1}(y - X\alpha) \) once \( \alpha \) is known, this equation relies on the estimates obtained from the Kalman Filter.
The estimate of \( \alpha \) can be found by GLS. To do this
\[ L^{-1}y, L^{-1}X_i, \ldots, L^{-1}X_p, \]
where \( X_i \) is the \( i \)'th column of \( X \), need to be computed. Referring to equation (6.4.3d) if the \( z_i \) is split into components
\[ z_i = z_{0i} - \sum_{k=1}^{p} \alpha_k z_{ki} \]
it can be rewritten as
\[ X_i* = (y_i - e_i^T z_{0i}) i-1 - \sum_{k=1}^{p} \alpha_k (X_i^k - e_i^T z_{ki}) i-1) / d_i. \]
This is linear in \( z_{0i} i-1 \ldots z_{pi} i-1 \). Making similar substitutions into the remaining equations in (6.4.3) it can also be seen that they are linear in \( z_{0i} i \ldots z_{pi} i \) and \( z_{0i} i-1 \ldots z_{pi} i-1 \). The equations for the filtered and predicted covariances remain the same. Thus the Kalman Filter can be used to obtain \( z_{ji} i, z_{ji} i-1, S_i i, S_i i-1 \) and \( \dot{X}_{ji} \) where \( z_{j1} i = 0 \) and \( S_1 i = 0 \) are obtained from (6.2.16). This leads to \( p+1 \) Kalman Filter recursions being used to form
\[ X_0 = L^{-1}y, X_1* = L^{-1}X_1 \ldots, X_p* = L^{-1}X_p \]
where \( X_0 = y \). The required recursions are
\[ z_{ji} i-1 = T_i z_{ji} i-1 i-1 \quad (6.4.5a) \]
\[ S_i i-1 = T_i S_i i-1 i-1 T_i^T + \tilde{U}_i \quad (6.4.5b) \]
\[ d_i = e_i^T S_i i-1 e_i + 1 \quad (6.4.5c) \]
\[ X_{ji} = (X_{ji} - e_i^T z_{ji} i-1) / d_i \quad (6.4.5d) \]
\[ z_{ji} i = z_{ji} i-1 + S_i i-1 e_i (X_{ji} - e_i^T z_{ji} i-1) / d_i \quad (6.4.5e) \]
and
\[ S_i i = S_i i-1 - S_i i-1 e_i e_i^T S_i i-1 (d_i)^{-1} \quad (6.4.5f) \]
\[ j=0, \ldots, p, \ i=1, \ldots, n. \]
Here \( S_i i-1 = \sigma^2 \text{Cov}(z_i i-1) \) and \( S_i i = \sigma^2 \text{Cov}(z_i i) \).
Now

$$|\Lambda|^{-1} = |L^{-1}|^2,$$

but $L^{-1}$ is lower triangular and has along its diagonal the elements $(1/\sqrt{d_i})$ hence

$$|\Lambda| = \prod_{i=1}^{n} d_i.$$

The quantity $\hat{\alpha}$ can be found by minimizing

$$\|L^{-1}y - L^{-1}x\alpha\|_2^2.$$  \hspace{1cm} (6.4.6)

This is now a linear least squares problem whose solution can be found efficiently by premultiplying (6.4.6) by an orthogonal matrix $Q^T=[Q_1|Q_2]^T$ which is chosen to transform $L^{-1}x$ into the upper triangular matrix $R$. Since the 2 norm is unaffected by orthogonal transformations

$$\hat{\alpha} = \text{arg min}_{\alpha} \| \begin{bmatrix} Q_1^T y_1 \\ Q_2^T y_2 \end{bmatrix} - \begin{bmatrix} R \\ 0 \end{bmatrix} \alpha \|_2^2$$

which gives

$$R\hat{\alpha} = Q_1^T y_1$$

where $\bar{y}_1$ is the first $p$ elements of $L^{-1}y$ and $\bar{y}_2$ is the remaining $n-p$ elements of $L^{-1}y$. The quantity $\alpha$ then can be obtained by back substitution. Also the residual sum of squares (RSS) is equal to

$$\|Q_2^T y_2\|_2^2.$$  \hspace{1cm} (6.4.7)

Hence

$$\hat{\sigma}^2 = \text{RSS}/n.$$  \hspace{1cm} (6.4.8)

A first estimate $z_i|_1$ of $z_i$ in (6.2.17b) is

$$z_i|_1 = z_0|_1 - \sum_{j=1}^{p} \alpha_j z_j|_1.$$  \hspace{1cm} (6.4.9)

This leaves the calculation of the point estimate $\hat{f}(t)$ which is

$$f(t) = e_1^T T(t,t_1)\alpha + e_1^T z(t).$$  \hspace{1cm} (6.4.9)

Here it is assumed that the estimates for $\alpha$, the initial estimates of the states $z_i$, $z_i|_1$, $i=1,...,n$, and its associated covariance $S_i|_1$
have been obtained as well as $z_{i+1|i}$ and its associated covariance $S_{i+1|i}$. These estimates of $z_i$ aren't based on all of the available data. This can be rectified by using the RTS Smoother ((3.3.4)-(3.3.6)) which for the state space formulation (6.2.17) is

$$
\begin{align*}
Z_k|n &= Z_k|k + A_k (z_{k+1|n} - z_{k+1|k}) \\
S_k|n &= S_k|k + A_k (S_{k+1|n} - S_{k+1|k})^T A_k^T
\end{align*}
$$

(6.4.10a)

(6.4.10b)

where

$$
A_k = S_k|k (A_{k+1} S_{k+1|k})^{-1}
$$

(6.4.10c)

$k=n,\ldots,1$.

The initial conditions which start the algorithm, $z_n|n$ and $S_n|n'$, are obtained after (6.4.5) has been run. Also wanted are estimates of $z(t)$ at points other than $t_i$ based on all of the available data. The Interpolation Smoother ((3.4.1)-(3.4.3)) provides these estimates. For the state space formulation (6.2.17) for $t_{i-1} \leq t \leq t_i$ they are

$$
\begin{align*}
z(t|n) &= T(t,t_{i-1})z_{i-1|i-1} + A(t_i,t)(z_i|n - z_i|i-1) \\
S(t|n) &= \Omega(t,t_{i-1}) + T(t,t_{i-1})S_{i-1|i-1}T(t,t_{i-1})^T \\
&\quad - A(t_i,t)(S_{i-1|i-1} - S_{i-1|1})A(t_i,t)^T
\end{align*}
$$

(6.4.11a)

(6.4.11b)

where

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

(6.4.11c)

and

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

(6.4.11d)

This reduces to the RTS Smoother when $t=t_i$. The estimate of $f(t)$, say $\hat{f}(t)$, based on all of the data can be found. For $t_{i-1} \leq t \leq t_i$, $i=2,\ldots,n$, it is,

$$
\hat{f}(t) = e_{t_1}^T (T(t,t_{1})\alpha + e_{t_1}^T z(t|n)).
$$

(6.4.12a)

For $t<t_1$

$$
\begin{align*}
f(t) &= e_{t_1}^T x(t) \\
&= e_{t_1}^T (T(t,t_{1})\alpha + u(t,t_{1}))
\end{align*}
$$
which gives
\[ \hat{f}(t) = e_1^T T(t,t_1) \hat{\alpha} \]  
(6.4.12b)
as the estimate of \( f(t) \) based on all of the data. For \( t > t_n \)
\[ f(t) = e_1^T x(t) \]
\[ = e_1^T [T(t,t_1) \alpha + z(t)] \]
\[ = e_1^T T(t,t_1) \alpha + e_1^T [T(t,t_n) z(t_n) + u(t,t_n)] \]
which yields
\[ \hat{f}(t) = e_1^T T(t,t_1) \hat{\alpha} + e_1^T T(t,t_n) z_{n|n} \]
(6.4.12c)
as the estimate of \( f(t) \) based on all of the data.

In the next section the continuity properties of \( \hat{f}(t) \) will be investigated.

6.5 Smoothness properties.

If \( \hat{f}(t) \) given by (6.4.12) satisfies \( LL^+ f(t) = 0 \) and has (2p-2) continuous derivatives at the knot points then it corresponds to an \( Lg \) Smoothing Spline. Since \( f(t) \) is a function of the elements of \( z(t|n) \) it makes sense to investigate the smoothness properties of this quantity. For \( \hat{f}(t) \) to possess (2p-2) continuous derivatives at the data points, \( z(t|n) \) has to possess (p-1) continuous derivatives at the knot points because of the structure imposed by the matrix companion form of (6.2.1). The following lemma provides the result necessary for \( \hat{f}(t) \), given by (6.4.12), to possess (2p-2) continuous derivatives at the knot points.
LEMMA 6.5.1 The first \((p-1)\) derivatives of \(z(t|n)\), given by (6.4.11), are continuous at the data points.

PROOF. Differentiating (6.4.11a), for \(t_{i-1} \leq t \leq t_i\), gives

\[
\frac{d}{dt} z(t|n) = MT(t, t_{i-1})z_{i-1|i-1} + \frac{d}{dt} A(t_i, t)(z_i|n - z_i|i-1)
\]  

(6.5.1)

where \(\Gamma(t_i, t)\) in the definition of \(A(t_i, t)\) is rewritten as

\[
\Gamma(t_i, t) = \int_{t_{i-1}}^{t} T(t,s)e^{T_{1}\cdot t} ds;
\]

(6.5.2)

and

\[
\frac{d}{dt} A(t_i, t)(MT(t, t_{i-1})S_{i-1|i-1} + \frac{d}{dt} \Gamma(t_i, t))S_{i|i-1}^{-1}
\]

(6.5.3)

which are obtained by differentiating (6.4.11b) and (6.4.11c) respectively. Substituting (6.5.2) and (6.5.3) into (6.5.1) and rearranging gives

\[
\frac{d}{dt} z(t|n) = Mz(t|n) + \lambda e e^T T(t_i, t) S_{i|i-1}^{-1}(z_i|n - z_i|i-1).
\]

(6.5.4)

As a consequence the states \(z(t|n)\) between the two data points have the same smoothness as the solution to the homogeneous differential equation. From this observation it follows that the states evaluated at the data points, \(t_i, i=1, \ldots, n\), are of most interest in determining the smoothness of \(z(t|n)\).

Let \(D^j\) be the jump in the \(j\)’th derivative of \(z(t|n)\) at the point \(t_i\). Then

\[
p^j = \frac{d}{dt} z(t_i^+|n) - \frac{d}{dt} z(t_i^-|n)
\]

\[
= 2\lambda e e^T T_{1+i|1+1} S_{1+i+1|i+1}^{-1}(z_{1+i+1|i+1} - z_{1|i+1} - z_{i+1|i} - z_{1|i-1}).
\]

(6.5.5)

Continuity of \(\frac{d}{dt} z(t|n)\) at \(t=t_i\) requires that \(D^1=0\). Substituting
t = t_i into (6.4.11a) and rearranging provides

\[ T_i + 1 = S_{i+1|i} = S_{i+1|i}^{-1}(z_{i+1|i} - z_{i+1|i-1}) = S_{i+1|i}^{-1}(z_{i+1|i} - z_{i+1|i-1}). \]  

(6.5.6)

Using the matrix inversion lemma on \( S_{i+1|i} \) defined by (6.4.3f) produces the following result

\[ S_{i+1|i}^{-1} = S_{i+1|i-1} + e_i e_i^T. \]  

(6.5.7)

Substituting (6.5.6) into (6.5.5) gives

\[ D^1 = \lambda e_p e_p^T S_{i+1|i}^{-1}(z_{i|n} - z_{i|i}) - S_{i+1|i-1}(z_{i|n} - z_{i|i-1}). \]  

(6.5.8)

From the Kalman Filter (6.4.3)

\[ z_{i|i} = z_{i|i-1} + S_{i|i-1} e_1(y_i - e_1^T z_{i|i-1}) / d_i \]

where

\[ d_i = e_1^T z_{i|i-1} + 1. \]

Substituting this into (6.5.8) and using (6.5.7) yields

\[ D^1 = \lambda e_p e_p^T e_1 e_1^T(z_{i|n} - z_{i|i-1}) - S_{i|i} e_1(y_i - e_1^T z_{i|i-1}) / d_i \]

\[ = \lambda e_p e_p^T e_1 e_1^T z_{i|n} - e_1 e_1^T z_{i|i-1} - e_1(y_i - e_1^T z_{i|i-1}) / d_i \]

\[ = \lambda e_p e_p^T e_1 e_1^T(z_{i|n} - e_1 e_1^T z_{i|i-1}) - e_1(y_i - e_1^T z_{i|i-1}) / d_i \]

\[ = \lambda e_p e_p^T e_1 e_1^T z_{i|n} - e_1 e_1^T z_{i|i-1} - e_1(y_i - e_1^T z_{i|i-1}) / d_i \]

\[ = \lambda e_p e_p^T e_1 e_1^T z_{i|n} - e_1 e_1^T z_{i|i-1} - e_1(y_i - e_1^T z_{i|i-1}) / d_i \]

\[ = \lambda e_p e_p^T e_1 e_1^T z_{i|n}. \]

Since \( e_p^T e_1 = 0, D^1 = 0 \), that is the first derivative of \( z(t|n) \) is continuous at the data points.

This result can be extended to higher derivatives. The first occurrence of discontinuity will appear in differentiating the term involving \( e_p^T T(t_i, t) \) in (6.5.4). The successive derivatives of \( T(t_i, t) \) can be represented by

\[ \frac{d^j}{d t^j} T(t_i, t) = T(t_i, t) P_j(M) \]  

(6.5.9)

where the \( P_j \) satisfy the recurrence
Parallelling the argument in the preceding paragraph provides the result that the first \( k \) derivatives of \( z(t|n) \) are continuous at \( t_i \) provided

\[
\mathbf{e}_P^T \mathbf{P}_{j-1}(M)^T \mathbf{e}_1 = 0, \quad j=1,\ldots,k. \tag{6.5.11}
\]

Suppose that the vectors \( \mathbf{P}_{j-1}(M)^T \mathbf{e}_1, \ j=1,\ldots,p \), are linearly independent. At most the first \((p-1)\) derivatives can be continuous. The \( p \)'th derivative being continuous is equivalent to the \((p+1)\)'th vector being orthogonal to \( p \) linearly independent vectors which leads to a contradiction. The matrix \( \mathbf{P}_i(M) \) can be written as

\[
\mathbf{P}_i(M) = \begin{bmatrix}
0 & \ldots & 0 & | & I_{p-i} \\
& & & \begin{bmatrix}
\mathbf{a}_{i1}^T \\
\vdots \\
\mathbf{a}_{ii}^T
\end{bmatrix}
\end{bmatrix} \tag{6.2.12}
\]

where the \( a_{ij} \) are functions of the coefficients in \( L \). Clearly

\[
\mathbf{e}_P^T \mathbf{P}_i(M)^T \mathbf{e}_1 = 0, \quad i=0,\ldots,p-2.
\]

\[
= 1, \quad i=p-1
\]

which makes \( z(t|n) \) together with its first \((p-1)\) derivatives continuous at the data points.

QED.

**Lemma 6.5.2** The curve (6.4.12) is indeed an \( L_g \) Smoothing Spine.

**Proof.** For the curve \( \hat{f}(t) \) given by (6.4.12) to be an \( L_g \) Smoothing Spline we must show that

\[
L^+ L \hat{f}(t) = 0,
\]

where \( L^+ \) is the formal adjoint of \( L \). From equation (6.5.4) we have that the data points \( z(t|n) \) have the same smoothness
properties as the solutions to the homogeneous differential equation (6.2.6). Hence

\[ L\hat{f}(t) = e^{T(t_1, t)}S_{-1}(z_1|n - z_{i-1}) = e^{T(t_1, t)}d. \]

But it is immediately verified that

\[ L^*L\hat{f}(t) = L^*e^{T(t_1, t)}d = 0 \]

for every \( d \). This derivation is essentially that found in Kohn and Ansley (1983).

As a consequence of LEMMA 6.5.1 \( \hat{f}(t) \) has \((2p-2)\) continuous derivatives at the knot points. All that remains to do is to check the boundary conditions and give an explicit form for the jump in the \((2p-1)'\)th derivative of \( f(t) \). By LEMMA 6.5.1 \( e_1z(t|n) \)

together with its first \((2p-2)\) derivatives are continuous. Now

\[ d^{2p-1} - \lambda e_p e_T(M)e_1(y_1 - e_1z_1|n) = -\lambda e_p(y_1 - e_1z_1|n) \]

which gives

\[ \frac{d^{2p-1}}{dt}(z_1(t_1|n)) - \frac{d^{2p-1}}{dt}(z_1(t_1^-|n)) = -\lambda(y_1 - e_1z_1|n) \quad (6.5.13) \]

as the jump in the \((2p-1)th\) derivative in \( \hat{f}(t) \) given by (6.4.12).

We now check the boundary conditions. For \( t < t_1 \)

\[ \hat{f}(t) = e_1T(t, t_1)\hat{\alpha}. \]

Now,

\[ \frac{d}{dt}\hat{f}(t) = e_1^T \frac{d}{dt}T(t, t_1)\hat{\alpha} \]

but

\[ \frac{d}{dt}T(t, t_1) = MT(t, t_1), \quad T(t_1, t_1) = I \]

where \( T(t, t_1) \) is the fundamental matrix solution to

\[ \dot{z}(t) = Mz \]
which can be written as

$$Lz_1 = 0,$$

that is

$$Lz_1(t|n) = 0, \ t<t_1.$$  

Similarly

$$Lz_1(t|n) = 0, \ t>t_n.$$  

hence the natural boundary conditions are satisfied.

QED.

As noted earlier the polynomial Smoothing Spline is obtained by modelling the signal by the stochastic differential equation

$$\frac{d^p x}{dt^p} = \sigma \sqrt{\frac{d\xi}{dt}}.$$  

This is a subcase of (6.1.4) so all the results obtained in this chapter also apply for this subcase.
CHAPTER 7
OTHER STOCHASTIC APPROACHES TO LG SMOOTHING SPLINES

7.1 Introduction.

Reinsch's (1967) algorithm is the main algorithm used to construct smoothing splines. For completeness a stochastic derivation of Reinsch's algorithm will be given for the more general model considered here. The argument which uses an application of the Projection Theorem (THEOREM 2.1.1) is a generalization of that given by Kohn and Ansley (1987) and is based on Osborne and Prvan (1987a). An estimate of the condition number for the usual implementation of the Reinsch algorithm will be included. De Hoag and Hutchinson (1987) suggested an alternative implementation of the Reinsch algorithm and this was noted also in Osborne and Prvan (1987a). A brief outline of it will be given together with an estimate of the condition number for this implementation.

Other approaches to Lg Smoothing Splines differ from the approach outlined in the previous chapter by assuming a diffuse prior on

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

(7.1.1)

the vector of initial conditions on

\[ Lx = \sigma \sqrt{\frac{dw(t)}{dt}}. \]

That is, they set \( x(t_1) \sim N(0, \gamma^2 I) \) and let \( \gamma^2 \to \infty \). Infinite covariance corresponds to lack of information about \( x(t_1) \). These approaches are related to the work done by Harvey and Phillips (1979), Weinert, Byrd and Sidhu (1980), Ansley and Kohn (1985), (1986) and Kohn and Ansley (1987).

Harvey and Phillips (1979) work explicitly with the diffuse prior. Osborne and Prvan (1987a) show that \( S_{p|p} \) is bounded as \( \gamma^2 \to \infty \).
provided

\[ \text{dim}\{v_i, v_i = T(t_i, t_1)\}^T h, i=1,\ldots,k\} = k, k=1,\ldots,p. \quad (7.1.2) \]

This result makes it possible to initialize the Kalman Filter at \( t=t_p \) instead of \( t=t_1 \). The estimate \( x(t|n), t\geq t_p \), is computed in the usual way. A smoothing step is required to compute \( x(t|n) \) for \( t<t_p \). Weinert, Byrd and Sidhu (1980) also initialize the Kalman Filter at \( t=t_p \).

