Biometrika (1989), **76**, 1, *pp.* 65–79
Printed in Great Britain

A fast algorithm for signal extraction, influence and cross-validation in state space models

BY ROBERT KOHN AND CRAIG F. ANSLEY
Australian Graduate School of Management, University of New South Wales, Kensington, N.S.W. 2033, Australia

SUMMARY

A fast algorithm is developed for computing the conditional mean and variance of the signal given the observations in a signal plus noise model. The resulting recursions can be applied immediately to provide new and efficient formulae for smoothing part or all of the state vector. The ideas of studentized residuals and leverage from regression analysis are generalized to state space models, and the new algorithm is used to compute the various measures. The results are also applied to obtain a new efficient algorithm for polynomial spline smoothing.

Some key words: Bayes estimate; Cross-validation; Generalized cross-validation; Leverage; Signal extraction; Spline smoothing; Studentized residual.

1. INTRODUCTION

Suppose observations are generated by a Gaussian signal plus noise process, with the signal described by a state space model. Such a model for the signal occurs often in practice, for example when the signal is the output of a stochastic difference or differential equation. This paper presents a new algorithm for signal extraction, that is, the computation of the conditional mean and variance of the signal given the observations. The usual method for signal extraction is to use the Kalman filter, see, for example, Anderson & Moore (1979, p. 105), followed by a smoothing step using an algorithm such as the fixed interval smoothing algorithm (Anderson & Moore, 1979, p. 187). Our approach also uses the Kalman filter, but for the smoothing step introduces a new set of recursions which are more efficient than those for the fixed interval smoothing algorithm. As for the fixed interval smoothing algorithm, recursions for estimating the signal and its variance are carried out separately, so that considerable additional savings can be made if the signal estimate alone is required. The development uses ideas introduced by Ansley & Kohn (1987a).

A remarkable property of the new recursions is that they can be applied immediately to provide new and efficient formulae for smoothing the entire state vector. For the state vector at a particular point in time, the basic recursions capture all information in future innovations of the process not already recovered by the Kalman filter.

If the stochastic process for the signal is viewed as a prior distribution for the signal, then signal extraction provides the posterior distribution of the signal conditional on the data and any parameters involved. If a prior is imposed on the unknown parameters, then our algorithm gives a fast way of computing the posterior distribution of the signal conditional on the data only. This is an important application of the algorithm because the integrations required to obtain the posterior from the prior and likelihood must

usually be done numerically, requiring computation of the conditional mean and variance of the signal at many different parameter values. A second important application where speed is crucial is in conducting simulation experiments to examine likelihood or Bayesian methods for signal estimation.

To illustrate the usefulness of our results we generalize the well-known ideas of studentized residuals and leverage from the linear regression context to the signal plus noise model, and show how both studentized residuals and points of high leverage may be determined efficiently using our algorithm. In particular, this allows us to investigate leverage in regression models with time series components. The discussion depends on the fundamental result that for any signal plus noise model an equivalent way to define leverage at a particular time point is as a multiple of the posterior variance of the signal. Such an interpretation is interesting even in the regression case. Because the observations are ordered, it is natural to consider also subsequences of points of high leverage, and our method extends easily to handle this problem. The theoretical discussion is illustrated with an example.

Finally, an important application of the new algorithm is to carry out cross-validation and generalized cross-validation efficiently in state space models. This again depends on the fundamental relationship between the conditional mean and the conditional variance of a signal plus noise model.

Our results provide a new way to compute efficiently polynomial smoothing splines of any order with the smoothness parameter estimated by either maximum likelihood, cross-validation or generalized cross-validation.

The paper is structured as follows. In § 2 we present the basic algorithm and extend it to handle the general smoothing problem, missing observations, diffuse initial conditions, and independent regressors. In § 3 we show how to apply the results of § 2 to compute studentized residuals and points of high leverage. An improved version of the Kohn & Ansley (1987a) spline smoothing algorithm is given in § 4.