Letting \( f(t_i) = \mu_i(f) \) in DEFN 6.1.1, their assumption that the first \( p \) functionals \( \mu_1, \ldots, \mu_p \) are linearly independent over \( H \), the Hilbert Space of functions \( f(t) \) whose \( p \)'th derivatives are square integrable, permits them to initialize the Kalman Filter in this way. Here it will be shown that working with condition (7.1.2) the same algorithm as that of Weinert, Byrd and Sidhu is obtained. Ansley and Kohn (1985) find Wienert, Byrd and Sidhu's condition of the first \( p \) functionals \( \mu_1, \ldots, \mu_p \) being linearly independent too restrictive. They obtain a more complicated algorithm by removing this condition. Ansley and Kohn (1986) show that the output from their more complicated algorithm is equivalent to that of Weinert, Byrd and Sidhu when the linear independence conditions holds.

7.2 Stochastic derivation of the generalized Reinsch algorithm.

Reinsch's (1967) algebraic derivation of his algorithm was outlined in CHAPTER 1. Here an alternative derivation of his algorithm will be given based on the stochastic model developed in CHAPTER 6.

Our attention is restricted to the equations (6.2.5) and (6.2.13). The following argument is a generalization of that given in Kohn and Ansley (1987).
Substituting $t_1 = t_i$ into (6.2.5) gives

$$x(t) = T(t, t_i)x(t_i) + \int_{t_i}^{t} T(t, s)e_p g(s)ds.$$  \hspace{1cm} (7.2.1a)

The observation equation is

$$y_i = e_1^T x(t_i) + e_i$$  \hspace{1cm} (7.2.1b)

where $e_i \sim N(0, \sigma^2)$ and is independent of the integral term in (7.2.1a). The following is proposed as a basis for the $p$-dimensional vector space containing $x(t)^T$.

$$e_1, e_1^T(t_{i+1}, t_i), ..., e_1^T(t_{i+p-1}, t_i).$$

We can choose

$$\Lambda = [\Lambda_{11}^p, ..., \Lambda_{pp+1}^p]$$  \hspace{1cm} (7.2.2a)

such that

$$\sum_{j=1}^{p+1} \Lambda_{ij}^P e_1^T(t_{i+j-1}, t_i) = 0$$  \hspace{1cm} (7.2.2b)

since $e_1^T(t_{i+p}, t_i)$ must be a linear combination of the proposed basis, moreover we can choose the coefficients such that

$$\sum_{j=1}^{p+1} |\Lambda_{ij}^P| = 1.$$  \hspace{1cm} (7.2.2c)

The following error contrasts are defined,

$$z_i = \Lambda_{1i}^Py_i = \Lambda_{1i}^P f(t_i) + \Lambda_{1i}^P e_i, \, i=1, ..., n.$$  \hspace{1cm} (7.2.3)

Now

$$\Lambda_{1i}^P f(t_i) = \Lambda_{1i}^P e_1^T x(t_i)$$

$$= \sum_{j=1}^{p+1} \Lambda_{ij}^P e_1^T x(t_{i+j-1})$$

$$= \sum_{j=1}^{p+1} \left\{ \Lambda_{ij}^P e_1^T(t_{i+j-1}, t_i)x(t_i) + \Lambda_{ij}^P e_1^T \int_{t_i}^{t_{i+j-1}} T(t_{i+j-1}, s)e_p g(s)ds \right\}$$

(by (7.2.1))

$$= \sum_{j=1}^{p+1} \Lambda_{ij}^P e_1^T \int_{t_i}^{t_{i+j-1}} T(t_{i+j-1}, s)e_p g(s)ds$$

(by (7.2.2b))
which is independent of \( x(t_i) \). Thus the error contrasts are independent of the \( x(t_i) \) and hence of the particular model of the initial conditions. The best estimate \( \hat{f} \) of the signal given the data is obtained by subtracting the best estimate of \( \varepsilon \) given the data from \( y \). The independent information is given by the error contrasts so that, by the Projection Theorem,

\[
\hat{f} = y - \text{cov}(\varepsilon, z)\text{var}(z)^{-1}z \tag{7.2.5}
\]

and this must correspond to the values of the Lg Smoothing Spline calculated at the data points. A detailed verification is given by Kohn and Ansley (1987) for the cubic polynomial smoothing spline case. To facilitate evaluating the quantities in (7.2.5) equation (7.2.4) is written in matrix form. Now

\[
z = \sigma \Lambda q + Je \tag{7.2.6}
\]

where \( J: \mathbb{R}^n \to \mathbb{R}^{n-p} \) is an upper triangular matrix formed from the \( \Lambda_i^P \) shifted so that \( J_{ii} = \Lambda_i^{P+1} \) and \( q \in \mathbb{R}^n \) is defined by (7.2.4) and has components defined by

\[
q_i = \sum_{j=1}^{p+1} \Lambda_i^{P+1} e_{t_i+j-1}^T T(t_i+j-1, t_i)e_{t_i+p}^T \frac{dw}{ds} ds. \tag{7.2.7}
\]

Thus

\[
\text{cov}(\varepsilon, z) = E(\varepsilon z^T) = E(\varepsilon(\sigma \Lambda q + Je)^T) = E(\varepsilon e^T J^T) \quad \text{(because} \ \varepsilon \ \text{is independent of} \ q) = \sigma^2 J^T \tag{7.2.8}
\]

and

\[
\text{var}(z) = E(zz^T) = E((\sigma \Lambda q + Je)(\sigma \Lambda q + Je)^T) = \sigma^2 \Lambda E(q q^T) + \sigma^2 JJ^T. \tag{7.2.9}
\]
Define

\[ V = E(qq^T). \]

There can be a contribution to \( E(qq^T) \) only when the intervals of integration in (7.2.7) overlap and in fact \( V \) is a \( 2p-1 \) banded matrix. This can be seen from the following example.

**EXAMPLE 7.2.1** Let \( p=2 \) and \( M = e_2^e_2^T \) in the underlying stochastic differential equation (6.2.1) generating the state equation. Then the state transition matrix is

\[ T(t,t_1) = \begin{bmatrix} 1 & (t-t_1) \\ 0 & 1 \end{bmatrix}. \]

Without loss of generality assume the data points are evenly spaced. Set

\[ h = t_{i+1} - t_i, \]

then condition (7.2.2b) yields

\[ \Delta^2_{11} = 0.25, \quad \Delta^2_{12} = -0.5, \quad \Delta^2_{13} = 0.25, \quad i=1,\ldots,n. \]

From (7.2.7) we have that

\[
q_i = -0.5xe_1^T \int_{t_i}^{t_{i+1}} T(t_{i+1},s)e_2 \, dw(s) + 0.25xe_1^T \int_{t_i}^{t_{i+2}} T(t_{i+2},s)e_2 \, dw(s)
\]

\[
= -0.5x \int_{t_i}^{t_{i+1}} (t_{i+1} - s) \, dw(s) + 0.25x \int_{t_i}^{t_{i+2}} (t_{i+2} - s) \, dw(s).
\]

Looking at the following,

\[
E(q_j q_1) = E(( \int_{t_j}^{t_{j+1}} -0.5x(t_{j+1} - s) \, dw(s) + \int_{t_j}^{t_{j+2}} 0.25x(t_{j+2} - s) \, dw(s))x_{t_{i+1}} \int_{t_i}^{t_{i+2}} (t_{i+2} - s) \, dw(s))
\]

\[
( \int_{t_i}^{t_{i+1}} -0.5x(t_{i+1} - s) \, dw(s) + \int_{t_i}^{t_{i+2}} 0.25x(t_{i+2} - s) \, dw(s))
\]

\[
= (A'(A)^T). \]
\[
\begin{align*}
&= E\left( 0.25 \int_j^{t_{j+1}} (t_{j+1} - s) d\omega(s) \right) + E\left( 0.25 \int_j^{t_{i+1}} (t_{i+1} - s) d\omega(s) \right) + \\
&\quad E\left( -0.125 \int_j^{t_{i+2}} (t_{i+1} - s) d\omega(s) \right) + E\left( -0.125 \int_j^{t_{i+2}} (t_{i+2} - s) d\omega(s) \right) + \\
&\quad E\left( 0.0625 \int_j^{t_{i+2}} (t_{i+2} - s) d\omega(s) \right).
\end{align*}
\]

For this to be non zero the limits of integration in at least one of the four expectations have to overlap. There are three ways of achieving this, \(i=j\), \(i=j-1\) or \(i=j+1\). Thus \(V\) is a 3 banded matrix.

This example can be extended to show for general \(p\) that \(V\) is a \(2p-1\) banded matrix.

We will now proceed to estimate the condition number for the usual implementation of this algorithm. The smoothed estimates are

\[\hat{f} = y - \sigma^2 J^T (\sigma^2 V + \sigma^2 J J^T)^{-1} J y\]

since

\[z = J y\]

using (7.2.3). The quantity

\[\text{var}(z) = (\sigma^2 V + \sigma^2 J J^T)\]

in standard implementations of the Reinsch algorithm (for example de Boor (1978)) appears to be constructed and factorized explicitly. Osborne and Prvan (1987a) consider this to be the obvious quantity to use in attaching a condition number to this method. The condition number for any matrix \(A\) when the Euclidean norm is being used in the perturbation analysis is

\[\text{cond}(A) = \sigma_n(A)/\sigma_1(A)\]
where $\sigma_n(A)$ is the largest singular value of $A$ and $\sigma_1(A)$ is the smallest singular value of $A$. If $A$ is a symmetric matrix the Rayleigh quotient can be employed to obtain estimates of the extreme eigenvalues and eigenvectors of $A$. The Rayleigh quotient for $A$ is

$$RQ(x) = \frac{x^T A x}{x^T x}$$

(this result can be found in Golub and Van Loan (1983)).

To estimate the condition number of $\text{var}\{z\}$ the quantities $V$ and $JJ^T$ have to be investigated. It is worthwhile noting here that the $2p-1$ banded matrix $V$ is composed of elements of the form

$$e^T \Omega(t+\delta,t)e_1.$$

Some information about the eigenvalues of $\Omega(t+\delta,t)$, given by (6.2.11), in the limit as $\delta \to 0$ will prove useful. Expanding $T(t+\delta,t)$ in (6.2.11) using (6.5.9) and (6.5.10) gives

$$\Omega(t+\delta,t) = \sigma^2 \int \sum_{i,j} \frac{(s-(t+\delta))^{i+j}}{i!j!} p_i(M)e_p^T p_j(M)^T ds.$$

Since the successive powers of $\delta$ are incommensurable as $\delta \to 0$, using result (6.5.11) (where $k=p$) in the calculation of the Rayleigh quotient of $\Omega(t+\delta,t)$ gives the largest eigenvalue being $O(\lambda \delta)$ with its associated eigenvector tending to $e_p$ as $\delta \to 0$ and the smallest eigenvalue being $O(\lambda \delta^{2p-1})$ with its associated eigenvector tending to $e_1$ as $\delta \to 0$. Since $V$ is composed of elements of the form $e_1^T \Omega(t+\delta,t)e_1$ by the above discussion these elements are $O(\lambda \delta^{2p-1})$ since $e_1$ in the limit is the eigenvector associated with the smallest eigenvalue of $\Omega(t+\delta,t)$. The matrix $V$ can be written more conveniently as $V = \lambda \delta^{2p-1}\tilde{V}$ where $\tilde{V}$ is an appropriately scaled matrix. We now investigate $JJ^T$.

There is a strong analogy between

$$\sum_{j=1}^{p+1} \Delta_{ij}^P e_1^T (t_i+j-1, t_i) = 0$$
satisfying (7.2.2c) and the definition of divided differences (this can be found in any elementary Numerical Analysis text book, for example, Conte and de Boor (1980)). Let \( D: \mathbb{R}^P \to \mathbb{R}^P \) be nonsingular then \( \Psi(t)^T = e_1^T \phi(t, t_1)D \)
is a vector of linearly independent functions which are sent to zero by \( \Lambda_i^P, i=1, ..., n-p \). The values of \( \Psi(t) \) at \( t_1, ..., t_p \), span the null space of \( J_j^T \). As a consequence any sufficiently smooth function \( m(t) \) can be written as
\[
m(t) = \sum_{i=1}^{p} \xi_i \Psi(t_i) + R(t)
\]
where \( \xi_i \in \mathbb{R}, i=1, ..., p \), are constants and \( R(t) \) is a remainder term. Using this it can be shown that
\[
\sum_{j=1}^{p+1} \Lambda_1^{P,j}m(t_1+j-1) = O(\delta^P)
\]
(see Householder (1953)). Thus \( J_j^T \) has an eigenvalue which is \( O(\delta^{2p}) \).
This can be seen by using the Rayleigh quotient. The argument must allow for zero modes, and then use sample values of a smooth function to generate the estimate. It follows from the singular value decomposition of \( J \) that the singular values of \( J_j^T \) are the non zero singular values of \( JJ^T \). Consequently the Rayleigh quotient of \( \text{var}(z) \) can be made as small as \( \max(\lambda \delta^{2p-1}, \delta^{2p}) \). Worked examples show that the correct order is \( O(\lambda \delta^{2p-1}) \). This is expected because if \( V \) in equation (7.2.10) has positive elements its contributions to the Rayleigh quotient will be small for the highly fluctuating eigenvectors of \( \text{var}(z) \). These arguments give \( O(1) \) contributions from \( JJ^T \) when \( m(t) \) is varying rapidly and are not relevant in estimating small eigenvalues.
The above argument indicates that the condition number of the Reinsch algorithm in its usual implementation is given by
\[
\text{cond}(\text{var}(z)) = O(\delta^{-2p+1} \lambda^{-1}).
\]
De Hoog and Hutchinson (1987) suggested an alternative implementation of the Reinsch algorithm which was also suggested by Osborne and Prvan (1987a). It is based on interpreting $\text{cov}(e,z)\text{var}(z)^{-1}J$ as the top left hand corner of a projection matrix which can be computed using the orthogonal transformation determined by the QR factorization of

$$
\begin{bmatrix}
J \\
\sqrt{\lambda}V^{1/2}
\end{bmatrix}.
$$

Recollecting that

$$
\hat{f} = y - J^T(\Lambda V + JJ^T)^{-1}Jy
$$

where $J$ is the upper triangular matrix defined after equation (7.2.6), and $V$ is a $2p-1$ banded matrix given by (7.2.10) and (7.2.7). This can be rewritten as

$$
\hat{f} = y - J^Tc
$$

where

$$
(\Lambda V + JJ^T)c = Jy. 
$$

Equation (7.2.10b) is just the normal equation associated with the Linear Least Squares problem which can be solved by orthogonal transformations. Hutchinson and de Hoog (1985) showed how the $2p-1$ banded matrix $V$ could be decomposed as $LL^T$ by a Cholesky decomposition in $O(n)$ operations. Thus the square root of $(\Lambda V + JJ^T)$ can be found by reducing

$$
\begin{bmatrix}
\sqrt{\Lambda L^T} \\
J
\end{bmatrix}
$$

to

$$
\begin{bmatrix}
R \\
0
\end{bmatrix}
$$

by a series of orthogonal transformations. The quantity $R \in \mathbb{R}^{n \times n}$ is upper triangular. The same transformations are applied to $Jy$. Backward and forward substitution can then be used to determine the vector $c$ and then the estimate $\hat{f}$ can be formed. De Hoag and Hutchinson (1987) show how the trace of the influence matrix needed for GCV can be found in
0(n) operations.

A condition number can be readily attached to this method. The observation that equation (7.2.10b) is the normal equation associated with the Linear Least Squares problem suggests how to do this. The quantity (7.2.10a) is the residual in the least squares problem. THEOREM 1 in Appendix 3 gives the perturbation result for the residual as being

\[ \frac{\| \hat{f} - f \|_2}{\| f \|} \leq \epsilon (1 + 2\text{cond}(J^T)\min\{1,n-p\} + (\epsilon^2)). \]

From earlier work we have that

\[ \text{cond}(JJ^T) = o(\delta^{-2p}) \]

hence

\[ \text{cond}(J^T) = o(\delta^{-p}). \]

This supports the claim by de Hoog and Hutchinson (1987) that this approach is more numerically stable than the usual implementation of the Reinsch algorithm. They only supported this claim with numerical results.

7.3 Working explicitly with the diffuse prior.

Harvey and Phillips (1979) cast the regression model with autoregressive-moving average disturbances in a form suitable for the Kalman Filter. They calculated the generalized least squares estimator using the Kalman Filter as a computational tool which circumvented calculating and factorizing the covariance matrix of disturbances. It also provided an O(n) algorithm.

Their approach can be used for Lg Smoothing Splines. We work with the following state space formulation which was obtained in Section 6.2,
\[ x_i = T_i x_{i-1} + u_i \quad (7.3.1a) \]
\[ y_i = e_i^T x_i + \varepsilon_i. \quad i = 1, \ldots, n. \quad (7.3.1b) \]

We now assume that \( x_1 \) has a diffuse prior, that is setting \( x_1 \sim N(0, \tau^2 I) \) and letting \( \tau^2 \to \infty \). Harvey and Phillips work explicitly with \( S_1 | 0^2 = \tau^2 I \). They advocate \( \tau^2 = 10^6 \) as a suitable choice. We are no longer estimating \( x(t_1) = \alpha \) as in CHAPTER 6. This means that the likelihood attached to the state space formulation (7.3.1) when it is written in matrix form is different. This will be seen in the following paragraph.

Here (7.3.1) will be written in matrix form. Let
\[ y^T = [y_1, \ldots, y_n] \]
and
\[ v^T = [v_1, \ldots, v_n] \]
where \[ v_i = e_i^T x_i + \varepsilon_i \]
then (7.2.1) can be written as
\[ y = v. \quad (7.3.2) \]
The covariance of \( v \) is \( V \) and is of the form \( \sigma^2 A \) as in section 6.3.

Since \( \varepsilon_i \) and \( x_i \) have known distributions a likelihood can be attached to (7.3.2). The likelihood is
\[ l = l(\sigma^2, \lambda | y) = \left[ (2\pi \sigma^2)^{n/2} |\Lambda|^{1/2} \right]^{-1} e^{-\frac{1}{2} y^T \Lambda y}. \quad (7.3.3) \]
The log likelihood is
\[ \ell = \ell(\sigma^2, \lambda | y) = \text{const} - \frac{n}{2} \log(\sigma^2) - \frac{1}{2} \log(|\Lambda|) - (2\sigma^2)^{-1} Q \quad (7.3.4) \]
where
\[ Q = y^T \Lambda y. \]
Maximizing (7.3.4) over \( \sigma^2 \) gives
\[ \sigma^2 = (1/n)Q. \]
Substituting this into (7.3.4) after absorbing all the constants yields

\[
\mathcal{L}^*(\lambda | y) = \text{const} - (n/2) \log(Q) - (1/2) \log(|\Lambda|). \tag{7.3.5}
\]

Again let L be the Cholesky factor of \(\Lambda\). As in section 6.3 \(L^{-1}y\) and \(|\Lambda|\) can be obtained from the Kalman Filter with an equation appended.

The following Kalman Filter is used

\[
\begin{align*}
x_{i|j} &= T_{i} \bar{x}_{i-1|j-1} + \bar{e}_{i} \\
S_{i|j} &= T_{i} S_{i-1|j-1} T_{i}^T + \bar{\Omega}_{i} \\
d_{i} &= e_{i}^T S_{i|j-1} e_{i} + 1 \\
X_{i}^* &= (y_{i} - e_{i}^T \bar{x}_{i-1|j-1}) / \sqrt{d_{i}} \\
x_{i|i} &= x_{i|i-1} + S_{i|i-1} e_{i} (y_{i} - e_{i}^T \bar{x}_{i-1|j-1}) / d_{i} \\
S_{i|i} &= S_{i|i-1} - S_{i|i-1} e_{i} e_{i}^T S_{i|i-1} e_{i} d_{i}^{-1}.
\end{align*}
\]

The sequence \(\{X_{i}^*\}\) supplies the \(i\)'th element of \(L^{-1}e\). The recursion is initiated with \(x_{1|0} = 0\) and \(S_{1|0} = \gamma^2 I\), where \(\gamma^2\) is chosen to be large. It will be shown in the next section that after \(p\) steps of the Kalman Filter the influence of the diffuse prior has vanished. It is necessary to exclude the estimates obtained from the first \(p-1\) steps of the Kalman Filter from the likelihood calculation since they are sensitive to the arbitrary starting value for \(\gamma^2 I\). The following partial log likelihood is advocated for determining the optimal \(\lambda\) by any standard non-linear optimization technique

\[
\mathcal{L}^*(\lambda | y) = \text{const} - [(n-p)/2] \log(Q_p) - (1/2) \sum_{i=p}^{n} \log(d_{i})
\]

where

\[
Q_p = (n-p)^{-1} \sum_{i=p}^{n} (X_{i}^*)^2.
\]

Note that for each value of \(\lambda\) the Kalman Filter has to be run. Once the optimal \(\lambda\) has been determined the RTS Smoother ((3.3.4)–(3.3.6))
and Interpolating Smoother ((3.4.1)-(3.4.3)) are applied to obtain $x(t|n)$. The curve being fitted is

$$
\hat{f}(t) = \begin{cases} 
T_{t_1}^T(t, t_1) x_1 | n & , t \leq t_1 \\
T_{t_1}^T x(t | n) & , t_1 \leq t \leq t_n \\
T_{t_1}^T(t, t_n) x_n | n & , t > t_n
\end{cases} \quad (7.3.7)
$$

Similar arguments to those given in Section 6.5 will show that the curve is a natural $L_g$ Smoothing Spline with $(2p-2)$ continuous derivatives at the data points.