2. A FAST ALGORITHM

2·1. *Standard case*

Consider the signal plus noise model

$$y(t) = f(t) + e(t), \tag{2·1}$$

where $f(t)$ is the scalar signal and $\{e(t)\}$ is a sequence of independent $N(0, \sigma^2)$ random variables. The signal is generated by

$$f(t) = h(t)'x(t), \quad x(t+1) = F(t)x(t) + u(t), \tag{2·2}$$

where $x(t)$ is the $q \times 1$ state vector, and $\{u(t)\}$ a sequence of $q \times 1$ independent $N(0, \sigma^2 Q(t)\}$ random variables. Assume that $x(0) \sim N(0, \sigma^2 S_0)$, with S_0 a given $q \times q$ matrix, and that $\{e(t)\}$ and $\{u(t)\}$ are independent of each other and of $x(0)$. Let $Y_t = \{y(1), \ldots, y(t)\}'$, so that Y_t is the vector of all observations available at time t, and for $1 \leqslant s, t \leqslant n$ define

$$\hat{f}(t) = E\{f(t) \mid Y_n\}, \quad S_f(t, s) = \sigma^{-2} \operatorname{cov}\{f(t), f(s) \mid Y_n\}. \tag{2·3}$$

Thus $\hat{f}(t)$ and $\sigma^2 S_f(t, t)$ are respectively the estimate of $f(t)$ and its conditional variance

given all n observations. Define

$$x(t|s) = E\{x(t)|Y_s\}, \quad S(t|s) = \sigma^{-2}\,\mathrm{var}\{x(t)|Y_s\}. \tag{2·4}$$

Note that $x(0|0) = 0$ and $S(0|0) = S_0$.

In this section we give a new algorithm for computing $\hat{f}(t)$ and $S_f(t, s)$ efficiently, in particular $S_f(t, t)$. The algorithm requires $x(t|t)$, $S(t|t)$, $x(t|t-1)$ and $S(t|t-1)$, for $t = 1, \ldots, n$, all of which can be obtained from the Kalman filter. Let

$$\varepsilon(t) = y(t) - E\{y(t)|Y_{t-1}\} = y(t) - h(t)'x(t|t-1) \tag{2·5}$$

be the tth innovation, with variance $\mathrm{var}\{\varepsilon(t)\} = \sigma^2 R(t)$, and define

$$K(t) = F(t)S(t|t-1)h(t)/R(t), \quad M(t) = F(t) - K(t)h(t)'. \tag{2·6}$$

Setting $\hat{e}(t) = E\{e(t)|Y_n\}$, note that

$$\hat{f}(t) = y(t) - \hat{e}(t), \quad \sigma^2 S_f(t, t) = \mathrm{var}\{f(t)|Y_n\} = \mathrm{var}\{e(t)|Y_n\}.$$

Our main result is given by Theorem 2·1. The proof is in the Appendix.

THEOREM 2·1. (i) *Define the sequence* $a(n), \ldots, a(1)$ *of* $q \times 1$ *random vectors by* $a(n) = 0$ *and*

$$a(t-1) = -h(t)\varepsilon(t)/R(t) + M(t)'a(t) \quad (t = n, \ldots, 2). \tag{2·7}$$

Then $\hat{e}(n) = \varepsilon(n)/R(n)$ *and, for* $t = n-1, \ldots, 1$,

$$\hat{e}(t) = \varepsilon(t)/R(t) + a(t)'K(t). \tag{2·8}$$

For $t = 1, \ldots, n$, $\hat{f}(t) = y(t) - \hat{e}(t)$.