The major disadvantage with initializing the Kalman Filter with $S_1|0 = \gamma^2 I$ is that this may lead to numerical stability problems. Harvey and Phillips (1979) report that they have not experienced this difficulty. Numerical results comparing the estimates obtained by Wecker and Ansley's (1983) approach with those obtained from the method outlined in this section agree to 4 significant figures when $t \leq t_1$ and then improve dramatically even for $\gamma^2$ as small as 10,000.

7.4 An alternative derivation of Weinert, Byrd and Sidhu's $L_g$ Smoothing Spline.

The following state space formulation is being considered

$$
y_i = h^T x_i + \epsilon_i \quad (7.4.1a)
$$

$$
x_i = T_{i-1} x_{i-1} + u_i \quad (7.4.1b)
$$

where $h, x_i, u_i \in \mathbb{R}^p$, $\epsilon_i, y_i \in \mathbb{R}$ and $T_i \in \mathbb{R}^{p \times p}$. The term $\epsilon_i$ is normally distributed with zero mean and variance $\sigma^2$. The term $u_i$ is normally distributed with zero mean and covariance given by

$$
\Omega_i = \text{cov}(u_i) = \sigma^2 \Lambda \int_{t_{i-1}}^{t_i} T(t_i, s) b b^T T(t_i, s)^T ds \quad (7.4.2)
$$

where $b \in \mathbb{R}^p$. Note that this is a slight generalization of the state space formulation (7.1.3) previously used. In the next chapter the curves that this particular formulation produces will be examined.
Osborne and Prvan (1987a) proved the following theorem which Ansley and Kohn (1985) also established. This theorem is pivotal to an alternative derivation of Weinert, Byrd and Sidhu’s (1980) Lg Smoothing Spline which circumvents using Reproducing Kernel Hilbert Spaces.

**THEOREM 7.4.1** For the state space formulation (7.4.1) with \( x_1 \sim N(0, \gamma^2) \), \( S_p|p \) is bounded as \( \gamma^2 \to \infty \) provided

\[
\text{dim}(v_i, v_i = T(t_i, t_1)h, i=1, \ldots, k) = k, \ k=1, \ldots, p. \quad (7.4.2b)
\]

**PROOF.** We are interested in the Kalman Filter recursions for the covariances. For the state space formulation (7.4.1) they are

\[
S_k|k-1 = T_kS_{k-1}|k-1T_k^T + \Omega_k 
\]

\[
d_k = h^TS_k|k-1h + \sigma^2 
\]

\[
S_k|k = S_k|k-1 - (S_k|k-1hh^TS_k|k-1)(d_k^{-1}). 
\]

\[
k=1, \ldots, n. 
\]

From our assumption about \( x_1 \) we have that

\[
S_1|0 = \gamma^2I.
\]

The first iteration of (7.4.3) yields

\[
d_1 = \gamma^2h^Th + \sigma^2
\]

and

\[
S_1|1 = \gamma^2I - \gamma^2hh^T(\gamma^2h^Th + \sigma^2)^{-1}
\]

\[
= \gamma^2I - \gamma^2(\gamma^2h^Th + \sigma^2)hh^T + \gamma^2\sigma^2hh^T
\]

\[
h^T(\gamma^2h^Th + \sigma^2)
\]

\[
= \gamma^2(I - \frac{hh^T}{h^Th}) + \sigma^2(\gamma^2h^Th + \sigma^2)hh^T + 0(\gamma^{-2})
\]

\[
= \gamma^2(I - \frac{hh^T}{h^Th}) + \sigma^2\frac{hh^T}{h^Th}2 + 0(\gamma^{-2}). 
\]

\[
(7.4.4)
\]

It is easy to check that \( h^Th \) is an orthogonal projection onto \( h \).

To be consistent with the notation employed in Osborne and Prvan...
(1987a) this orthogonal projection is denoted by $Q_1$. Also let
\[ w_1 = v_1 = h \]  
(7.4.5a)
so that
\[ Q_1 = \frac{w_1^T w_1}{w_1^T w_1}. \]  
(7.4.5b)
Define
\[ R_1 = \sigma^2 \frac{T}{(w_1^T w_1)^2}. \]  
(7.4.5c)
then (7.4.4) can be written as
\[ S_1|_1 = T(t_1, t_1)[\gamma^2(I - Q_1) + R_1]T(t_1, t_1)^T + O(\gamma^{-2}). \]  
(7.4.6)
We now execute the second iteration. Equation (7.4.3a) becomes
\[ S_2|_1 = T(t_2, t_1)[\gamma^2(I - Q_1) + R_1 + T(t_2, t_1)^{-1} \Omega_2 T(t_2, t_1)^{-T}]T(t_2, t_1)^T + O(\gamma^{-2}) \]
which after setting
\[ E_2 = R_1 + T(t_2, t_1)^{-1} \Omega_2 T(t_2, t_1)^{-T} \]
becomes
\[ S_2|_1 = T(t_2, t_1)[\gamma^2(I - Q_1) + E_2]T(t_2, t_1)^T + O(\gamma^{-2}). \]
Equation (7.4.3b) for $k=2$ after substituting for $S_2|_1$ is up to $O(\gamma^{-2})$ terms
\[ d_2 = h^T T(t_2, t_1)[\gamma^2(I - Q_1) + E_2]T(t_2, t_1)^T h + \sigma^2. \]
Setting
\[ v_2 = T(t_2, t_1)h \]
this expression simplifies to
\[ d_2 = v_2^T[\gamma^2(I - Q_1) + E_2]v_2 + \sigma^2. \]
Equation (7.4.3c) up to $O(\gamma^{-2})$ terms becomes
\[ S_2|_2 = T(t_2, t_1)[\gamma^2(I - Q_1) + E_2 - \frac{(\gamma^2(I - Q_1)v_2 + E_2 v_2)(\gamma^2(I - Q_1)v_2 + E_2 v_2)^T}{v_2[\gamma^2(I - Q_1) + E_2]v_2 + \sigma^2}]T(t_2, t_1)^T. \]
Define
\[ w_2 = (I - Q_1) v_2 \]
then the above expression simplifies to
\[
S_{2|2} = T(t_2, t_1) [\gamma^2 (I - Q_1) + E_2 - \frac{(\gamma^2 w_2 + E_2 v_2)(\gamma^2 w_2 + E_2 v_2)^T}{\gamma^2 w_2 + v_2 E_2 v_2 + \sigma^2}] T(t_2, t_1)^T
\]
\[
= T(t_2, t_1) [\gamma^2 (I - Q_1) + E_2 - \frac{\gamma^2 w_2 v_2^T + \gamma^2 w_2 v_2^T + \gamma^2 w_2 v_2^T}{\gamma^2 w_2 + v_2 E_2 v_2 + \sigma^2}] T(t_2, t_1)^T + O(\gamma^{-2})
\]
\[
= T(t_2, t_1) [\gamma^2 (I - Q_1) + E_2 - \frac{\gamma^4 w_2 v_2^T}{\gamma^2 w_2 + v_2 E_2 v_2 + \sigma^2}] T(t_2, t_1)^T + O(\gamma^{-2})
\]
\[
= T(t_2, t_1) [\gamma^2 (I - Q_1) + E_2 - \frac{\gamma^2 v_2^T}{\gamma^2 w_2 + v_2 E_2 v_2 + \sigma^2}] T(t_2, t_1)^T + O(\gamma^{-2})
\]
\[
= T(t_2, t_1) [\gamma^2 (I - Q_1) + E_2 - \frac{\gamma^2 v_2^T}{\gamma^2 w_2 + v_2 E_2 v_2 + \sigma^2}] T(t_2, t_1)^T + O(\gamma^{-2})
\]
\[
= T(t_2, t_1) [\gamma^2 (I - Q_1) + E_2 + \sigma^2] T(t_2, t_1)^T + O(\gamma^{-2})
\]
\[
= T(t_2, t_1) [\gamma^2 (I - Q_1) + E_2 + \sigma^2] T(t_2, t_1)^T + O(\gamma^{-2})
\]

Defining
\[ Q_2 = Q_1 + \frac{w_2^T v_2}{w_2^T w_2} \]
this simplifies to the following expression (up to \(O(\gamma^{-2})\) terms)
\[
S_{2|2} = T(t_2, t_1) [\gamma^2 (I - Q_2) + R_2] T(t_2, t_1)^T
\]
where

\[
R_2 = (I - \frac{T}{w_2w_2^T})E_2(I - \frac{T}{w_2w_2^T})^T + \sigma^2 \frac{w_2w_2^T}{(w_2w_2^T)^2}.
\]

It is also worthwhile noting that

\[Q_2v_1 = v_1\]

and

\[Q_2v_2 = v_2.\]

In general we need to show that for \(k \leq p\), (up to first order terms)

\[S_{k|k-1} = T(t_k^T t_1)(\gamma^2(I - Q_{k-1}) + E_k)T(t_k^T t_1)^T \quad (7.4.7)\]

where \(E_k\) is 0(1) and \(Q_k\) is the orthogonal projection of rank \(k\) defined by

\[Q_k v_k = v_k = [v_1, \ldots, v_k]. \quad (7.4.8)\]

This result has been shown to be true when \(k = 1\) and \(k = 2\). The proof of (7.4.7) is by induction. Assume that the result holds for the first \(k\) steps, it is necessary to verify the result for the \((k+1)\)'st step based on (7.4.7). Let

\[w_k = (I - Q_{k-1})v_k \quad (7.4.8a)\]

so that

\[Q_k = Q_{k-1} + \frac{T}{w_k w_k^T}w_k w_k^T.\]

Substituting (7.4.7) into (7.4.3c) gives

\[S_{k|k} = T(t_k^T t_1) \times
\]

\[
(\gamma^2 w_k^T + E_k v_k)(\gamma^2 w_k^T + E_k v_k)^T
\]

\[
(\gamma^2(I - Q_{k-1}) + E_k - \sigma^2 + \gamma^2 \frac{T}{w_k^T w_k} + \frac{T}{v_k^T v_k})T(t_k^T t_1)^T
\]

which following lines of argument similar to those used in obtaining \(S_{2|2}\) simplifies to

\[S_{k|k} = T(t_k^T t_1)(\gamma^2(I - Q_k) + R_k)T(t_k^T t_1)^T \quad (7.4.9)\]
where, up to terms of $O(\gamma^{-2})$.

\[
R_k = (I - \frac{\mathbf{w}_k^T \mathbf{w}_k}{\sigma^2}) \mathbf{E}_k (I - \frac{\mathbf{w}_k^T \mathbf{w}_k}{\sigma^2})^T + \sigma^2 \frac{\mathbf{w}_k^T \mathbf{w}_k}{\sigma^2}^2.
\]  

(7.4.10)

The term $R_k$ is $O(1)$ and is positive definite if $E_k$ is.

Substituting (7.4.9) into (7.4.3a.) gives

\[
S_{k+1|k} = T(t_{k+1}, t_k) S_k T(t_{k+1}, t_k)^T + \Omega(t_{k+1}, t_k)
\]

\[
= T(t_{k+1}, t_1) \{ \gamma^2 (I - Q_k) + R_k + T(t_{k+1}, t_1)^{-1} \Omega_{k+1} T(t_{k+1}, t_1)^{-T} \} T(t_{k+1}, t_1)^T
\]

where

\[
E_{k+1} = R_k + T(t_{k+1}, t_1)^{-1} \Omega_{k+1} T(t_{k+1}, t_1)^{-T}
\]

\[
= R_k + \sigma^2 \int \frac{T(t_1, s) \mathbf{b} \mathbf{b}^T T(t_1, s)^T ds}{t_{k+1}}
\]

which shows that the result holds for the $(k+1)$st step so the induction has been established. Now after $p$ steps $Q_p = I$ so $I - Q_p = 0$ and

\[
S_p|p = T(t_p, t_1) R_p T(t_p, t_1)^T = O(1).
\]

This is the desired result.

QED.

It is necessary to obtain $x_p|p$ and $S_p|p$ to initialize the Kalman Filter at $t=t_p$. Once these quantities have been found the smoothed estimate $x_{i|p}$, $i \geq p$, and its associated covariance can be computed in the usual way using the RTS Smoother. Modifications are necessary in computing these quantities when $i < p$.

First of all the $\lim_{\gamma^2 \to \infty} x_k|k$ for $k \leq p$ will be found. For reference the Kalman Filter for the state space formulation (7.4.1) will be
given. It is

\[ x_{i|i-1} = T_i x_{i-1|i-1} \]  
\[ S_{i|i-1} = T_i S_{i-1|i-1} T_i^T + \Omega_i \]  
\[ d_i = h^T S_{i|i-1} h + \sigma^2 \]  
\[ x_{1|i} = x_{1|i-1} + S_{1|i-1} h(y_i - h^T x_{1|i-1})/d_i \]  
\[ S_{1|i} = S_{1|i-1} - S_{1|i-1} h h^T S_{1|i-1} d_i^{-1} \]  
\[ \text{for } i=1, \ldots, n. \]

From THEOREM 7.4.1 \( x_k|k \) can be expanded as follows

\[ x_k|k = x_k^0 + \gamma^{-1} x_k|k + \ldots \]  

Substituting (7.4.7) into (7.4.11d) using (7.4.11a), (7.4.2b) and (7.4.8a) gives

\[ x_{1|i} = T_i x_{i-1|i-1} + \gamma^2 T(t_1, t_1) w_1 (\gamma w_1^T v_i^T v_i + \sigma^2) + O(\gamma^{-2}). \]  

Setting

\[ x_{1|i} = T(t_1, t_1) \xi_i \]

simplifies (7.4.12) to

\[ \xi_i = \xi_{i-1} + \gamma^2 w_i (\gamma w_i^T v_i^T v_i + \sigma^2) + O(\gamma^{-2}) \]

\[ = \xi_{i-1} + w_i (y_i - v_i^T \xi_{i-1})/(w_i^T w_i) + O(\gamma^{-2}). \]

Letting \( \gamma \to \infty \) produces

\[ \xi_i = \xi_{i-1} + w_i (y_i - v_i^T \xi_{i-1})/(w_i^T w_i) \]

\[ = (I - \frac{w_i v_i^T}{w_i^T w_i}) \xi_{i-1} + w_i y_i/(w_i^T w_i). \]

Comparing (7.4.14) with (7.4.10) it is seen that \( \text{cov}(\xi_i) \) equals \( R_i \).

Equation (7.4.14) can also be written as
In particular

\[ y_p = V_p^T \xi_p \]

but since \( V \in \mathbb{R}^{p \times p} \) and has full rank we have that

\[ \xi_p = V_p^{-T} y_p. \]

Substituting this result into (7.4.13) gives

\[
\begin{align*}
    x_p^0 | p &= T(t_p, t_1) V_p^{-T} y_p \\
    &= (V_p^T t_p, t_1)^{-1} y_p \\
    &= \begin{bmatrix}
        h^T(t_1, t_p) T(t_p, t_1)^{-1} y_p \\
        h^T(t_2, t_p) T(t_p, t_1)^{-1} y_p \\
        \vdots \\
        h^T(t_p, t_1) T(t_p, t_1)^{-1} y_p
    \end{bmatrix}
\end{align*}
\]

which is the estimate used by Weinert, Byrd and Sidhu (1980) to initialize their recursion.

From THEOREM 7.4.1 we have that

\[
S_p | p = T(t_p, t_1) R_p T(t_p, t_1)^T.
\]

Now \( R_p \) is the covariance of \( \xi_p \) which is also given by

\[
\text{Cov}(\xi_p) = \text{Cov}(V_p^{-T} y_p) = V_p^{-T} \text{Cov}(y_p) V_p^{-1}.
\]
Substituting this into (7.4.18) gives
\[ S_p|p = (V_p^T(t_p,t_1)^{-1} - I) - l \text{ Cov}(y_p)(T(t_p,t_1)^{-T} V_p)^{-1} \]
\[ = \bar{M}^{-1}\text{ Cov}(y_p)\bar{M}^{-T}. \]

From (7.4.1a)
\[ y_i = h^T x_i + \epsilon_i \]
but \( x_i \) using result (6.2.5) can be written as
\[ x(t_i) = T(t_i, t_p) x(t_p) + \sigma^2 \lambda \int T(t_i, s)b \frac{dw}{ds} ds. \]

Setting
\[ z_i = \sigma^2 \lambda \int T(t_i, s)b \frac{dw}{ds} ds \]
the observation equation can be rewritten as
\[ y_i = h^T_i T(t_i, t_p) x(t_p) + h^T_i z_i + \epsilon_i. \]

Let \( \Lambda_{ij} \) denote the \((i,j)\)th element of \( \text{ Cov}(y_p) \). After some manipulation, for \( j \geq i \),
\[ \Lambda_{ij} = \delta_{ij} \sigma^2 + \sigma^2 \lambda h^T_i T(t_i, t_j) \int T(t_i, s)bb^T(t_i, s)^T ds h. \quad (7.4.19) \]

Thus
\[ \Lambda = \sigma^2 I + H \]
where \( H \) is defined by (7.4.19). Using this result gives
\[ S_p|p = \bar{M}^{-1}(\sigma^2 I + H)\bar{M}^{-T}. \]

When \( b = e_p \) and \( \sigma^2 I = Q_p \) this is equivalent to the covariance used by Weinert, Byrd and Sidhu (1980) to initialize their recursion.

The following lemma provides the recursion for the first \((p-1)\) smoothed estimates. These are required to obtain the smoothing spline.
LEMMA 7.4.1 Letting $\gamma^2 \to \infty$, the following equations provide the smoothed estimate $x_i|n$ and its associated covariance, i.e., for the state space formulation (7.4.1) with $x_1 \sim N(0, \gamma^2 I)$. They are

$$x_{i-1}|n = T(t_{i-1}, t_1)\xi_{i-1} + T(t_{i-1}, t_1)(I+\Delta_i X_i)T(t_{i-1}, t_1)^T \times (x_i|n - T(t_i, t_1)\xi_{i-1})$$

(7.4.20a)

and

$$S_{i-1}|n = T(t_{i-1}, t_1)[(I+\Delta_i X_i)S_i|n(I+\Delta_i X_i)^T - \Delta_i - \Delta_i X_i^T \Delta_i]T(t_{i-1}, t_1)^T$$

(7.4.20b)

where

$$\Delta_i = R_{i-1} - E_i$$

(7.4.20c)

the quantities $E_i$ and $R_i$ given by (7.4.10a) and (7.4.10b), and $X_i$ satisfies the following equations

$$X_i = X_i Q_{i-1} X_i$$

$$Q_{i-1} E_i X_i Q_{i-1} = Q_{i-1}$$

This specifies $X_i$ uniquely.