(ii) *Let* $b(t) = \sigma^{-2}\,\mathrm{var}\{a(t)\}$. *Then* $b_n = 0$ *and*

$$b(t-1) = h(t)h(t)'/R(t) + M(t)'b(t)M(t) \quad (t = n, \ldots, 2). \tag{2·9}$$

Furthermore, $S_f(n, n) = 1 - 1/R(n)$ *and, for* $t = n-1, \ldots, 1$,

$$S_f(t, t) = 1 - 1/R(t) - K(t)'b(t)K(t). \tag{2·10}$$

For $s > 0$,

$$S_f(t, t+s) = -\frac{\varepsilon_t(t+s)}{R(t+s)} - K(t)'M(t+1)' \ldots M(t+s)'b(t+s)K(t+s), \tag{2·11}$$

where $\varepsilon_t(t+1) = -h(t+1)'K(t+1)$ *and for* $s > 1$

$$\varepsilon_t(t+s) = -K(t)'M(t+1)' \ldots M(t+s-1)'h(t+s). \tag{2·12}$$

In most applications the computed matrices $b(t)$, that is, including numerical error, will all be positive-semidefinite as required. However, it is straightforward to rewrite (2·9) in square root form as Morf & Kailath (1975) did, thus ensuring that each $b(t)$ is positive-semidefinite, and at the same time obtaining greater computational accuracy.

To illustrate the savings possible by using the new algorithm, Table 1 shows the number of multiplications and divisions required per step of the Kalman filter, the fixed interval smoothing algorithm and the new algorithm when the dimension of the state vector takes the values $1, \ldots, 6$ and 15. Thus if $q = 5$ and we are interested only in the signal estimate, we need only 7% of the work required by the fixed interval smoothing algorithm, whereas if we want both the signal estimate and conditional variance we need 70%. For $q = 15$ the corresponding percentages are 2% and 49%. As for the fixed interval smoothing algorithm, the new algorithm does not require the conditional variance to be computed in order to compute the signal estimate.

Table 1. *Multiplications and divisions per step; q, dimension*

q	Kalman filter	Fixed int. smoothing	Signal (2·7)-(2·8)	Variance (2·9)-(2·10)	Total (2·7)-(2·10)
1	4	11	4	11	15
2	38	37	9	40	49
3	91	119	16	97	113
4	179	274	25	191	216
5	311	525	36	331	367
6	496	895	49	526	575
15	6031	13375	256	6241	6497

Table 2. *Multiplications and divisions per step taking into account special structure of $F(t)$ and $h(t)$; q, dimension*

Model	q	Kalman filter	Fixed int. smoothing	Signal (2·7)-(2·8)	Variance (2·9)-(2·10)	Total (2·7)-(2·10)
1	3	12	83	8	13	21
2	5	47	385	14	52	66
3	15	271	9385	36	359	395

It is often possible to exploit special structure in the $F(t)$ matrices or $h(t)$ vectors to make additional computational savings. For example, consider the class of structural time series models discussed by Kitagawa & Gersch (1984) and Harvey & Todd (1983); such models have sparse $F(t)$ matrices with several elements identically 1. Table 2 gives operation counts, multiplications or divisions, for three such models. Model 1 has a quarterly seasonal component, but no trend component, so that the dimension of the state vector is $q=3$. Model 2 has a quarterly seasonal component and a second difference trend component so that $q=5$. Model 3 has a second-order autoregressive component, a monthly seasonal component and a second difference trend component so that $q=15$. Table 2 is similar to Table 1 with the operation counts taking into account the special structure of $F(t)$.

Our results are easily generalized to the case where $y(t)$ is a vector and $\mathrm{var}\{e(t)\}$ is a matrix which may vary with t. The derivation in the Appendix is virtually unchanged, and we omit details.

Because $h(t)$, $F(t)$ and $Q(t)$ are all functions of t, our results apply when there are missing observations. In this case we observe $y(t_i)$ for $i=1,\ldots,n$, where $t_1<\ldots<t_n$ and the t_i are not necessarily consecutive integers.

2·2. *Diffuse initial conditions*

In many applications, $x(0)$ is diffuse or partially diffuse. Instead of being a fixed matrix, $S_0=kS_0^{(1)}+S_0^{(0)}$, with $k>0$ and $S_0^{(1)}$ and $S_0^{(0)}$ independent of k, so that, as $k\to\infty$, $x(0)$ becomes diffuse. See Kitagawa & Gersch (1984) for an application to structural time series models, and Kohn & Ansley (1987) for an application to spline smoothing.