PROOF. Referring to the RTS Smoother given by (5.2.7) we have that

$$x_{i-1}|n = x_{i-1}|i-1 + A_{i-1}(x_i|n - T_i x_{i-1}|i-1)$$

(7.4.21a)

and

$$S_{i-1}|n = S_{i-1}|i-1 + A_{i-1}(S_i|n - S_i|i-1)A_{i-1}^T$$

(7.4.21b)

where

$$A_{i-1} = S_{i-1}|i-1 T_i S_{i} T_i^T$$

(7.4.21c)

The quantities $S_i|i-1$ and $S_{i-1}|i-1$ from THEOREM 7.4.1 can be written as

$$S_i|i-1 = T(t_i, t_1)\{\gamma^2(I - Q_{i-1}) + E_i + \gamma^{-2}E_i + \cdots\}T(t_i^T, t_1)$$

(7.4.22a)
and
\[ S_{i-1|i-1} = T(t_{i-1}, t_1)\{\gamma^2(I - Q_{i-1}) + R_{i-1} + \gamma^{-2}R_{i-11} + \cdots \} \times T(t_{i-1}, t_1)^T \quad (7.4.22b) \]

Substituting (7.4.22a) and (7.4.22b) into the expression for \( A_{i-1} \) given by (7.4.21c) yields after simplifying
\[
A_{i-1} = T(t_{i-1}, t_1)\{\gamma^2(I - Q_{i-1}) + R_{i-1} + \gamma^{-2}R_{i-11} + \cdots \} \times \\
{\gamma^2(I - Q_{i-1}) + E_i + \gamma^{-2}E_{i1} + \cdots}^{-1}T(t_{i}, t_1)^{-1} \\
= T(t_{i-1}, t_1)\{\gamma^2(I - Q_{i-1}) + R_{i-1} + \gamma^{-2}R_{i-11} + \cdots \} \times \\
\{X_i + \gamma^{-2}X_{i1} + \gamma^{-4}X_{i2} + \cdots \}T(t_{i}, t_1)^{-1} \\
\]
where
\[
\{\gamma^2(I - Q_{i-1}) + E_i + \gamma^{-2}E_{i1} + \cdots\}\{X_i + \gamma^{-2}X_{i1} + \gamma^{-4}X_{i2} + \cdots\} = I \\
\]
and
\[
\{X_i + \gamma^{-2}X_{i1} + \gamma^{-4}X_{i2} + \cdots\}\{\gamma^2(I - Q_{i-1}) + E_i + \gamma^{-2}E_{i1} + \cdots\} = I. \\
\]
The above two relationships give
\[
(I - Q_{i-1})X_i = 0 \Rightarrow X_i = Q_{i-1}X_i \quad (7.4.23a) \\
X_i(I - Q_{i-1}) = 0 \Rightarrow X_i = X_iQ_{i-1} \quad (7.4.23b) \\
E_iX_i + (I - Q_{i-1})X_{i1} = I \Rightarrow (I - Q_{i-1})X_{i1} = I - E_iX_i \quad (7.4.24a) \\
X_iE_i + X_{i1}(I - Q_{i-1}) = I \Rightarrow X_{i1}(I - Q_{i-1}) = I - X_iE_i. \quad (7.4.24b) \\
\]
Postmultiplying (7.4.23b) and using (7.4.23a) gives
\[
X_i = Q_{i-1}X_iQ_{i-1}. \\
\]
Premultiplying (7.4.24a) by \( Q_{i-1} \) yields
\[
Q_{i-1}E_iX_i = Q_{i-1}. \\
\]
Using (7.4.23b) this can be written as
\[
Q_{i-1}E_iX_iQ_{i-1} = Q_{i-1}. \\
\]
Consider
\[
\{\gamma^2(I - Q_{i-1}) + R_{i-1} + \gamma^{-2}R_{i-11} + \cdots\}\{X_i + \gamma^{-2}X_{i1} + \gamma^{-4}X_{i2} + \cdots\} \]
\[= \gamma^2(I - Q_{i-1})X_i + (I - Q_{i-1})X_{i1} + \gamma^{-2}(I - Q_{i-1})X_{i2} + R_{i-1}X_i + \]
\[
\gamma^{-2}R_{i-11} + \gamma^{-2}R_{i-11}X_i + \cdots \\
\]
\[ \begin{align*}
= I - E_1 X_1 + R_{i-1} X_1 + \gamma^{-2}(R_{i-1} X_1 + R_{i-1} X_1 - E_1 X_1 - E_1 X_1) + \cdots \\
= I + (R_{i-1} - E_1)X_1 + \gamma^{-2}[(R_{i-1} - E_1)X_1 + (R_{i-1} - E_1)X_1] + \cdots \\
= I + \Delta_{i} X_1 + \gamma^{-2}(\Delta_{1} X_{i1} + \Delta_{1} X_{i1}) + \cdots \\
\end{align*} \]

where

\[ \Delta_i = R_{i-1} - E_1 \]

and

\[ \Delta_{11} = (R_{i-1} - E_{i1}) \]

thus

\[ A_{i-1} = T(t_{i-1}, t_1)(I + \Delta_{i} X_1 + \gamma^{-2}(\Delta_{1} X_{i1} + \Delta_{1} X_{i1}) + \cdots)T(t_{i-1}, t_1)^{-1}. \]

(7.4.25)

Recall that by construction

\[ x_{i1}^0 = T(t_{i1}, t_1) \xi_i \]

where \( \xi_i \) is furnished by (7.4.15). Substituting (7.4.25) into (7.4.21a) and letting \( \gamma^2 \to \infty \) provides the recursion

\[ x_{i-1 | n} = T(t_{i-1}, t_1) \xi_{i-1} + T(t_{i-1}, t_1)(I + \Delta_{i} X_1)T(t_{i1}, t_1)^{-1} \]

\[ \times \{x_{i1 | n} - T(t_{i1}, t_1) \xi_{i-1}\}. \]

In the expression for the smoothed covariance given by (7.4.21b) the term of interest is

\[ S_{i-1 | i-1} - A_{i-1} S_{i-1 | i-1} A_{i1}^T \]

since \( S_p | n \) is bounded. Referring to (7.4.7), (7.4.9) and (7.4.25) the above expression can be rewritten as

\[ T(t_{i-1}, t_1)[\gamma^2(I - Q_{i-1}) + R_{i-1} + \gamma^{-2}R_{i-1} + \cdots] \]

\[ - \{I + \Delta_{i} X_1 + \gamma^{-2}(\Delta_{1} X_{i1} + \Delta_{1} X_{i1}) + \cdots\}(\gamma^2(I - Q_{i-1}) + E_{i1} \]

\[ + \gamma^{-2}E_{i1} + \cdots \}

\[ \times \{I + \Delta_{i} X_1 + \gamma^{-2}(\Delta_{1} X_{i1} + \Delta_{1} X_{i1}) + \cdots\}^T]T(t_{i-1}, t_1)^T. \]
Consider
\[\gamma^2(I - Q_{-1}) + R_{-1} + O(\gamma^{-2}) - \{I + \Delta_{i}X_{i+} \gamma^{-2}(\Delta_{i}X_{i} + \Delta_{i}X_{1}) + \cdots \}\{\gamma^2(I - Q_{-1}) + E_{11} + \gamma^{-2}E_{11} + \cdots \}\]
\[\times \{I + \Delta_{i}X_{i+} \gamma^{-2}(\Delta_{i}X_{i} + \Delta_{i}X_{1}) + \cdots \}^T\]
\[= \gamma^2(I - Q_{-1}) + R_{-1} + O(\gamma^{-2})
- \{\gamma^2(I - Q_{-1}) + E_{11} + \gamma^{-2}(\Delta_{i}X_{i} + \Delta_{i}X_{1}) + \cdots \}\{\gamma^2(I - Q_{-1}) + \gamma^{-2}E_{11} + \cdots \}\]^T
\[\times \{I + \Delta_{i}X_{i+} \gamma^{-2}(\Delta_{i}X_{i} + \Delta_{i}X_{1}) + \cdots \}^T\]

which using the results (7.4.23b) and (7.4.24b) simplifies to
\[\gamma^2(I - Q_{-1}) + R_{-1} + O(\gamma^{-2}) - \{\gamma^2(I - Q_{-1}) + E_{11} + \gamma^{-2}(\Delta_{i}X_{i} + \Delta_{i}X_{1}) + \cdots \}\{\gamma^2(I - Q_{-1}) + \gamma^{-2}E_{11} + \cdots \}\]^T
\[= \gamma^2(I - Q_{-1}) + R_{-1} - \gamma^2(I - Q_{-1}) - \gamma^2(I - Q_{-1})X_{i\Delta_{i}}^T
- \gamma^2(I - Q_{-1})(\Delta_{i}X_{i} + \Delta_{i}X_{1})^T
- E_{11} - \gamma^2E_{11}^T - \gamma^2E_{11}^T - \Delta_{i} - \Delta_{i}X_{i}^T\Delta_{i}^T
+ O(\gamma^{-2})\]
\[= (R_{i-1} - E_{1}) - (I - Q_{-1})X_{i\Delta_{i}}^T - E_{11}^T - \Delta_{i}X_{i}^T\Delta_{i}^T + O(\gamma^{-2})\]
\[= \Delta_{i} - (I - E_{11}X_{i}^T)\Delta_{i}^T - E_{11}^T - \Delta_{i}X_{i}^T\Delta_{i}^T + O(\gamma^{-2})\]
\[= -\Delta_{i} - \Delta_{i}X_{i}^T\Delta_{i}^T + O(\gamma^{-2})\]

thus
\[S_{i-1|n-1} - \Delta_{i}^T S_{i|i-1} - \Delta_{i}X_{i}^T\Delta_{i}^T + \cdots \]T
\[= T(t_{i-1}, t_1)[-\Delta_{i} - \Delta_{i}X_{i}^T\Delta_{i}^T + \cdots \]T
\[= T(t_{i-1}, t_1)[(I + \Delta_{i}X_{i})S_{i|i} - \Delta_{i}X_{i}^T\Delta_{i}^T]T(t_{i-1}, t_1).\] (7.4.26)

Substituting this result into (7.4.28) gives the recursion
\[S_{i-1|n} = \Delta_{i}X_{i}^T\Delta_{i}^T T(t_{i-1}, t_1)^T.\]

QED.

The following lemma provides the recursion for x(t|n) and its associated covariance S(t|n) for t≤p.
LEMMA 7.4.2 Letting $\tau \to 0$, the following equations provide the smoothed estimate $x(t|n)$ and its associated covariance, $t_{i-1} \leq t \leq t_i$, $i \leq p$, for the state space formulation (7.4.1) with $x_1 \sim N(0, \tau^2 I)$. The recursion is

$$x(t|n) = T(t, t_1)\breve{x}_{i-1} + (I + \breve{A}_i X_i)(x_i|n - T(t_i, t_1)\breve{x}_{i-1})$$

and

$$S(t|n) = T(t, t_1)[((I + \breve{A}_i X_i)S_i|n(I + \breve{A}_i X_i)^T - \bar{f}(t_i, t) - E_i X_i^T \breve{A}_i - \breve{A}_i X_i \breve{A}_i^T]T(t, t_1)^T$$

where

$$\breve{A}_i = R_{i-1} + \bar{f}(t_i, t_1) - E_i$$

and $X_i$ satisfies

$$X_i = Q_{i-1} X_i Q_{i-1}$$

$$Q_{i-1} E_i X_i Q_{i-1} = Q_{i-1}.$$ 

The quantities $E_i$ and $R_i$ are given by (7.4.10a) and (7.4.10b).

The proof of LEMMA 7.4.2 is in Appendix 2.

The curve being fitted is

$$\hat{f}(t) = \begin{bmatrix} h^T T(t, t_1) x_{1}|n & t \leq t_1 \\ h^T x(t|n) & t_1 \leq t \leq t_n \\ h^T T(t, t_n) x_{n}|n & t_n < t \end{bmatrix}$$

Paralleling the argument given in the proof of LEMMA 6.5.1 for the state space formulation (7.4.1) the first $k$ derivatives of $x(t|n)$ are continuous at the data points provided

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

where the $P_{j-1}$ satisfy the recurrence
The state space formulation (7.4.1) will only provide an Lg Smoothing Spline if \( h = e_1 \) and \( b = e_p \) which is borne out in LEMMA 6.5.1.

If \( h = e_1 \) and \( b = e_p \) then there exists an alternative to using LEMMA 7.4.2 for calculating \( x(t|n), t < t_p \). It is a consequence of the variational derivation that Lg Smoothing Splines solve

\[
L^+ L x(t) = 0
\]

(for example see Prenter (1969)) where \( x(t) \) is the first component of the vector \( x(t) \) and \( L^+ \) is the formal adjoint of \( L \). The discontinuity at the \((2p-1)\)st derivative at the data points is used to successively generate terminal conditions for \( x(t|n) \) at \( t_1 = p, \ldots, 2 \), and then to integrate (7.4.32) back to \( t_1-1 \). Here the smoothing formula for \( x(t|n), t_{i-1} \leq t \leq t_i \), provides an explicit solution to the boundary value problem for (7.4.32) with boundary conditions at \( x(t_{i-1}|n) \) and \( x(t_i|n) \).
8.1 Introduction.

In CHAPTER 6 it was shown that Lg Smoothing Splines are related to the following stochastic differential equation

\[ \dot{x}(t) = Mx(t) + \sigma \sqrt{\lambda} \, \frac{dw}{dt} \, e_p, \]  

(8.1.1)

where \( x \) is a constant vector to be determined, \( x(t) \in \mathbb{R}^p \), \( M \in \mathbb{R}^{p \times p} \) and \( \omega(t) \) is a Wiener process with unit dispersion. The data points \( y_1, \ldots, y_n \) are assumed to be expressible as

\[ y_i = e_1^T x_i + e_i. \]  

(8.1.2)

where the \( e_i \) are independent of the Wiener process \( \omega(t) \) and are normally distributed with zero mean and variance \( \sigma^2 \). The solution to the stochastic differential equation (8.1.1) can be written in recursive form

\[ x_{i+1} = T_{i+1} x_i + u_{i+1} \]  

(8.1.3)

where \( T_{i+1} = T(t_{i+1}, t_i) \) is the fundamental matrix solution satisfying

\[ T(t, \xi) = M T(t, \xi), \quad T(t_i, t_i) = I, \]

and \( u_{i+1} \) is given by

\[ u_{i+1} = \sigma \sqrt{\lambda} \int_{t_i}^{t_{i+1}} T(t_{i+1}, s) e_p \frac{dw(s)}{dt} \, ds. \]

which has associated covariance

\[ \Omega_{i+1} = \sigma^2 \int_{t_i}^{t_{i+1}} T(t_{i+1}, s) e_p e_p^T T(t_{i+1}, s)^T ds. \]

The quantity \( u_{i+1} \) also has zero mean and is normally distributed. The state space formulation (8.1.2) and (8.1.3) facilitates the use of recursive algorithms to solve for \( x(t|n) \). This consists of a forward pass of the Kalman Filter followed by a backward pass of the RTS.
Smoother initialized with the final estimates of the forward filter before using the Interpolation Smoother to evaluate $x(t|n)$ and its associated error covariance $S(t|n)$ for $t_{i-1} \leq t \leq t_i$ which is required in calculating formal confidence intervals for the curve being fitted. A similar result to that contained in Lemma 6.5.1 can be obtained for the state space formulation (8.1.2) and (8.1.3) considered here.

In Section 7.4 a more general form was assumed for the observations $y_i$ and vector $u_i$. It is

$$y_i = h^T x_i + e_i$$  \hspace{1cm} (8.1.4a)

$$x_i = T_i x_{i-1} + u_i$$  \hspace{1cm} (8.1.4b)

where

$$u_i = \sigma \sqrt{\int_{t_{i-1}}^{t_i} T(t_i,s)b \frac{d\omega(s)}{ds} ds},$$

$b \in \mathbb{R}^p$ and the remaining quantities are defined to be the same as in Section 7.4. This is the state space formulation considered by Osborne and Prvan (1987a). There it was shown that the first $k$ derivatives of $x(t|n)$ are continuous at the knot points provided

$$b^T P_{j-1}(M)^T h = 0, \ j=1, \ldots, k,$$  \hspace{1cm} (8.1.5a)

where the $P_{j-1}$ satisfy the recurrence

$$P_0 = I, \ P_j = \frac{d}{dt}(P_{j-1}) - MP_{j-1}, \ j=1,2,\ldots$$  \hspace{1cm} (8.1.5b)

The Smoothing Spline being fitted coincides with the usual definition of an $L_g$ Smoothing Spline if $h = e_1$ and $b = e_p$. The state transition equation (8.1.4a) arises from the stochastic differential equation

$$\dot{x}(t) = Mx(t) + \sigma \sqrt{\frac{d\omega}{dt}} b.$$  \hspace{1cm} (8.1.6)

This was our first attempt at Generalized Smoothing Splines.
REMARK 8.1.1 If $b$ has a zero element in its last position the covariance $\Omega_j$ is no longer positive definite. Information implementations of the Kalman Filter cannot be used since the square root of the inverse of $\Omega_j$ is required.

Condition (8.1.5) provides a mechanism for generating curves which possess a variety of smoothness possibilities at the knot points. Some examples are given below.

EXAMPLE 8.1.1. If $p = 3$, $b = e_p$, $M = e_1 e_2^T + e_2 e_3^T$ and $h = e_1 + e_2$ then $h^T x(t|n)$ is a quintic polynomial with its first derivative continuous at the data points.

EXAMPLE 8.1.2. If $p = 3$, $b = e_p$, $M = e_1 e_2^T + e_2 e_3^T$ and $h = e_1 + e_2 + e_3$ then $h^T x(t|n)$ is a quintic polynomial with no continuous derivatives at the data points.

EXAMPLE 8.1.3. If $p = 3$, $b = e_2$, $M = e_1 e_2^T + e_2 e_3^T$ and $h = e_1$ then $h^T x(t|n)$ is a quintic polynomial with two continuous derivatives at the data points.

The stochastic differential equation (8.1.6) can be made more general. This generality permits a wider range of smoothness possibilities for $h^T x(t|n)$ at the data points. The above represent our first attempt at Generalized Smoothing Splines and correspond to a subcase of Generalized Smoothing Splines which will be developed in the next section.
8.2 Generalized Smoothing Splines.

The data \( y_i \) is still assumed to be expressible as

\[
y_i = h^T x_i + \epsilon_i, \quad i = 1, \ldots, n, \tag{8.2.1}
\]

where \( h, x_i \in \mathbb{R}^p \), \( y_i, \epsilon_i \in \mathbb{R} \), and the noise \( \epsilon_i \) is normally distributed with zero mean and variance \( \sigma^2 \). We want to generalize the stochastic differential equation (8.1.6) which furnishes the state equation, and it is the stochastic forcing term \( \frac{d\omega}{dt} \) \( \alpha \Lambda \) which begs to be generalized. Suppose that \( x(t) \) now satisfies the stochastic differential equation

\[
\dot{x}(t) = M(t)x(t) + \sigma \Lambda \frac{d\omega}{dt} \tag{8.2.2}
\]

where \( M(t) \) is assumed to possess enough continuous derivatives for our purposes and \( \omega(t) \) is a Wiener process, independent of the \( \epsilon_i \) in (8.2.1), satisfying, for positive semi definite \( V: \mathbb{R}^p \rightarrow \mathbb{R}^p \) given,

\[
E((\omega_i(t+\delta) - \omega_i(t))(\omega_j(t+\delta) - \omega_j(t))) = V_{ij}\delta. \tag{8.2.3}
\]

Assuming that the vector of initial conditions \( x(t_1) \) is constant then the solution to equation (8.2.2) is

\[
x(t) = T(t,t_1)x(t_1) + \sigma \Lambda \int_{t_1}^{t} T(t,s) \frac{d\omega(s)}{ds} \, ds \tag{8.2.4}
\]

where \( T(t,t_1) \) is the fundamental matrix solution for

\[
\dot{x}(t) = M(t)x(t). \tag{8.2.5}
\]

That is, \( T(t,t_1) \) satisfies the initial value problem

\[
\frac{d}{dt} T(t,t_1) = MT(t,t_1), \quad T(t_1,t_1) = I. \tag{8.2.6}
\]

Equation (8.2.4) can be written in recursive form

\[
x_{i+1} = T_{i+1}x_i + u_{i+1} \tag{8.2.7}
\]

where

\[
u_{i+1} = \sigma \Lambda \int_{t_i}^{t_{i+1}} T(t_{i+1},s) \frac{d\omega(s)}{ds} \, ds. \tag{8.2.8}
\]

By definition \( u_{i+1} \) is independent of \( x(t) \) for \( t \leq t_i \) and is normally
distributed with zero mean and covariance given by

\[ \Omega_{i+1} = \sigma^2 \lambda \int_{t_i}^{t_{i+1}} T(t_{i+1}, s) V T(t_{i+1}, s)^T ds. \]  \hspace{1cm} (8.2.9)

The state space formulation (8.2.1) and (8.2.7) is that considered by Osborne and Prvan (1987b).

The following lemma provides the order of magnitude of the smallest and largest eigenvalues of \( \Omega_{i+1} \) when the data points are equally spaced \( \delta \) apart and \( \delta \to 0 \).

**Lemma 8.2.1** 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 \( v \) such that

\[ VP_j(M)v = 0, \quad j=0,1,\ldots,k-1, \]  \hspace{1cm} (8.2.10)

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

**Proof.** The key step is to make a Taylor series expansion of \( \Omega(t+\delta, t) \). This is obtained by using

\[ \frac{d}{ds}^j (T(t,s)) = T(t,s)P_j(M) \]  \hspace{1cm} (8.2.11)

where \( P_j(M) \) is defined by

\[ P_0(M) = I, \quad P_1(M) = \frac{d}{dt}(P_{i-1}) - MP_{i-1}, \quad i=1,2,\ldots \]  \hspace{1cm} (8.2.12)

The expansion is

\[ \Omega(t+\delta, t) = \sigma^2 \lambda \int_{t_i}^{t_{i+1}} (s - (t+\delta))^i j P_j(M) V P_j(M)^T ds. \]  \hspace{1cm} (8.2.13)

An estimate of the order of magnitude of the largest eigenvalue of \( \Omega(t+\delta, t) \) is obtained by maximizing the Rayleigh quotient

\[ \frac{r^T \Omega(t+\delta) r}{r^T r}. \]

The \( O(\delta) \) term must be picked up. This is achieved by setting
i=j=0 in (8.2.13) and using $P_0(M) = I$. From (8.2.10) its associated eigenvector has to be asymptotic to the orthogonal complement of the kernel of $V$ in (8.2.10). For reference we denote this eigenvector by $\gamma$. The smallest eigenvalue is found by minimizing the Rayleigh quotient and noting that distinct powers of $\delta$ are incommensurate as $\delta \to 0$. Forming the Rayleigh quotient using the vector $v$ from (8.2.10) the first non-zero contribution occurs when $i=j=k$ giving $0(\lambda \delta^{2k+1})$ as the required size of the smallest eigenvalue.

QED.

How to proceed from the state space formulation (8.2.1) and (8.2.7) depends upon the assumptions placed on $x(t_1)$, the vector of initial conditions on (8.2.2). If $x(t_1)$ is a constant vector to be determined then the approach outlined in CHAPTER 6 is advocated. This generates a slightly different state space formulation to (8.2.1) and (8.2.7). If $x(t_1)$ has a diffuse prior then we work directly with the state space formulation (8.2.1) and (8.2.7). The continuity properties of the curves being fitted remain the same regardless of the initial assumptions placed upon $x(t_1)$. This being the case, without loss of generality, we will look at these curves when a diffuse prior is assumed for $x(t_1)$. The Generalized Smoothing Spline in this instance is

$$s(t) = \begin{cases} h^T(t, t_1)x_1|n & , t < t_1 \\ h^T x(t|n) & , t_1 \leq t < t_n \\ h^T(t, t_n)x_n|n & , t > t_n \end{cases} \quad (8.2.14)$$

The Interpolation Smoother furnishes $x(t|n)$, it requires the quantities obtained from the Kalman Filter and RTS Smoother.
For the state space formulation (8.2.1) and (8.2.7) the Kalman Filter equations ((2.3.5)-(2.3.10)) become

\[ x_{i|i-1} = T_{i}x_{i-1|i-1} \]  
\[ S_{i|i-1} = T_{i}S_{i-1|i-1}T_{i}^T + \Omega_{i} \]  
\[ d_{i} = h^TS_{i|i-1}h + \sigma^2 \]  
\[ x_{i|i} = x_{i|i-1} + S_{i|i-1}h(y_{i} - h^Tx_{i|i-1})/d_{i} \]

and

\[ S_{i|i} = S_{i|i-1} - S_{i|i-1}hh^TS_{i|i-1}d_{i}^{-1}. \]

The RTS Smoother ((3.3.4)-(3.3.6)) becomes

\[ x_{i-1|n} = x_{i|i-1} + A_{i-1}(x_{i|n} - x_{i|i-1}) \]  
\[ S_{i-1|n} = S_{i-1|i-1} + A_{i-1}(S_{i|n} - S_{i|i-1})A_{i-1}^T \]

where

\[ A_{i-1} = S_{i-1|i-1}T^TS_{i-1|i-1}^{-1}. \]

The Interpolation Smoother ((3.4.1)-(3.4.3)) is for \( t_{i-1} < t < t_{i} \)

\[ x(t|n) = T(t,t_{i-1})x_{i-1|i-1} + A_{i-1}(x_{i|n} - x_{i|i-1}) \]  
\[ S(t|n) = \Omega(t,t_{i-1}) + T(t,t_{i-1})S_{i-1|i-1}T(t,t_{i-1})^T - A(t_{i-1},t)(S_{i|n} - S_{i|i-1})A(t_{i-1},t)^T \]

where

\[ A(t_{i-1},t) = [T(t,t_{i-1})S_{i-1|i-1}T_{i} + \Gamma(t,t_{i-1})]S_{i|i-1}^{-1} \]

with

\[ \Gamma(t,t_{i-1}) = \Omega(t,t_{i-1})T(t_{i},t). \]

8.3 Smoothness properties.

Since the curve being fitted (8.2.14) is a function of the elements of \( x(t|n) \) it makes sense to investigate the smoothness properties of this quantity. THEOREM 8.3.1 states the necessary
conditions for \( x(t|n) \) to possess \( k \) continuous derivatives.

**THEOREM 8.3.1** The first \( k \) derivatives of \( x(t|n) \), which is given by (8.2.17), are continuous only if

\[
V P_j(M)^T h = 0, \; j=0,\ldots,k-1, \quad (8.3.1a)
\]

where the \( P_j \) satisfy the recurrence

\[
P_0 = I, \; P_j = \frac{d}{dt}(P_{j-1}) - MP_{j-1}, \; j=1,2,\ldots. \quad (8.3.1b)
\]

**PROOF** This proof is similar in spirit to that furnished for **LEMMA 6.5.1**.

Differentiating (8.2.17a) for \( t_{i-1} \leq t \leq t_i \) gives

\[
\dot{x}(t) = M T(t,t_{i-1}) x_{i-1|i-1} + \frac{d}{dt} A(t_i,t)(x_i | n - x_{i|i-1}) \quad (8.3.2)
\]

where

\[
\Gamma(t_i,t) = \sigma \Lambda \int_{t_{i-1}}^{t} T(t,s) VT_s ds,
\]

\[
\frac{d}{dt} A(t_i,t) = \left[ M T(t,t_{i-1}) S_{i-1|i-1} \right] + \frac{d}{dt} \Gamma(t_i,t) S_{i|1|i-1} \quad (8.3.3)
\]

and

\[
\frac{d}{dt} \Gamma(t_i,t) = M \Gamma(t_i,t) + \sigma^2 \Lambda V T(t_i,t) \quad (8.3.4)
\]

which are obtained by differentiating (8.2.17b) and (8.2.17c) respectively. Substituting (8.3.3) and (8.3.4) into (8.3.2) after simplification yields

\[
\dot{x}(t) = M x(t|n) + \sigma^2 \Lambda V T(t_i,t) S_{i|1|i-1} (x_i | n - x_{i|i-1}) \quad (8.3.5)
\]

Let \( D_i^1 \) represent the jump in the \( i \)th derivative of \( x(t|n) \). Then

\[
D_i^1 = \frac{d}{dt} x_{i+1|i} - \frac{d}{dt} x_{i|i} \nonumber
\]

\[
= \sigma^2 \Lambda V [T_{i+1} T_{i|i} S_{i|i-1}(x_{i+1|i} - x_{i+1|i}) - S_{i|i-1}(x_i | n - x_{i|i-1})]. \quad (8.3.6)
\]

The quantity \( \frac{d}{dt} x(t|n) \) is continuous at \( t=t_i \) if \( D_i^1 = 0 \). Rearranging (8.2.16a) furnishes

\[
T_{i+1} T_{i|i} S_{i|i-1}(x_{i+1|i} | n - x_{i+1|i}) = S_{i|i}(x_i | n - x_{i|i}). \quad (8.3.7)
\]
Applying the matrix inversion lemma (Appendix 1) to the formula for \( S_i^1 \) given by (8.2.15e) gives
\[
S_i^1 = S_i^1 + \sigma^{-2}hT.
\] (8.3.8)
Substituting (8.3.7) into (8.3.6) gives
\[
D^1 = \sigma^2V[S_i^1|_i(x_i|_n - x_i|i_1) - S_i^1|i-1(x_i|_n - x_i|i_1)].
\] (8.3.9)
Inserting (8.2.11d) into the above equation and using (8.3.8) yields
\[
D^1 = \sigma^2V[\sigma^{-2}hh^T(x_i|_n - x_i|i_1) - S_i^1|i-1h(y_i - h^T\xi_i|i-1)/d_i] \nonumber
\]
\[
= \sigma^2V[\sigma^{-2}hh^Tx_i|_n - \sigma^{-2}hh^Tx_i|i_1 - h(y_i - h^T\xi_i|i-1)/d_i - \nonumber
\]
\[
\sigma^{-2}hh^TS_i|i-1h(y_i - h^T\xi_i|i-1)/d_i] \nonumber
\]
\[
= \sigma^2V[\sigma^{-2}hh^Tx_i|_n - \sigma^{-2}hh^Tx_i|i_1 - \sigma^{-2}(\sigma^2 + h^TS_i|i-1h)y_i/d_i + \nonumber
\]
\[
\sigma^{-2}hh^Tx_i|i-1(\sigma^2 + h^TS_i|i-1h)/d_i].
\]
but
\[
d_i = \sigma^2 + h^TS_i|i-1h
\]
so this expression simplifies to
\[
D^1 = -\lambda Vh(y_i - h^T\xi_i|_n).
\]
For the first derivatives of \( x(t|n) \) to be continuous at the data points we require that \( D^1 = 0 \). The bracketed term will be non zero with probability one.

This result can be extended to higher derivatives. The first occurrence of a discontinuity will appear in differentiating the term involving \( VT(t_i,t)^T \) in (8.3.4). The successive derivatives of \( T(t_i,t) \) can be represented by
\[
\frac{d^j}{dt^j}T(t_i,t) = T(t_i,t)P_j(M) \] (8.3.10)
where the \( P_j \) satisfy the recurrence
\[
P_0 = I, P_j = \frac{d}{dt}(P_{j-1}) - MP_{j-1}, j=1,2,\ldots
\] (8.3.11)
Parallelling the argument in the preceeding paragraph provides the result that the first \( k \) derivatives of \( x(t|n) \) are continuous
at the knot points provided
\[ V p_{j-1}(M)^T h = 0, \quad j=1, \ldots, k. \]

In particular
\[ k \leq \dim \ker(V). \quad (8.3.12) \]

QED.

8.4 Roles that \( h \) and \( V \) play.

THEOREM 8.3.1 provides the necessary condition for the first \( k \) derivatives of \( x(t|n) \) to be continuous. This condition provides a mechanism for constructing algorithms for curves given by \( h^T x(t|n) \) which possess a variety of smoothness possibilities.

By assumption \( h \neq 0 \) in the observation equation (8.1.4a). If \( V \) has full rank then \( x(t|n) \) has a derivative that is discontinuous at the knot points regardless of what form \( h \) takes. In this instance \( h \) determines whether the curve \( h^T x(t|n) \) being fitted possesses zero through to \( (p-1) \) continuous derivatives at the knot points.

When \( V = e_p e_p^T \) and \( h = e_1 \) the Lg Smoothing Spline is recovered. Our first attempt at Generalized Smoothing Splines is recovered by setting \( V = bb^T \).

Recollecting that by definition \( V \) is positive semi definite, the necessary condition that the covariance \( \Omega_j \) given by (8.2.9) is non singular is that the \( p' \)th row and column of \( V \) is not equal to the zero vector. As mentioned earlier, problems arise when the state equation covariance is singular. In particular, the information implementations of the Kalman Filter cannot be used.

The following examples illustrate the flexibility of Generalized Smoothing Splines.
EXAMPLE 8.4.1. If \( M = \sum_{i=1}^{p-1} e_i e_{i+1}^T, \quad V = \sum_{i=p}^{k} e_i e_i^T, \quad k=p, \ldots, 1, \) and \( h = e_1 \) then a polynomial of degree \( 2p-1 \) is being fitted which possesses \( p+k-2 \) continuous derivatives at the data points.

EXAMPLE 8.4.2. If \( p=3, \quad M = \sum_{i=1}^{p-1} e_i e_{i+1}^T, \quad h = e_2 e_2^T + e_1 e_1^T \) then \( h^T x(t|n) \) is a quintic polynomial. Three choices of \( V \) are considered here.

(i) If \( V = e_3 e_3^T \), then \( h^T x(t|n) \) has two continuous derivatives at the knot points.

(ii) If \( V = e_2 e_2^T + e_3 e_3^T \), then \( h^T x(t|n) \) has one continuous derivative at the knot points.

(iii) If \( V = I_3 \), then \( h^T x(t|n) \) has one continuous derivative at the knot points.
CHAPTER 9

ALGORITHMS FOR GENERALIZED SMOOTHING SPLINES

9.1 Introduction.

Two different algorithms for Generalized Smoothing Splines will be given using the maximum likelihood approach which was proposed by Wecker and Ansley for the polynomial Smoothing Spline case for determining the smoothing parameter \( \lambda \). Here the vector of initial conditions on the underlying stochastic differential equation is a constant vector which has to be calculated. These algorithms can easily be adapted to obtaining \( \lambda \) by GCV using Ansley and Kohn's (1987) stochastic approach.

Ansley and Kohn (1987) produced an \( O(n) \) algorithm for GCV within a stochastic framework. They show it is possible to use the Kalman Filter and Fixed Point Smoother to evaluate the trace of the influence matrix, \( s(\lambda) \), in the GCV estimator which is minimized over \( \lambda \). The GCV estimator is given by

\[
GCV(\lambda) = \frac{\| (I - s(\lambda))y \|^2}{\| \text{tr}(I - s(\lambda)) \|^2}
\]

where \( s(\lambda) \) is defined in Section 1.3. The term \( s(\lambda)y \) is obtained efficiently either by Kohn and Ansley (1985) or Wecker and Ansley (1983). The central idea in evaluating the trace of the influence matrix \( s(\lambda) \) is to realize that the \( i \)'th element of \( s(\lambda)y \) is by definition \( h^Tx(t_i|n) \) for given \( \lambda \) hence \( s(\lambda) \) can be found by running the Kalman Filter and the RTS Smoother on the vectors \( e_1, \ldots, e_n \). To evaluate \( s(\lambda) \) in this fashion would require \( O(n^2) \) operations but we want the \( \text{tr}(s(\lambda)) \) which requires only the elements on the principal diagonal of \( s(\lambda) \). Ansley and Kohn (1987) use the RTS Smoother expressed in terms of the Fixed Point Smoother (as seen in its
derivation in Section 3.3) and the Kalman Filter to evaluate $\sigma(\lambda)_{ii}$, $i=1, \ldots, n$, in $O(n)$ operations and hence the $\text{tr}(\sigma(\lambda))$.

The order of magnitude of the condition numbers for the two methods will be estimated. The first method considered is square root information filter based on Paige and Saunders' (1977) approach which was given in Section 4.3 and used in Osborne and Prvan (1987a) to evaluate Lg Smoothing Splines. The second method uses a recursive square root covariance filter based on Osborne and Prvan's (1987b) approach outlined in Section 4.4 followed by a smoothing step using Prvan and Osborne's (1987) approach given in CHAPTER 5. If maximum smoothness is required then for the Generalized Smoothing Spline outlined in CHAPTER 8 a Covariance Filter approach is desirable. This is predicted by the condition number estimate for a related approach and supported by the error estimates produced by running the algorithms in single and double precision on a UNIVAC 1100 for different values of $p$ and different values of $\lambda$ which straddle the optimal $\lambda$. Machine single precision is $10^{-8}$. The problem of determining which $p$ to choose or the form of $V$ will not be addressed in this thesis. Which method to use in determining $\lambda$ will also not be considered. The two examples that are used to illustrate the two algorithms are the Gallant data (Gerig and Gallant (1975)) and the Sunspot data found in Pandit and Wu (1983). These two data sets represent the extremes of smoothing. Gallant's data is fitted by a very smooth curve, while the Sunspot data is fitted by a curve which nearly interpolates the data.
9.2 Extending the Ansley and Wecker approach to Smoothing Splines to handle Generalized Smoothing Splines.

Even though the Wecker and Ansley approach to Smoothing Splines extended to $Lg$ Smoothing Splines was considered in detail in CHAPTER 6 and extending this to handle Generalized Smoothing Splines is straightforward, its extension is included here as a convenient reference for the two algorithms to be given in sections 9.3 and 9.4.

The observations $y_i$ are assumed to be of the form

$$y_i = h^T x_i + \epsilon_i, \quad i = 1, \ldots, n. \quad (9.2.1)$$

where $h, x_i \in \mathbb{R}^p$ and $y_i, \epsilon_i \in \mathbb{R}$ with $\epsilon_i \sim N(0, \sigma^2)$. The following stochastic differential equation generates the states $x_i$.

$$\dot{x}(t) = M(t)x(t) + \sigma \sqrt{\omega}(t) \quad (9.2.2)$$

where $M(t)$ is assumed to be smooth enough for our purposes and $\omega(t)$ is a Wiener process, independent of $\epsilon_i$ in (9.2.1), satisfying for positive semi definite $V: \mathbb{R}^p \to \mathbb{R}^p$ given,

$$E((\omega_i(t+\delta) - \omega_i(t))(\omega_j(t+\delta) - \omega_j(t))) = V_{ij}\delta.$$ 

The solution to (9.2.2) can be written recursively as

$$x_{i+1} = T_{i+1}x_i + u_{i+1} \quad (9.2.3)$$

where

$$u_{i+1} = u(t_{i+1}, t_i) = \sigma \sqrt{\omega} \int_{t_i}^{t_{i+1}} T(t_{i+1}, s) \frac{d\omega(s)}{ds} ds \quad (9.2.4)$$

is independent of $x(t)$ for $t \leq t_i$ and is normally distributed with zero mean and covariance given by

$$\Omega_{i+1} = \sigma \sqrt{\omega} \int_{t_i}^{t_{i+1}} T(t_{i+1}, s)VT(t_{i+1}, s)T ds. \quad (9.2.5)$$

By considering the vector of initial conditions $x(t_i) = \alpha$ on (9.2.2) to be deterministic we can rewrite (9.2.1) as

$$y_i = h^T T(t_i, t_1)\alpha + h^T u(t_i, t_1) + \epsilon_i. \quad (9.2.6)$$
Setting
\[ z_i = z(t_i) = u(t_i, t_1) \] (9.2.7)
equation (9.2.6) can be rewritten as
\[ y_i = h^T(t_i, t_1)\alpha + h^T z_i + \epsilon_i \] (9.2.8)
and the following recursion is obtained for \( z_{i+1} \) using the argument
given in Section 6.2
\[ z_{i+1} = T_{i+1} z_i + u_{i+1} \] (9.2.9)
where \( z_i \) is independent of \( u_{i+1} \) by definition. Equations (9.2.8) and
(9.2.9) furnish the state space formulation for Wecker and Ansley's
approach.

As explained in Section 6.3 maximum likelihood estimation can be
used to determine optimal \( \lambda \). The state space formulation (9.2.8) and
(9.2.9) can be written in matrix form as
\[ y = X\alpha + v \] (9.2.10)
where
\[ y^T = [y_1, \ldots, y_n]^T, \]
\[ X = [r_1, \ldots, r_n]^T, \]
where
\[ r_i^T = h^T(t_i, t_1). \]
and
\[ v^T = [v_1, \ldots, v_n]^T, \]
where
\[ v_i = h^T z_i + \epsilon_i. \]
As outlined in the latter half of Section 6.3 the likelihood to be
maximized over \( \lambda \) is
\[ Z^*\star(\lambda | y, X) = -(n/2)\log(a^T a) - (1/2)\log(|A|) \] (9.2.11)
where
\[ a = L^{-1}(y - \hat{X}\alpha) \]
and

\[ \text{cov}(v) = \sigma^2 \Lambda = \sigma^2 \mathbf{L} \mathbf{L}^T. \]

For each different value of \( \lambda \) the quantities \( \hat{\alpha}, a \) and \( |\Lambda| \) have to be evaluated.