Now (Ansley & Kohn, 1985; Kohn & Ansley, 1986) for fixed $k>0$, (2·4)-(2·6) can be written

$$x(t\,|\,s;\,k)=x^{(0)}(t\,|\,s)+O(1/k),\quad S(t\,|\,s;\,k)=kS^{(1)}(t\,|\,s)+S^{(0)}(t\,|\,s)+O(1/k),$$
$$\varepsilon(t;\,k)=y(t)-h(t)'x^{(0)}(t\,|\,t-1)+O(1/k)=\varepsilon^{(0)}(t)+O(1/k),$$
$$R(t;\,k)=kR^{(1)}(t)+R^{(0)}(t)+O(1/k),$$

where each of the superscripted terms is independent of k. For $1 \leqslant t \leqslant n$ and $s = t$ or $s = t-1$, each such term can be obtained with the modified Kalman filter of Ansley & Kohn (1985) or Kohn & Ansley (1986). Moreover, if $R^{(1)}(t) = 0$ then $R^{(0)}(t) > 0$.

We now define the signal estimate and its conditional variance as

$$\hat{f}(t) = \lim_{k \to \infty} \hat{f}(t; k) = h(t)' x^{(0)}(t \mid n),$$

$$\sigma^2 S_f(t, t) = \lim_{k \to \infty} \sigma^2 S_f(t, t; k) = \sigma^2 h(t)' S^{(0)}(t \mid n) h(t).$$

Similarly, the associated conditional covariances are defined as $\sigma^2 S_f(t, s) = \lim \sigma^2 S_f(t, s; k)$, as $k \to \infty$.

Theorem 2·1 extends to the partially diffuse case by letting $k \to \infty$ in (2·6)–(2·12). Specifically, we replace $\varepsilon(t)$ by $\varepsilon^{(0)}(t)$ in (2·7) and (2·8), redefine $K(t)$ in (2·6) as

$$K(t) = \begin{cases} F(t) S^{(1)}(t \mid t-1) h(t) / R^{(1)}(t) & (R^{(1)}(t) > 0), \\ F(t) S^{(0)}(t \mid t-1) h(t) / R^{(0)}(t) & (R^{(0)}(t) > 0), \end{cases}$$

and replace $1/R(t)$ in (2·7)–(2·11) by 0 if $R^{(1)}(t) > 0$, and by $1/R^{(0)}(t)$ if $R^{(1)} = 0$. The last replacement means that, for values of t for which $R^{(1)}(t) > 0$, the first term in (2·7)–(2·9) and the second term in (2·10) is removed. Similarly, if $R^{(1)}(t+s) > 0$, the first term in (2·11) is removed.

2·3. *Additive regressors*

Consider the observation equation

$$y(t) = z(t)' \beta + f(t) + e(t),$$

with $f(t)$ generated by (2·2), $z(t)$ an $r \times 1$ vector of independent regressors and β an $r \times 1$ vector of regression coefficients. The total signal is now $z(t)' \beta + f(t)$, which can be estimated in one of two equivalent ways. The first way is to apply Theorem 2·1 and the results of Kohn & Ansley (1985). The second is to augment the state vector $x(t)$ defined in § 2·1 by redefining it as $\{x(t)', \beta'\}'$, and taking β diffuse. In the latter case the state transition equation becomes

$$\begin{bmatrix} x(t+1) \\ \beta \end{bmatrix} = \begin{bmatrix} F(t) & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} x(t) \\ \beta \end{bmatrix} + \begin{bmatrix} u(t) \\ 0 \end{bmatrix},$$

where $\beta \sim N(0, kI)$. The results of § 2·2 for the partially diffuse case can now be applied.