In Section 6.4 it was illustrated how the Kalman Filter can be employed as a computational device to evaluate \( \hat{\alpha}, a^T a, L^{-1} y, L^{-1} x \) and \( |\Lambda| \). The required recursive equations for our state space formulation (9.2.8) and (9.2.9) have the initial conditions \( z_{ji}|_{i-1} = 0 \) and \( S_{i}|_{i-1} = 0 \) and are

\[
\begin{align*}
    z_{ji}|_{i-1} &= T_i z_{ji-1}|_{i-1} \\
    S_{i}|_{i-1} &= T_i S_{i-1}|_{i-1} T_i^T + \bar{\Omega}_i \\
    d_i &= h^T S_{i}|_{i-1} h + 1 \\
    X_{ji}^* &= (X_{ji} - h^T z_{ji}|_{i-1}) / \sqrt{d_i} \\
    z_{ji}|_{i} &= z_{ji}|_{i-1} + S_{i}|_{i-1} h (X_{ji} - h^T z_{ji}|_{i-1}) / \sqrt{d_i}
\end{align*}
\]

and

\[
S_{i}|_{i} = S_{i}|_{i-1} - S_{i}|_{i-1} h h^T S_{i}|_{i-1} d_i^{-1}
\]

where \( \bar{\Omega}_i = \sigma^{-2} \Omega_i \). We have that

\[
|\Lambda| = \prod_{i=1}^{n} d_i
\]

hence

\[
\log(|\Lambda|) = \sum_{i=1}^{n} \log(d_i).
\]

The quantity \( \hat{\alpha} \) is found by minimizing

\[
\| L^{-1} y - L^{-1} x \alpha \|_2^2
\]

as outlined in Section 6.4. As a byproduct \( a^T a \), the RSS, is also obtained.

Once an optimal \( \lambda \) has been found either by MLE given in CHAPTER 6 or GCV given in CHAPTER 1 we can proceed to fitting the Generalized
Smoothing Spline. The quantities

\[ y'_i = y_i - \mathbf{h}^T(t_j, t_{1})\alpha \quad i=1, \ldots, n, \]

can now be found. The Kalman Filter for the state space formulation

\[ y'_i = \mathbf{h}^T z_i + \epsilon_i \]
\[ z_i = T_i z_{i-1} + u_i \]

\[ i=1, \ldots, n, \]

where \( \epsilon_i \sim N(0, \sigma^2) \) and \( u_i \sim N(0, \Omega_i) \) is

\[ z_j|j-1 = T_j z_{j-1}|j-1 \]
\[ S_j|j-1 = T_j S_{j-1}|j-1 T_j + \Omega_j \]
\[ d_j = \mathbf{h}^T S_j|j-1 \mathbf{h} + 1 \]
\[ z_j|j = z_j|j-1 + S_j|j-1 \mathbf{h}(y'_j - \mathbf{h}^T z_j|j-1)/d_j \]

and

\[ S_j|j = S_j|j-1 - S_j|j-1 \mathbf{h} \mathbf{h}^T S_j|j-1 d_j^{-1} \]

\[ j=1, \ldots, n, \]

initialized with \( z_1|1 = 0 \) and \( S_1|1 = 0 \). It is worthwhile remarking here that \( S_j|j-1 \) and \( S_{j-1}|j-1 \) above are the covariances of \( \sigma^{-1} z_j|j-1 \) and \( \sigma^{-1} z_{j-1}|j-1 \) respectively and that \( d_j \) is the covariance of \( \sigma^{-1}(y'_j - \mathbf{h}^T z_j|j-1) \). The RTS Smoother for the above state space formulation is given by (6.4.10) and the Interpolation Smoother is given by (6.4.11).

The Generalized Smoothing Spline is given by

\[ f(t) = \begin{cases} 
\mathbf{h}^T(t, t_1)\hat{\alpha} & t<t_1 \\
\mathbf{h}^T(t, t_1)\hat{\alpha} + \mathbf{h}^T z(t|n) & t_1<t\leq t_n \\
\mathbf{h}^T(t, t_1)\hat{\alpha} + \mathbf{h}^T(t, t_n) z_n|n & t>t_n 
\end{cases} \]

(9.2.15)

This curve's smoothness properties were discussed in Section 8.3.
9.3 Information filter algorithm for Generalized Smoothing Splines.

According to Sorensen and Wets (1982) two practical aspects to consider when designing numerical algorithms are most importantly numerical stability and to a lesser extent efficiency of the algorithms. It isn't usual for the Kalman Filter (9.2.12) to be implemented directly because it lacks numerical stability. Numerical stability according to Sorensen and Wets (1982) is very important in filtering problems because they are recursive in nature and not self correcting. This provides opportunities for calculations to become meaningless due to error growth if the algorithms are not formulated stably. Since the Generalized Smoothing Spline procedure outlined in Section 9.2 relies on recursive algorithms such as the Kalman Filter and RTS Smoother we want to employ a method which is numerically stable. One such candidate is the square root information filter developed by Paige and Saunders (1977) outlined in Section 4.3. This approach has the advantage that it provides the smoothed values as an integral part of the algorithm instead of as a separate procedure. In developing their approach they used Duncan and Horn's (1972) equivalence between Generalized Linear Least Squares and the recursive estimation problem that the Kalman Filter solves. The Generalized Linear Least Squares problem is a problem in numerical linear algebra for which reliable solution techniques are available (see, for example, Lawson and Hanson (1974)).

In the Wecker and Ansley approach the state vector $x_1$ and its associated covariance are zero thus $x_1|_1 = x_1|_n = 0$ and $S_1|_1 = S_1|_n = 0$. The Paige and Saunders' approach, as outlined in Section 4.3, in the first instance is applied to the following state space formulations
\[ z_{ji} = T_i z_{ji-1} + u_i \]
\[ X_{ji} = h^T z_{ji} + \epsilon_i \]

where

\[ X_{ji} = \begin{bmatrix} y_j \\ h^T(t_i, t_j) \end{bmatrix}, \quad j=0, \ldots, p, \quad i=1, \ldots, n. \]

\[ u_i \sim N(0, \Omega_i) \] and \[ \epsilon_i \sim N(0, \sigma^2) \] where \( \Omega_i \) is given by (9.2.5). Their method is initiated at \( \hat{R}_2 = L_2^T \) and \( \hat{b}_2 = 0 \) where

\[ (\sigma^{-2} \Omega_j)^{-1} = L_j L_j^T \]

with \( L_j \) being a Cholesky factor. The \( p+1 \) state space formulations are rewritten as

\[
0 = L_j^T z_{ji} + F_j z_{ji-1} + u_i' \\
X_{ji} = h^T z_{ji} + \epsilon_i
\]

where \( F_i = -L_i^T T_i \) and \( u_i' = -L_i^T u_i \). Note that this is slightly different from (4.3.1) and (4.3.2) in that we are not transforming the state and observation equations into having identity covariance and unit variance but covariance \( \sigma^2 I \) and variance \( \sigma^2 \). As shown in Section 4.3 the \( p+1 \) rewritten state space formulations can be written as \( p+1 \) systems of equations. We now treat these \( p+1 \) systems of equations as \( p+1 \) weighted least squares problems to be solved, as outlined in Section 4.3, for \( z_{ji|i-1} \). Its associated covariance is also available.

These quantities are required to form the matrix \( X^* \) where

\[ X^*_{ji} = (X_{ji} - h^T z_{ji|i-1})/d_i \]

with

\[ d_i = h^T S_i |i-1| h + 1. \]

Thus as outlined in the preceding section the quantities \( \hat{a} \) and RSS are obtained which are required to evaluate the log likelihood (9.2.11) for given \( \lambda \). The optimal smoothing parameter \( \lambda \) can be found
by any non linear optimization method. Every evaluation of (9.2.11) requires that the p+1 weighted linear least squares problems be solved.

Once the optimal $\lambda$ has been determined the quantities

$$y'_i = y_i - \sum_{j=1}^{p} h^T(t_j, t_1) \alpha$$  \hspace{1cm} i=1, \ldots, n,$$

are formed and the state space formulation

$$y'_i = h^T z_i + e_i$$
$$z_i = T_i z_{i-1} + u_i$$

$$i=2, \ldots, n,$$

is used. The step (4.3.11) is performed (n-2) times saving the upper block diagonals $R_k$, $R_{k, k+1}$ and $b_k$ and then step (4.3.8) is performed where $\tilde{R}_n$ and $\tilde{b}_n$ are stored. Thus the following system of equations can be solved by back substitution to obtain $z_{i|n}$, $i=2, \ldots, n$.

\[
\begin{bmatrix}
    b_2 \\
    b_3 \\
    \vdots \\
    b_{n-1} \\
    \tilde{b}_n
\end{bmatrix}
\begin{bmatrix}
    R_2 \\
    R_3 \\
    \vdots \\
    R_{n-1} \\
    \tilde{R}_n
\end{bmatrix}
\begin{bmatrix}
    R_{2,3} \\
    R_{3,4} \\
    \vdots \\
    R_{n-1,n} \\
    \tilde{R}_{n|n}
\end{bmatrix}
\begin{bmatrix}
    z_2|n \\
    z_3|n \\
    \vdots \\
    z_{n-1}|n \\
    \tilde{z}_{n|n}
\end{bmatrix}
\]

Obtaining the smoothed covariances is outlined in (4.3.13). Along the way the filtered and predicted estimates are stored along with the square roots of the covariance inverses. Now the Interpolation Smoother (6.4.11) can be employed to obtain $z(t|n)$ which is required to obtain the Generalized Smoothing Spline point estimates which are given by (9.2.15).

We now want to obtain the condition number for our implementation of Paige and Saunders' (1977) least squares approach. The condition
number is a measure of the inherent sensitivity of this approach to the actual estimation problem. This is derived for the Lg Smoothing Spline case in Osborne and Prvan (1987a) and is intimated for the Generalized Smoothing Spline case in Osborne and Prvan (1987b). The Paige and Saunders least squares approach to the Generalised Smoothing problem is to solve

$$\min_{x} \| r \|_2^2 : \quad r = Cx - c$$

(9.3.1)

where

$$C = \begin{bmatrix} 0 & -L_1^T & 0 & \cdots & 0 \\
\sigma^{-1}h & 0 & L_2^T & \sigma^{-1}h & \cdots \\
0 & \sigma^{-1}h & 0 & L_3^T & \cdots \\
\cdots & \cdots & \cdots & \cdots & \cdots \\
0 & \cdots & \cdots & \cdots & L_{n-1}^T \sigma^{-1}h \\
\sigma^{-1}h & \cdots & \cdots & \cdots & 0 \\
\sigma^{-1}h & \cdots & \cdots & \cdots & 0 \\
\sigma^{-1}h & \cdots & \cdots & \cdots & 0 \\
\cdots & \cdots & \cdots & \cdots & \cdots \\
0 & \cdots & \cdots & \cdots & L_n^T \\
0 & \cdots & \cdots & \cdots & \cdots \\
0 & \cdots & \cdots & \cdots & \cdots \\
0 & \cdots & \cdots & \cdots & \cdots \\
\end{bmatrix}, \quad c = \begin{bmatrix} y_1 \sigma^{-1} \\
y_2 \sigma^{-1} \\
y_3 \sigma^{-1} \\
\vdots \\
y_n \sigma^{-1} \end{bmatrix}$$

For our analysis the assumptions that $\delta \to 0$ and $n\delta$ is bounded as $n \to \infty$ and that the data points are equally spaced $\delta$ apart are being employed. They are reasonable assumptions to make since the aim is to smooth the given set of data. Theorem 1 in Appendix 3 gives the basic perturbation result for the linear least squares problem. It is

$$\frac{\| \bar{x} - x \|}{\| x \|} \leq \epsilon \left( \frac{2 \text{cond}(C)}{\cos(\theta)} + \tan(\theta) \text{cond}(C)^2 \right) + O(\epsilon^2)$$

(9.3.2)

where $\bar{x}$ is the solution to the perturbed least squares problem. The quantity $\epsilon$ is the maximum of the relative errors and is bounded by $\sigma_1(C)/\sigma_n(C)$. Now

$$\sin(\theta) = \frac{\| r \|}{\| c \|}.$$
(1985) and Wahba (1985) assume that \( \lambda \to 0 \) like \( \delta^p \), \( 0 < p < 1 \), as \( \delta \to 0 \). We will adopt this assumption. In the Lg Smoothing Spline case this assumption can be seen to be appropriate by considering the extreme values of \( \phi \). Recollecting that an Lg Spline minimizes (6.1.2) where \( \mu = \lambda^{-1} \), when \( \phi = 0 \) the approximation term dominates (6.1.2) for large \( n \) unless the curve being fitted tends to interpolate the data, while when \( \phi = 1 \) the penalty function dominates. The extreme eigenvalues of \( \Omega(t+\delta,t) \) are useful in obtaining the extreme eigenvalues of the data matrix \( C^T C \). LEMMA 8.3.1 proves useful in determining these extreme eigenvalues. It is worthwhile making the remark that if there exists a \( v \) satisfying (8.3.10) then the choice of \( h = v \) where \( v \) is chosen for optimal smoothness ensures that \( z(t|n) \) has \( k \) continuous derivatives. If this is the case then the smallest eigenvalue of \( \Omega(t+\delta,t) \) has its eigenvector tending to \( h \).

The following sequences aid in finding the extreme singular values of \( C \). They are

\[
\begin{align*}
\alpha_i &= h, \quad \alpha_{i+1} = T_{i+1} \alpha_i + h, \quad i = 1, \ldots, n-1, \\
\beta_i &= \gamma, \quad \beta_{i+1} = T_{i+1} \beta_i + \gamma, \quad i = 1, \ldots, n-1.
\end{align*}
\]

(9.3.3) (9.3.4)

If

\[
\tilde{\alpha}_i = \int_{t_{i-1}}^{t_i} T(t_i,s)h \, ds
\]

\[
\sum_{j=2}^{n} T(t_i,t_j) \int_{t_{j-1}}^{t_j} T(t_j,s)h \, ds
\]

then letting \( \delta \to 0 \) gives

\[
\tilde{\alpha}_i = \sum_{j=2}^{i} T(t_i,t_j) \int_{t_{j-1}}^{t_j} T(\delta,s)h \, ds \]

\[
\sum_{j=2}^{i} T(t_i,t_j) \int_{t_{j-1}}^{t_j} (h + O(\delta)) \, ds
\]
\[
\alpha_i = \sum_{j=2}^{i} T(t_i, t_j) \delta h + O(\delta^2). \tag{9.3.5}
\]

From the sequence (9.3.3) we have that

\[
\alpha_i = \sum_{j=2}^{i} T(t_i, t_j) h
\]

thus for (9.3.5) to equal \(\alpha_i\) we need to multiply it by \(n\) hence

\[
\alpha_i \sim n\bar{\alpha}_i + O(1) \tag{9.3.6}
\]

and similarly it may be shown that

\[
\beta_i \sim n\bar{\beta}_i + O(1) \tag{9.3.7}
\]

where

\[
\bar{\beta}_i = \int_{t_1}^{t_i} T(t_i, s) h \ ds.
\]

An estimate of the smallest and largest eigenvalue of \(C^T C\) is required to estimate \(\text{cond}(C^T C)\). The Rayleigh quotient is used to estimate these eigenvalues. Here it is

\[
\text{RQ}(x) = \frac{\bar{H}(x)}{\sum_{i=1}^{n} \| x_i \|^2}
\]

where

\[
\bar{H}(x) = \sigma^{-2}(h^T x_1)^2 + \| L_1^T x_1 \|^2 + \sum_{i=2}^{n} \| L_i^T x_{i-1} - L_{i-1}^T x_i \|^2 + \sigma^{-2}(h^T x_i)^2. \tag{9.3.10}
\]

It is worthwhile recollecting that \(L_1\) is the Cholesky factor of \(\Omega_1\) thus the extreme eigenvalues of \(C^T C\) can be estimated using the results for the extreme eigenvalues of \(\Omega(t+\delta, t)\) which are given in LEMMA 8.3.1. Exciting the smallest eigenvalue \(\omega_1\) of \(\Omega(t+\delta, t)\) which is \(O(\lambda \delta^{2k+1})\) and has associated eigenvector tending to \(h\), the large values of the RQ are obtained. One way of doing this is selecting \(x_i = \alpha_i\), \(i = 1, \ldots, n\) by (9.3.3). This gives

\[
\bar{H}(\alpha) = \sum_{i=1}^{n} \sigma^{-2}(h^T \alpha_i) + \| L_1^T h \|^2 + \sum_{i=2}^{n} \| L_i^T \alpha_{i-1} - L_{i-1}^T \alpha_i \|^2
\]
\[ n \sigma^{-2}(h^T \alpha_1)^2 + \| L^T_1 h \|^2 + \sum_{i=2}^{n} \left( \| L^T_1 T_1 \alpha_{i-1} + L^T_1 h - L^T_1 T_1 \alpha_{i-1} \|^2 \right) \]
\[ = \sum_{i=1}^{n} h^T \Omega^{-1}_1 h + \sigma^{-2}(h^T \alpha_1)^2 \]
\[ = \sum_{i=1}^{n} \left( \omega_1(t_i)^{-1} \| h \|^2(1 + O(\delta)) + \sigma^{-2}(h^T \alpha_1)^2 \right) \]  
(9.3.11)

but
\[ \sum_{i=1}^{n} \alpha_i^T \alpha_i \sim n^2 \sum_{i=1}^{n} \| \tilde{\alpha}(t_i) \|^2 \sim n^3 \int \| \tilde{\alpha}(s) \|^2 ds \]

so
\[ \text{RQ}(\alpha) = O(\lambda^{-1} \delta^{-2k+1}) + O(1). \]

The other possibility is to choose \( x \) such that the small eigenvalues are only excited in a small number of terms in (9.3.10) thus avoiding the \( O(n^3) \) term in (9.3.11). Choosing \( x_i = h, x_j = 0, j \neq i \), gives
\[ \text{RQ}(x) = O(\lambda^{-1} \delta^{-2k-1}) \]  
(9.3.12)

which is larger than the estimate above and cannot be improved upon.

Small values of the RQ are obtained by choosing \( x \) to only excite the largest eigenvalue of \( \Omega(t+\delta,t) \). The choice \( x_i = \alpha, x_j = 0, j \neq i \), does not work. It yields
\[ \bar{H}(x) = \omega_p(t_i) \| \alpha \|^2(1 + O(\delta)) + \| L^T_1 + T_1 + \| \alpha \|^2 + O(1) \]

and the second term excites other eigenvalues excepting \( \omega_1 \) because its associated eigenvector in the limit is \( h \). Thus the smallest \( \bar{H}(x) \) with this choice is \( O(\delta^{-1} \lambda^{-1}) \). Alternatively taking \( x_i = \beta_i \) for all \( i \) gives
\[ \bar{H}(\beta) = \sum_{i=1}^{n} \left( \omega_p(t_i)^{-1} \| \beta \|^2(1 + O(\delta)) + \sigma^{-2}(h^T \beta_i)^2 \right). \]  
(9.3.13)

Using (9.3.3) gives
\[ h^T \beta_{i+1} = h^T \beta_i + \delta h^T M(t_i) \beta_i + O(\delta^2 \beta_i) \]
\[ \approx \sum_{j=1}^{t_i} \delta h^T M(t_j) n \int T(t_j,s) \alpha ds \]
thus the second term in $\bar{H}(\beta)$ is $O(n^3)$ but
\[
\sum_{i=1}^{n} \beta_i^T \beta_i \approx n^2 \sum_{i=1}^{n} \| \bar{\beta}(t_i) \|^2 \approx n^3 \int \| \bar{\beta}(s) \|^2 \, ds
\]
so that
\[
RQ(\beta) = O(\delta^{-1}) + O(1) = O(1).
\] (9.4.14)
A smaller order is not possible.