2·4. *More general smoothing results*

The recursions (2·7) and (2·9) and their derivations in the Appendix are of independent interest because they enable efficient computation of the conditional mean $E(g \mid Y_n)$ and conditional variance $\operatorname{var}(g \mid Y_n)$ for most vector random variables g of interest in state space models. Specifically, suppose g is a vector random variable which is independent of $\{e(s), s > t\}$ and $\{u(s), s \geqslant t\}$, and that $E(g \mid Y_t)$, $\operatorname{var}(g \mid Y_t)$ and $\Delta(t) = \sigma^{-2} \operatorname{cov}\{g, x(t) \mid Y_t\}$ are known. Then

$$E(g \mid Y_n) = E(g \mid Y_t) - \Delta(t) F(t)' a(t), \tag{2·13}$$

$$\operatorname{var}(g \mid Y_n) = \operatorname{var}(g \mid Y_t) - \sigma^2 \Delta(t) F(t)' b(t) F(t) \Delta(t)', \tag{2·14}$$

where the vectors $a(t)$ and matrices $b(t)$ are obtained from Theorem 2·1. We obtain (2·13) and (2·14) for the important case when $g = x(t)$, but the derivation for general g

is obtained similarly by filtering the augmented state vector $\{x(s)', g'\}'$ for $s \geqslant t$, as Ansley & Kohn (1987a).

From equations (2·10)–(2·12) of Ansley & Kohn (1987a)

$$x(t|j) = x(t|j-1) - S(t|t)F(t)'a_t(j) \quad (j > t),$$

where $a_t(j)$ is defined in (A·2) in the Appendix. Thus from (A·3),

$$x(t|n) = x(t|t) - S(t|t)F(t)'a(t), \tag{2·15}$$

which is a special case of (2·13). From (2·15)

$$\text{var}\{x(t|n)\} = \text{var}\{x(t|t)\} + \sigma^2 S(t|t)F(t)'b(t)F(t)S(t|t).$$

Because

$$\text{var}\{x(t)\} = \sigma^2 S(t|n) + \text{var}\{x(t|n)\} = \sigma^2 S(t|t) + \text{var}\{x(t|t)\},$$

we obtain

$$S(t|n) = S(t|t) - S(t|t)F(t)'b(t)F(t)S(t|t), \tag{2·16}$$

which is the appropriate version of (2·14).

Once we have the $a(t)$'s and $b(t)$'s from the recursions (2·7) and (2·9), we can obtain the conditional means and variances in (2·13)–(2·16) with very little additional computation. In particular, if $x_i(t)$ is the ith element of $x(t)$, we can readily obtain $E\{x_i(t)|Y_n\}$ and $\text{var}\{x_i(t)|Y_n\}$ by setting $g = x_i(t)$. Using (2·13) and (2·14), $E\{x_i(t)|Y_n\}$ can be computed with at most q^2+q operations in addition to those required to obtain the $a(t)$'s, where q is the dimension of the state vector. If in addition $\text{var}\{x_i(t)|Y_n\}$ is required, it can be computed with at most a further q^2+q operations given the $b(t)$'s. It is clear from Tables 1 and 2 in § 2·1 and Table 3 in § 4 below that this is a far more efficient way to smooth $x_i(t)$ than the fixed interval smoothing algorithm.

3. STUDENTIZED RESIDUALS AND LEVERAGE

3·1. *General*

Studentized residuals and leverage have been discussed extensively in the regression literature; see, for example, Cook & Weisberg (1982, p. 10). In this section we extend these ideas to signal plus noise models by means of a fundamental relationship between the conditional mean and conditional variance of the signal estimate, given in Lemma 3·1 below. This result also leads to efficient computation of both studentized residuals and measures of leverage via Theorem 2·1.

3·2. *Studentized residuals*

Consider first the regression model

$$y(t) = z(t)'\beta + e(t), \tag{3·1}$$

where $z(t)$ is a sequence of $r \times 1$ fixed regressors, β is a vector of unknown coefficients and $e(t)$ is a sequence of $N(0, \sigma^2)$ independent random variables. This model can be thought of as a signal plus noise model with signal $f(t) = z(t)'\beta$. Let

$$Y_n = \{y(1), \ldots, y(n)\}', \quad Z = \{z(1), \ldots, z(n)\}'.$$

If $\hat{\beta}$ is the least-squares estimate of β, then the estimate of the signal at time t is $\hat{f}(t) = z(t)'\hat{\beta}$. Writing $\hat{f} = \{\hat{f}(1), \ldots, \hat{f}(n)\}'$,