Some knowledge about $\theta$ is required before we can evaluate (9.4.1). The assumption of independence and unit covariance for $n$ effective degrees of freedom gives
\[
\sin(\theta) \sim \sqrt{n} / \sqrt{\left( \sum \frac{y_i^2}{\sigma^2} \right)}
\] (9.4.15)
but
\[
y_i = \eta(t_i) + \varepsilon_i
\]
thus by the law of large numbers the linear term in $\varepsilon$ can be ignored
\[
\sin(\theta) \sim \sqrt{n} / (\sigma^{-2} \int \eta(t)^2 \, dt + \sigma^{-2} \sum \varepsilon_i^2)
\]
\[
\sim 1 / (\sigma^{-2} \int \eta(t)^2 \, dt + 1)
\]
\[
\sim \sigma / (\int \eta(t)^2 \, dt + 1)
\]
and so tends to a limit $> 0$ as $n \to \infty$. From (9.4.12) and (9.4.14) it follows that $\text{cond}(C) \to \infty$ as $\delta \to 0$ so that (9.4.1) must eventually be dominated by the term involving $\text{cond}(C)^2$. Thus asymptotically, the condition number of the Paige and Saunders algorithm applied to Generalized Smoothing Splines (denoted by $\text{condPS}$) is of the following order providing the assumption about $\lambda$ is true.
It was shown in CHAPTER 6 that the largest value that \( k \) can take is \( p-1 \). In this case Paige and Saunders algorithm applied to the Generalized Smoothing Spline Problem possesses a condition number estimate of the same order as the usual implementation of the Reinsch algorithm as \( \delta \to 0 \). Thus the two algorithms would appear to be essentially equivalent on the basis of stability. Both algorithms are suspect for values of \( p \), as small as 5, and modest values of \( n \), this is illustrated in TABLE 9.5.1 and TABLE 9.5.2 for the Paige and Saunders implementation.

9.4 Covariance filter algorithm for Generalized Smoothing Splines.

The Kalman Filter equations used in the approach to Generalized Smoothing Splines inspired by Wecker and Ansley (1983) can be implemented more stably as the Generalized Least Squares recursive algorithm outlined in Section 4.4. We use the Square root Fixed - interval, Discrete time Smoother developed by Prvan and Osborne (1987) to calculate the smoothed estimates.

In this section the estimate of the condition number for calculating the Generalized Smoothing Spline using Generalized Least Squares will be given because we can make progress with it. It suggests that the behaviour of the method outlined in the preceeding paragraph could be satisfactory. The Generalized Least Squares problem becomes

\[
\min_{\mathbf{s}} \mathbf{T}^\top \mathbf{s} \quad : \quad \mathbf{j}^{1/2} \mathbf{s} = \mathbf{C} \mathbf{x} - \mathbf{f}
\]  

(9.4.1)

where
\[
C = \begin{bmatrix}
    I & T_2 & \\
    h^T & & \\
    & T_2 & I \\
    h^T & & \\
    & & \ddots & I \\
    & & & 0 & 1
\end{bmatrix}
\]

(9.4.2)

\[
f^T = [x_1^T \, 0 \, y_1^T \, 0 \, y_2^T \, \ldots \, 0 \, y_n^T]^T.
\]

\[
x^T = [x_1^T \, \ldots \, x_n^T]^T.
\]

and

\[
J = \text{diag}(J_i, \, i=1, \ldots, n)
\]

with the \(J_i \in \mathbb{R}^{(p+1) \times (p+1)}\) which is defined by

\[
J_1 = \begin{bmatrix}
S_1 & 0 \\
0 & 1
\end{bmatrix}, \quad J_i = \begin{bmatrix}
\sigma^{-2} \Omega_i & 0 \\
0 & 1
\end{bmatrix}, \quad i=2, \ldots, n.
\]

Let \(L_i\) be the Cholesky factor of \(\sigma^{-2} \Omega_i\) where \(\Omega_i\) is given by (9.2.5) then

\[
J_i^{1/2} = \begin{bmatrix}
L_i & \\
& 1
\end{bmatrix}.
\]

By choosing a matrix \(Z^T\) of maximum rank such that

\[
Z^T C = 0
\]

the problem (9.4.1) is equivalent to

\[
\min_{T} \quad Z^T J_i^{1/2} s = -Z^T f
\]

where for our situation one suitable choice of \(Z\) is

\[
Z^T = \begin{bmatrix}
    h^T & -1 \\
    h^T T_2 & 0 & h^T & -1 \\
    h^T T_2 & 0 & h^T T_3 & 0 & h^T & -1 \\
    h^T T_2 & 0 & h^T T_3 & 0 & h^T T_4 & 0 \\
    \vdots & \vdots & \vdots & \vdots & \ddots & \ddots \\
    h^T T_2 & 0 & h^T T_3 & 0 & h^T T_4 & 0 & \ldots & h^T T_n & 0 & h^T & -1
\end{bmatrix}
\]

(9.4.4)

THEOREM 2 in Appendix 3 gives the following condition number for
the method outlined above, which we have denoted by \( \text{condP} \), provided \( \| s \|^2 \neq 0 \), and \( \text{cond} \{ Z^T J^{1/2} \} \) and \( \text{cond} \{ C \} \) get large as \( \delta \to \infty \).

\[
\text{condP} = \frac{\| \bar{x} - x \|}{\| x \|} = O(\text{cond} \{ Z^T J^{1/2} \} \text{cond} \{ C \}^2 \| s \| \| q \|) \quad (9.4.5)
\]

where \( \bar{x} \) is the solution to the perturbed generalized least squares problem and

\[
q = f + r
\]

is the estimate of the signal. There is no direct dependence on \( \text{cond} \{ J^{1/2} \} \). As mentioned in Section 4.4 the Generalized Least Squares problem is equivalent to

\[
\min_r r^T J^{-1} r
\]

where

\[
r = Cx - f.
\]

Thus \( s = J^{-1/2} r \) is a vector of independent zero mean normal random variables with identity covariance. From this it follows, almost surely, that

\[
\| s \| = O(\delta^{-1/2}) \quad (9.4.6)
\]

consequently

\[
\| s \| \| q \| = O(1). \quad (9.4.7)
\]

First of all we will look at estimating \( \text{cond} \{ Z^T J^{1/2} \} \) which is the square root of the condition number of \( Z^T J Z \). This latter term is easier to estimate. Assume for convenience that the data are equally spaced. The \((i,j)\)th element of the matrix \( Z^T J Z \) is given by

\[
(Z^T J Z)_{ij} = \sum_{k=1}^{\min(i-1,j-1)} h^T T_{k+1} \Omega T_{k+1} h + h^T \Omega h + \sigma^2 \delta_{ij}
\]

\((S_1|0 = 0 \text{ here})\). This can be rewritten as

\[
(Z^T J Z)_{ij} = n \int_{t_1}^{t_\tau} h^T T(\tau,s) \Omega T(\tau,s)^T h \, ds + \sigma^2 \delta_{ij} + O(\delta \| \Omega \|)
\]

where.
\[ \bar{t} = \min(t_i, t_j). \]

Let
\[ \psi(t_i, t_j) = \int_{t_1}^{\bar{t}} h(T(\bar{t}, s))^T \Omega T(\bar{t}, s) h \, ds \]
and
\[ \psi_{i,j} = \psi(t_i, t_j) + O(\delta \| \Omega \|) \]
so that
\[ Z^T J Z = \sigma^2 I + n \psi. \] (9.4.8)

Calculating the condition number of \( Z^T J Z \) requires the extreme eigenvalues of (9.4.8) since the condition number is the ratio of the largest eigenvalue to the smallest eigenvalue. In general \( \psi(s, t) \) can be expected to be a slowly varying function. Since \( \psi(s, t) \) is a slowly varying function it is reasonable to expect that the order of magnitude of the largest eigenvalue can be estimated by forming the Rayleigh quotient with the test vector
\[ 1^T = [1, 1, \ldots, 1]^T, \]
thus
\[ RQ(1) \approx \sigma^2 + \sum \sum \psi_{i,j} \]
\[ \approx \sigma^2 + n^2 \int \int \psi(s, t) \, ds \, dt \]
\[ \approx \sigma^2 (1 + O(\lambda \delta^{-1})) \]
since \( \| \Omega \| = \sigma^2 \lambda O(\delta) \). Some cancellation has to be introduced into the Rayleigh quotient to estimate the order of magnitude of the small eigenvalues. This is achieved by choosing a vector which differences adjacent elements giving
\[ RQ \approx \sigma^2 + \sum \sum (-1)^{i+j} \psi_{i,j} \]
\[ \approx \sigma^2 + n^2 \int \int \psi(s, t) \, ds \, dt \times \delta \text{(differencing)} \]
\[ \approx \sigma^2 + n^2 \delta^2 \sigma \lambda \]
\[ \approx \sigma^2(1 + O(\lambda)) \]

where \( \sigma^2 \) comes from \( \| \Omega \| \) in the term \( \psi(s,t) \). There is no advantage in developing a test vector to give a higher order of differencing because the \( O(1) \) term will dominate. Again, if as in Section 9.3, the assumption that \( \lambda = O(\delta^\phi) \), \( 0 < \phi < 1 \), is made, then the estimates of the smallest and largest eigenvalues of \( Z^TJZ \) above provide the following strictly lower bound condition number

\[
\text{cond}(Z^TJZ) = O(\lambda \delta^{-1})
\]

hence

\[
\text{cond}(Z^Tj^{1/2}) = O(\lambda^{1/2} \delta^{-1/2}). \tag{9.4.9}
\]

The Rayleigh quotient approach can be used once more to estimate the order of magnitude of the condition number of \( C \). The Rayleigh quotient for this matrix is

\[
\text{RQ}(x) = \frac{\| x \|_2^2 + \sum \| w_i \|^2 + \sum (h^T x_i)}{\sum \| x_i \|^2} \tag{9.4.10}
\]

where

\[ w_i = x_{i+1} - T_{i+1} x_i. \tag{9.4.11} \]

To minimize the RQ it is necessary that both \( w_i \) and \( h^T x_i \) be made small while simultaneously making \( x_i \) large. There is no restriction in scaling \( w_i \) so that \( \| w_i \| = 1 \). Recollecting that \( T(t_{i+1}, t_i) \) is the fundamental matrix solution to the following homogeneous differential equation

\[ \dot{x}(t) = M(t)x(t) \]

we have that

\[ T(t_{i+1}, t_i) = I + \delta M(t_i) + O(\delta^2). \tag{9.4.12} \]

If at each step \( w_i \) is chosen to nearly maximally increase \( \| x_i \| \) then it follows that \( w_i \) will be slowly varying and as a consequence its
norm will be close to 1. As mentioned above to minimize the RQ it is necessary to keep $h^T x_i$ small. If we choose 

$$h^T x_k = 0, \; k=1,\ldots,i+1,$$

then

$$h^T x_{i+1} = 0 = h^T (w_i + T(t_{i+1}, t_i) x_i) = h^T (w_i + \delta M(t_i) x_i) + O(\delta^2 \|x_i\|). \; (9.4.13)$$

This suggests that $w_i$ can be chosen recursively by

$$w_i = -\delta M(t_i) x_i + z_i, \quad h^T z_i = 0 \quad (9.4.14)$$

where $z_i$ is chosen to increase $x_{i+1}$ and $w_i$ has been made small as required. From (9.4.11)

$$x_{i+1} = w_i + T_{i+1} x_i.$$

But using (9.4.14) this becomes

$$x_{i+1} = -\delta M(t_i) x_i + z_i + T_{i+1} x_i = z_i + (T_{i+1} - \delta M(t_i)) x_i = z_i + (I + \delta M(t_i)) x_i = x_i + z_i + O(\delta^2 \|x_i\|)$$

$$= \sum_{j=1}^{i} z_i + O(\delta^2 \|x_i\|) = n \int z(t) dt + O(1). \quad (9.4.15)$$

We have that $\|w_i\| = O(1)$ as required and $h^T x_i = O(1)$. Now

$$\sum_{i=1}^{n} \|x_i\|^2 = \sum_{i=1}^{n} \int_{t_1}^{t_i} z(t) dt + O(1)$$

$$= n^2 \int_{t_1}^{t_n} z(t) dt ds + O(\delta^{-1}) = O(\delta^{-2})$$

hence

$$\text{RQ}(x) = O(\delta^2).$$

The largest singular value of $C$ also has to be estimated in order to
get an idea of the condition number of $C$. To get an estimate of the largest eigenvalue of $C$ we apply Gershgorin's theorem (it can be found in, for example, Stoer and Bulirsch (1980) p385) to $C^T C$. This matrix has band structure. The individual entries of $C^T C$ are all $O(1)$ hence by Gershgorin's theorem and the band structure of this matrix the largest eigenvalue of $C^T C$ is $O(1)$. Thus

$$\text{cond}(C) = O(\delta^{-2}).$$

Using this and (9.4.9) in (9.4.5) gives

$$\text{cond}P = O(\lambda^{1/2} \delta^{-5/2}) \quad (9.4.16)$$

since $\text{cond}(Z^T, 1/2)$ and $\text{cond}(C)$ get large as $\delta \to 0$. This condition number estimate is independent of $p$.

Now

$$\frac{\text{cond}P}{\text{cond}PS} = O(\lambda^{3/2} \delta^{2k-3/2}),$$

thus the condition number estimate of the algorithm outlined in this section is generally smaller than the condition number estimate for the method outlined in the last section provided that the assumption about $\lambda$ given earlier is realistic (all the condition numbers quoted are lower bounds). In particular, for the popular case of cubic smoothing splines the ratio of the condition numbers is

$$\frac{\text{cond}P}{\text{cond}PS} = \lambda^{3/2} \delta^{1/2}$$

which differentiates in favour of the algorithm outlined here provided the assumption about $\lambda$ mentioned earlier is valid. This estimate is for the non recursive form based on Paige's device but it suggests that the recursive form will have advantages over the method outlined in the preceeding section. Numerical results are given in the next section which support the suggestion that the recursive algorithm based on Paige's (1979) device might well be superior to the algorithm based on Paige and Saunders' (1977) approach. The recursive algorithm
based on Paige's (1979) device is a square root covariance filter while the algorithm based on Paige and Saunders' (1977) approach is a square root information filter. Since the state transition equation covariances are given and not the inverses of the covariances we would expect the algorithm based on Paige's device to perform better.

As mentioned in the introduction the algorithms in this section and the last one can in essence also be applied to the diffuse prior algorithms outlined in CHAPTER 7 and can also be employed in using GCV instead of MLE to obtain the optimal smoothing parameter $\lambda$.

9.5 Numerical Results.

The least squares implementation based on Paige and Saunders' (1977) method and the generalized recursive least squares implementation based on Paige's (1979) device of the Generalized Smoothing Spline was used on two data sets, the Gallant data (Gerig and Gallant (1975))(here $n=72$) and the Sunspot data (Pandit and Wu (1987) p487)(where $n=176$) for $h = e_1$, for a range of values of $p$, $\lambda$ and $V$. These two data sets represent the extremes of smoothing. The Gallant data is fitted by a very smooth curve which appears to require a lower order of smoothing spline (there is overfitting for higher values of $p$), while the Sunspot data is fitted by a Smoothing Spline which tries very hard to interpolate the data. The estimates of the errors in the computation are obtained by running the algorithms in both single and double precision on a UNIVAC 1100 computer which has $10^{-8}$ accuracy in single precision. Clearly the results in the accompanying tables illustrate that the recursive algorithm based on Paige's (1979) device performs better than the algorithm based on Paige and Saunders' (1977) method for different $V$. Results are given
for a range of values of $\lambda$ which straddle the optimal $\lambda$ computed by the maximum likelihood procedure discussed in Section 9.3 and CHAPTER 6. As predicted by the condition number for Paige and Saunders algorithm, the accuracy improves when the continuity conditions are relaxed. This corresponds to increasing the rank of $V$. This does not happen when Paige's device is implemented. For the Sunspot data $\lambda$ is very large and is more or less independent of the continuity conditions. This is consistent with interpolating the data (see Wecker and Ansley (1983)). Evidence for oversmoothing with the Gallant data is provided by the optimal $\lambda$ decreasing rapidly (from $10^{-2}$ to $10^{-6}$) as $p$ is increased from 2 to 4. In summary, the algorithm proposed in Section 9.5 performs better than the existing algorithms. The two algorithms mentioned above also have the additional advantage that the continuity conditions can be relaxed when we possess a priori information which suggests that $2p-2$ continuous derivatives is too stringent a constraint.

A few graphs are included to illustrate that for that for optimal $\lambda$ the Gallant data is fitted with a very smooth curve and the Sunspot data is fitted with a curve which nearly interpolates the data. Formal confidence intervals are attached to the graphs.
Table 9.5.1. Gallant Data error estimates.

<table>
<thead>
<tr>
<th>$p$</th>
<th>$V$</th>
<th>$\lambda$</th>
<th>Paige error</th>
<th>Paige &amp; Saunders error</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>$e_2^T e_2$</td>
<td>0.0001</td>
<td>$0.12 \times 10^{-8}$</td>
<td>$0.24 \times 10^{-6}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.001</td>
<td>$0.18 \times 10^{-8}$</td>
<td>$0.20 \times 10^{-6}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.01</td>
<td>$0.63 \times 10^{-8}$</td>
<td>$0.31 \times 10^{-7}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.1</td>
<td>$0.44 \times 10^{-8}$</td>
<td>$0.41 \times 10^{-7}$</td>
</tr>
<tr>
<td></td>
<td>$I_2$</td>
<td>0.0001</td>
<td>$0.75 \times 10^{-8}$</td>
<td>$0.14 \times 10^{-6}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.001</td>
<td>$0.87 \times 10^{-8}$</td>
<td>$0.11 \times 10^{-6}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.01</td>
<td>$0.61 \times 10^{-8}$</td>
<td>$0.43 \times 10^{-7}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.1</td>
<td>$0.41 \times 10^{-8}$</td>
<td>$0.34 \times 10^{-7}$</td>
</tr>
<tr>
<td>3</td>
<td>$e_3^T e_3$</td>
<td>0.0001</td>
<td>$0.23 \times 10^{-7}$</td>
<td>$0.30 \times 10^{-5}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.001</td>
<td>$0.41 \times 10^{-7}$</td>
<td>$0.20 \times 10^{-5}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.01</td>
<td>$0.25 \times 10^{-7}$</td>
<td>$0.31 \times 10^{-6}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.1</td>
<td>$0.10 \times 10^{-7}$</td>
<td>$0.36 \times 10^{-7}$</td>
</tr>
<tr>
<td></td>
<td>$I_3 - e_1^T e_1$</td>
<td>0.0001</td>
<td>$0.26 \times 10^{-7}$</td>
<td>$0.30 \times 10^{-5}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.001</td>
<td>$0.40 \times 10^{-7}$</td>
<td>$0.20 \times 10^{-5}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.01</td>
<td>$0.22 \times 10^{-7}$</td>
<td>$0.31 \times 10^{-6}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.1</td>
<td>$0.84 \times 10^{-7}$</td>
<td>$0.36 \times 10^{-7}$</td>
</tr>
<tr>
<td></td>
<td>$I_3$</td>
<td>0.0001</td>
<td>$0.23 \times 10^{-7}$</td>
<td>$0.13 \times 10^{-5}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.001</td>
<td>$0.31 \times 10^{-7}$</td>
<td>$0.33 \times 10^{-6}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.01</td>
<td>$0.20 \times 10^{-7}$</td>
<td>$0.24 \times 10^{-6}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.1</td>
<td>$0.87 \times 10^{-8}$</td>
<td>$0.28 \times 10^{-7}$</td>
</tr>
<tr>
<td></td>
<td>$e_1^T e_4$</td>
<td>$e_3^T e_3$</td>
<td>$e_4^T e_4$</td>
<td>$I_4 - e_1^T e_1$</td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
<td>-------------</td>
<td>-------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>0.0001</td>
<td>$0.80 \times 10^{-7}$</td>
<td>$0.39 \times 10^{-7}$</td>
<td>$0.38 \times 10^{-7}$</td>
<td>$0.33 \times 10^{-7}$</td>
</tr>
<tr>
<td>0.001</td>
<td>$0.31 \times 10^{-7}$</td>
<td>$0.27 \times 10^{-7}$</td>
<td>$0.22 \times 10^{-7}$</td>
<td>$0.25 \times 10^{-7}$</td>
</tr>
<tr>
<td>0.01</td>
<td>$0.13 \times 10^{-6}$</td>
<td>$0.10 \times 10^{-6}$</td>
<td>$0.10 \times 10^{-6}$</td>
<td>$0.13 \times 10^{-6}$</td>
</tr>
<tr>
<td>0.1</td>
<td>$0.46 \times 10^{-6}$</td>
<td>$0.39 \times 10^{-6}$</td>
<td>$0.39 \times 10^{-6}$</td>
<td>$0.30 \times 10^{-6}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.0001</td>
<td>$0.53 \times 10^{-3}$</td>
<td>$0.14 \times 10^{-3}$</td>
<td>$0.43 \times 10^{-5}$</td>
<td>$0.75 \times 10^{-6}$</td>
</tr>
<tr>
<td>0.001</td>
<td>$0.19 \times 10^{-4}$</td>
<td>$0.19 \times 10^{-5}$</td>
<td>$0.13 \times 10^{-6}$</td>
<td>$0.64 \times 10^{-7}$</td>
</tr>
<tr>
<td>0.01</td>
<td>$0.37 \times 10^{-3}$</td>
<td>$0.20 \times 10^{-4}$</td>
<td>$0.22 \times 10^{-5}$</td>
<td>$0.99 \times 10^{-6}$</td>
</tr>
<tr>
<td>0.1</td>
<td>$0.27 \times 10^{-3}$</td>
<td>$0.24 \times 10^{-4}$</td>
<td>$0.41 \times 10^{-5}$</td>
<td>$0.16 \times 10^{-5}$</td>
</tr>
</tbody>
</table>
Table 9.5.2. Sunspot Data error estimates.

<table>
<thead>
<tr>
<th>p</th>
<th>V</th>
<th>λ</th>
<th>Paige error</th>
<th>Paige &amp; Saunders error</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>$e_2 e_2^T$</td>
<td>0.1</td>
<td>$0.27 \times 10^{-5}$</td>
<td>$0.27 \times 10^{-4}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.0</td>
<td>$0.14 \times 10^{-5}$</td>
<td>$0.18 \times 10^{-4}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10.0</td>
<td>$0.39 \times 10^{-6}$</td>
<td>$0.14 \times 10^{-5}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100.0</td>
<td>$0.13 \times 10^{-5}$</td>
<td>$0.17 \times 10^{-5}$</td>
</tr>
<tr>
<td>$I_2$</td>
<td></td>
<td>0.1</td>
<td>$0.18 \times 10^{-5}$</td>
<td>$0.21 \times 10^{-4}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.0</td>
<td>$0.11 \times 10^{-5}$</td>
<td>$0.66 \times 10^{-5}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10.0</td>
<td>$0.91 \times 10^{-6}$</td>
<td>$0.12 \times 10^{-5}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100.0</td>
<td>$0.11 \times 10^{-5}$</td>
<td>$0.15 \times 10^{-5}$</td>
</tr>
<tr>
<td>3</td>
<td>$e_3 e_3^T$</td>
<td>0.1</td>
<td>$0.21 \times 10^{-4}$</td>
<td>$0.75 \times 10^{-3}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.0</td>
<td>$0.83 \times 10^{-4}$</td>
<td>$0.33 \times 10^{-3}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10.0</td>
<td>$0.26 \times 10^{-3}$</td>
<td>$0.17 \times 10^{-2}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100.0</td>
<td>$0.60 \times 10^{-3}$</td>
<td>$0.10 \times 10^{-2}$</td>
</tr>
<tr>
<td>$I_3 - e_1 e_1^T$</td>
<td></td>
<td>0.1</td>
<td>$0.10 \times 10^{-4}$</td>
<td>$0.59 \times 10^{-4}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.0</td>
<td>$0.52 \times 10^{-4}$</td>
<td>$0.15 \times 10^{-3}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10.0</td>
<td>$0.19 \times 10^{-3}$</td>
<td>$0.65 \times 10^{-4}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100.0</td>
<td>$0.37 \times 10^{-3}$</td>
<td>$0.96 \times 10^{-4}$</td>
</tr>
<tr>
<td>$I_3$</td>
<td></td>
<td>0.1</td>
<td>$0.10 \times 10^{-4}$</td>
<td>$0.67 \times 10^{-4}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.0</td>
<td>$0.37 \times 10^{-4}$</td>
<td>$0.95 \times 10^{-4}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10.0</td>
<td>$0.10 \times 10^{-3}$</td>
<td>$0.77 \times 10^{-4}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100.0</td>
<td>$0.99 \times 10^{-4}$</td>
<td>$0.10 \times 10^{-3}$</td>
</tr>
<tr>
<td>(4e_4^T)</td>
<td>0.1</td>
<td>(0.55 \times 10^{-2})</td>
<td>(0.32 \times 10^1)</td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td>(0.39 \times 10^{-2})</td>
<td>(0.88 \times 10^0)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10.0</td>
<td>(0.81 \times 10^{-2})</td>
<td>(0.39 \times 10^{-1})</td>
<td></td>
<td></td>
</tr>
<tr>
<td>100.0</td>
<td>(0.28 \times 10^{-1})</td>
<td>(0.10 \times 10^0)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(e_3^Te_3^T)</td>
<td>0.1</td>
<td>(0.38 \times 10^{-2})</td>
<td>(0.26 \times 10^0)</td>
<td></td>
</tr>
<tr>
<td>(e_4^T e_4^T)</td>
<td>1.0</td>
<td>(0.32 \times 10^{-2})</td>
<td>(0.62 \times 10^{-1})</td>
<td></td>
</tr>
<tr>
<td>10.0</td>
<td>(0.87 \times 10^{-2})</td>
<td>(0.95 \times 10^{-1})</td>
<td></td>
<td></td>
</tr>
<tr>
<td>100.0</td>
<td>(0.23 \times 10^{-1})</td>
<td>(0.68 \times 10^{-1})</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(I_4 - e_1^T e_1^T)</td>
<td>0.1</td>
<td>(0.33 \times 10^{-2})</td>
<td>(0.48 \times 10^{-1})</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>(0.24 \times 10^{-2})</td>
<td>(0.13 \times 10^{-1})</td>
<td></td>
</tr>
<tr>
<td></td>
<td>10.0</td>
<td>(0.73 \times 10^{-2})</td>
<td>(0.12 \times 10^{-1})</td>
<td></td>
</tr>
<tr>
<td></td>
<td>100.0</td>
<td>(0.12 \times 10^{-1})</td>
<td>(0.86 \times 10^{-2})</td>
<td></td>
</tr>
<tr>
<td>(I_4)</td>
<td>0.1</td>
<td>(0.33 \times 10^{-2})</td>
<td>(0.19 \times 10^{-1})</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>(0.24 \times 10^{-2})</td>
<td>(0.81 \times 10^{-2})</td>
<td></td>
</tr>
<tr>
<td></td>
<td>10.0</td>
<td>(0.49 \times 10^{-2})</td>
<td>(0.44 \times 10^{-2})</td>
<td></td>
</tr>
<tr>
<td></td>
<td>100.0</td>
<td>(0.57 \times 10^{-2})</td>
<td>(0.30 \times 10^{-2})</td>
<td></td>
</tr>
</tbody>
</table>
FIGURE 9.5.1. Weight-Height Ratio of Preschool Boys Estimated Using $p = 2$ and $V = e^{\frac{T}{p}}$. (A triangle indicates observations, the middle broken line indicates the estimate and the outer two lines are the 95% confidence limits.)
FIGURE 9.5.2. Weight-Height Ratio of Preschool Boys Estimated Using $p = 3$ and $V = e_p e_p^T$. (A triangle indicates observations, the middle broken line indicates the estimate and the outer two lines are the 95% confidence limits.)
FIGURE 9.5.3. Sunspot Data Estimated Using $p = 2$ and $V = e_p e_p^T$. (A triangle indicates observations, the middle broken line indicates the estimate and the outer two lines are the 95% confidence limits.)
FIGURE 9.5.4. Sunspot Data Estimated Using $p = 3$ and $V = e_p e_p^T$. (A triangle indicates observations, the middle broken line indicates the estimate and the outer two lines are the 95% confidence limits.)
APPENDIX 1.

LEMMA 1. If $\text{tr}(AB) = 0 \ \forall A$ then $B = 0$, where $A$ is an $m \times n$ matrix and $B$ is an $n \times m$ matrix.

PROOF. Let $A = [a_{ij}]$ and $B = [b_{ij}]$. Now

$$\text{tr}(AB) = \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij} b_{ji}.$$ 

If $\text{tr}(AB) = 0 \ \forall A$ then we have that

$$\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij} b_{ji} = 0 \ \forall A.$$ 

Specializing $A$ to be $e_i e_j^T$

$$b_{ji} = 0, \ \text{i=1, ..., m, j=1, ..., n}.$$ 

QED.

MATRIX INVERSION LEMMA. For $R, W \in \mathbb{R}^{p \times n}$ and a positive definite $Q \in \mathbb{R}^{n \times n}$ the following relationship holds

$$RW(WRW^T + Q)^{-1} = (W^TQ^{-1}W + R^{-1})^{-1}W^TQ^{-1}.$$ 

PROOF. This is proved by first postmultiplying the above expression by $(WRW^T + Q)$ and then premultiplying by $(W^TQ^{-1}W + R^{-1})$. 

QED.
LEMMA 7.4.2. Letting \( \gamma^2 \rightarrow \infty \), the following equations provide the smoothed estimate \( x(t|n) \) for \( t_{i-1} \leq t \leq t_i \) and its associated covariance for the state space formulation (7.4.1) with \( x_1 \sim N(0, \gamma^2 I) \). The recursion is

\[
x(t|n) = T(t, t_1)\mathcal{E}_{i-1} + (I + \bar{A}_i X_i)(x_{i|n} - T(t_i, t_1)\mathcal{E}_{i-1})
\]

(7.4.27a)

and

\[
S(t|n) = T(t, t_1)[(I + \bar{A}_i X_i)S_{i|n}(I + \bar{A}_i X_i)^T - \bar{F}(t_i, t) - E_i X_i^T \bar{A}_i^T
- \bar{A}_i X_i^T \bar{A}_i]T(t_i, t_1)^T
\]

where

\[
\bar{A}_i = R_{i-1} + \bar{F}(t_i, t) - E_i
\]

with

\[
\bar{F}(t_i, t) = T(t, t_1)^{-1}F(t_i, t)T(t_i, t_1)^{-T}
\]

(7.4.27d)

and \( X_i \) satisfies

\[
X_i = Q_{i-1}X_i Q_{i-1} \quad (7.4.23)
\]

\[
Q_{i-1}E_i X_i Q_{i-1} = Q_{i-1} \quad (7.4.24)
\]

The quantities \( E_i \) and \( R_i \) are given by (7.4.10a) and (7.4.10b).

PROOF. It is worthwhile noting that \( X_i \) can be determined and is symmetric. Referring to the Interpolation Smoother given by (3.4.1) - (3.4.3) we have that

\[
x(t|n) = T(t, t_{j-1})x_{1-1|i-1} + A(t_i, t)(x_{i|n} - x_{i|i-1})
\]

(A2.1)

where

\[
A(t_i, t) = [T(t, t_{i-1})S_{i-1|i-1}^{-1} + \Omega(t, t_{i-1})T(t_i, t)^T]S_{i|i-1}^{-1}
\]

(A2.2)

and
\[ S(t|n) = T(t,t_{i-1})S_{i-1}|_{i-1} T(t,t_{i-1})^T - A(t_i,t)(S_{i}|_{i-1} - S_{i}|_n)A(t_i,t)^T \]  
(A2.3)

for \( t_{i-1} \leq t < t_i \).

The quantities \( S_{i}|_{i-1} \) and \( S_{i-1}|_{i-1} \) from THEOREM 7.4.1 and LEMMA 7.4.1 can be written as

\[ S_{i}|_{i-1} = T(t_{i-1},t_1)\{\gamma^2(1 - Q_{i-1}) + E_i + \gamma^{-2}E_{i+1} + \cdots \}T(t_1,t_{i-1})^{-1}. \]  
(A2.4)

and

\[ S_{i-1}|_{i-1} = T(t_{i-1},t_1)\{\gamma^2(1 - Q_{i-1}) + R_{i-1} + \gamma^{-2}R_{i-11} + \cdots \} \times T(t_{i-1},t_1)^T. \]  
(A2.5)

In the proof to LEMMA 7.4.1 it was shown that

\[ X_1 = Q_{i-1}X_1Q_{i-1} \]

and

\[ Q_{i-1}E_{i+1}Q_{i-1} = Q_{i-1}. \]

Using (A2.4) and (A2.5) in (A2.2) we get

\[ A(t_i,t) = (T(t,t_{i-1})T(t_{i-1},t_1)\{\gamma^2(1 - Q_{i-1}) + R_{i-1} + \gamma^{-2}R_{i-11} + \cdots \} \times T_{i-1}^T \times T(t_{i-1},t_1)^T + \Gamma(t_i,t))T(t_i,t_1)^{-1} \times \]

\[ \{\gamma^2(1 - Q_{i-1}) + E_i + \gamma^{-2}E_{i+1} + \cdots \}^{-1}T(t_1,t_{i-1})^{-1}. \]

Setting

\[ \bar{R}_{i-1} = R_{i-1} + \bar{R}(t_i,t) \]

we look at

\[ \{\gamma^2(1 - Q_{i-1}) + \bar{R}_{i-1} + \gamma^{-2}\bar{R}_{i-11} + \cdots \} \times \{\gamma^2(1 - Q_{i-1}) + E_i + \gamma^{-2}E_{i+1} + \cdots \}^{-1}. \]  
(A2.6)

As seen in LEMMA 7.4.1 the inverse term can be written as

\[ X_1 + \gamma^{-2}X_{i+1} + \gamma^{-4}X_{i+2} + \cdots , \]

thus equation (A2.6) becomes

\[ \{\gamma^2(1 - Q_{i-1}) + \bar{R}_{i-1} + \gamma^{-2}\bar{R}_{i-11} + \cdots \}(X_1 + \gamma^{-2}X_{i+1} + \gamma^{-4}X_{i+2} + \cdots) \]
\[ = \gamma^2 (I - Q_{i-1})X_i + \bar{R}_{i-1}X_i + (I - Q_{i-1})X_{i1} + \gamma^{-2}(R_{i-11}X_{i1} + \bar{R}_{i-1}X_{i1} + (I - Q_{i-1})X_{i2} + O(\gamma^{-4}). \]

Using the relationships (7.4.23) and (7.4.24) this simplifies to

\[ I + (\bar{R}_{i-1} + E_1) + \gamma^{-2}((R_{i-11} - E_{11})X_i + (\bar{R}_{i-1} - E_1)X_{i1}) + O(\gamma^{-4}). \]

Setting

\[ \bar{A}_i = \bar{R}_{i-1} - E_i \]

and

\[ A_{11} = R_{i-11} - E_{11} \]

the above expression can be written more compactly as

\[ I + \bar{A}_iX_i + \gamma^{-2}(A_{11}X_i + \bar{A}_{11}X_{i1}) + O(\gamma^{-4}) \]

thus

\[ A(t_{1.i},t) = T(t_{1.i},t_{i-1})(I + \bar{A}_iX_i + \gamma^{-2}(\bar{A}_{11}X_i + \bar{A}_{11}X_{i1}) + O(\gamma^{-4})T(t_{1.i},t_{i-1})^{-1}. \]  

(A2.7)

Taking the limit as \( \gamma^2 \to \infty \) the interpolation smoothed estimate for \( t_{i-1} \leq t \leq t_i \) is given by

\[ x(t|n) = T(t_{1.i-1})x_{i-1|i-1} + (I + \bar{A}_i)(x_i|n - x_i|i-1) \]

where \( x_{i-1|i-1} \) is obtained from solving

\[ (V_{1-i}^T)^+ y_{1-i} = \xi_{1-i} \]

(which is defined in Section 7.4) since

\[ x_{i-1|i-1} = T(t_{i-1}.t_{1.i})\xi_{1-i}, \]

thus (7.4.27a) is obtained.

The Interpolation Smoothed estimate's covariance for \( t_{i-1} \leq t \leq t_i \) is given by (A2.3). In this expression the possible unbounded terms are

\[ T(t_{1.i-1})S_{i-1|i-1}T(t_{i-1},t_{1.i}) - A(t_{1.i},t)S_{i|i-1}A(t_{1.i},t)^T \]

which using (A2.4), (A2.5) and (A2.7) become
Inspecting the following term gives using (7.4.23) and (7.4.24)

\[ T(t, t_1) \{ (I - Q_{t-1}) + R_{t-1} + \gamma^{-2} \}
\]

\[ \times \{ (I + A_{t-1}) + E_{t-1} + \gamma^{-2} (I_{t-1} + A_{t-1}) \}
\]

\[ \times \{ T(t, t_1) \{ I + A_{t-1}X_{t-1} + (A_{t-1}X_{t-1} + A_{t-1}X_{t-1}) + \ldots \} \}
\]

After letting \( \gamma^2 \to \infty \) the possible unbounded terms become

\[ T(t, t_1) \{ -f(t_1, t) - E_1X_1T_{t-1} - \Delta_1X_1T_{t-1} + O(\gamma^{-2}) \}
\]

This is inserted into equation (A2.3), after some manipulation equation (7.4.27b) is obtained. QED.
APPENDIX 3
PERTURBATION RESULTS FOR LEAST SQUARES PROBLEMS

The following theorem found in Golub and Van Loan (1983) provides the perturbation results for the least squares problem.

**THEOREM 1.** Suppose \( x, r, \hat{x}, \text{ and } \hat{r} \) satisfy

\[
\| Ax - b \|_2 = \min, \quad r = b - Ax.
\]

\[
\| (A + \delta A)x - (b + \delta b) \|_2 = \min, \quad \hat{r} = (b + \delta b) - (A + \delta A)\hat{x},
\]

where \( A, \delta A \in \mathbb{R}^{n \times p}, (n \geq p) \) and \( b \neq 0, \delta b \in \mathbb{R}^n \). Assume that

\[
\varepsilon = \max \left\{ \frac{\| \delta A \|_2}{\| A \|_2}, \frac{\| \delta b \|_2}{\| b \|_2} \right\} < \frac{\sigma_1(A)}{\sigma_2(A)}
\]

where \( \sigma_1(A) \) is the smallest singular value of \( A \) and \( \sigma_n(A) \) is the largest singular value of \( A \) so that

\[
\sin(\theta) = \frac{\| Ax - b \|_2}{\| b \|_2} \neq 1.
\]

Then

\[
\frac{\| \hat{x} - x \|_2}{\| x \|_2} \leq \varepsilon \left( \frac{2\text{cond}(A)}{\cos(\theta)} + \tan(\theta)\text{cond}(A)^2 \right) + O(\varepsilon^2)
\]

(A3.1)

\[
\frac{\| \hat{r} - r \|_2}{\| r \|_2} \leq \varepsilon \left( 1 + 2\text{cond}(A)\min(1,n-p) \right) + O(\varepsilon^2).
\]

(A3.2)

The Generalized Linear Least Squares problem is

\[
\min \| W^{-1/2}(y - Cx) \|_2, \quad W \in \mathbb{R}^{n \times n}, \quad C \in \mathbb{R}^{n \times p}, \quad x
\]

where the data is assumed to be of the form

\[
y = Cx + w
\]

where \( w \) is an unknown noise vector with zero mean and covariance \( W \).

Paige (1979) reformulated this problem as

\[
\min \, v^T v : Bv = y - Cx, \quad v, x
\]

where
\[ W = BB^T, \quad w = Bv, \quad E(v) = 0, \text{ and } E(vv^T) = I. \]

It was seen in Section 4.4 that by choosing a matrix \( Z^T \) of maximum rank such that

\[ Z^T C = 0 \]

the equivalent problem is

\[ \min_{v} v^T v : Z^T Bv = -Z^T y. \]

Paige (1979) furnishes the following perturbation result.

**THEOREM 2.** Let \( \delta_i = t_i - t_{i-1} \). If \( \|v\|_2 \neq 0 \), and if \( \text{cond}(Z^T B) \) and \( \text{cond}(C) \) get large as \( \delta_i \to 0 \), \( \forall i \), then the condition number associated with the GLS estimate using Paige's device is

\[ \frac{\|\hat{x} - x\|_2}{\|x\|_2} \leq \alpha \text{cond}(Z^T B) \text{cond}(C) \frac{2\|s\|_2}{\|q\|_2} \quad (A3.3) \]

where \( \hat{x} \) is the solution to the perturbed Generalized Linear Least Squares problem and

\[ q = f + r \]

is an estimate of the signal.
REFERENCES.