$$\hat{f} = AY_n, \tag{3·2}$$

where $A = Z(Z'Z)^{-1}Z'$ is called the hat matrix or influence matrix. The tth residual is $\hat{e}(t) = y(t) - \hat{f}(t)$, with variance $\sigma^2\{1 - A(t, t)\}$, where $A(t, t)$ is the tth diagonal element of A. To judge whether a particular residual is unusually large we consider its size relative to its standard deviation under the assumed model. One way of doing this is to construct the studentized residuals (Cook & Weisberg, 1982, p. 18),

$$\hat{e}(t)/[\hat{\sigma}\{1 - A(t, t)\}^{\frac{1}{2}}], \tag{3·3}$$

where $\hat{\sigma}$ is an estimate of σ. It is evident from (3·3) that, unless $A(t, t)$ is small, the ratio $\hat{e}(t)/\hat{\sigma}$ may be a misleading indicator of whether $\hat{e}(t)$ is unusually large.

The idea of studentized residuals can be generalized to any signal plus noise model in a straightforward manner. Writing $f = \{f(1), \ldots, f(n)\}'$, (3·2) holds for the signal plus noise model (2·1), with $A = \text{cov}\,(f, Y_n)\{\text{var}\,(Y_n)\}^{-1}$; for models with diffuse initial conditions, a definition in terms of limits is necessary. Define $\hat{e} = y - \hat{f}$ to be the $n \times 1$ vector of residuals and S_f to be the $n \times n$ matrix with stth element $S_f(t, s)$ given by (2·3), so that $\sigma^2 S_f = \text{var}\,(f \mid Y_n)$. The following lemma establishes a basic relationship between the matrix A and the variance matrix S_f.

LEMMA 3·1. *For the signal plus noise model* (2·1)

$$S_f = \sigma^{-2}\,\text{var}\,(f \mid Y_n) = A, \tag{3·4}$$

and in particular $S_f(t, t) = A(t, t)$. *Further,*

$$\text{var}\,(\hat{e}) = \sigma^2(I - A), \tag{3·5}$$

and in particular $\text{var}\,\{\hat{e}(t)\} = \sigma^2\{1 - A(t, t)\}$.

Proof. Write $\text{var}\,(f) = \sigma^2 V$, so that $\text{var}\,(Y_n) = \sigma^2(I + V)$ and $\text{cov}\,(f, Y_n) = \sigma^2 V$. Then by simple algebra

$$A = \text{cov}\,(f, Y_n)\{\text{var}\,(Y_n)\}^{-1} = V(I + V)^{-1}, \quad \text{var}\,(\hat{f}) = \sigma^2 V(I + V)^{-1}V.$$

Then

$$\text{var}\,(e \mid Y_n) = \text{var}\,(f \mid Y_n) = \text{var}\,(f) - \text{var}\,(\hat{f}) = \sigma^2\{V - V(I + V)^{-1}V\} = \sigma^2 A,$$

$$\text{var}\,(\hat{e}) = \text{var}\,(e) - \text{var}\,(e \mid Y_n) = \sigma^2(I - A). \qquad \square$$

Lemma 3·1 was obtained by Craven & Wahba (1979) for the special case of polynomial smoothing splines, using an algebraic proof.

We can now define the tth studentized residual for the signal plus noise model by (3·3). Computation of all the $A(t, t)$ would appear to require $O(n^2)$ operations in general; see, for example, the discussion of Kohn & Ansley (1987a, p. 43). However, it follows from Lemma 3·1 that $A(t, t) = S_f(t, t)$ so that all the $A(t, t)$ can be computed efficiently in $O(n)$ operations using Theorem 2·1.

The relationship between the regression model (3·1) and the signal plus noise model (2·1) can be made clear by noting that an alternative way of estimating the fitted values $\hat{f}(t) = z(t)'\hat{\beta}$ in the regression model (3·1) is to place a diffuse prior on β, that is, to take

$$\beta \sim N(0, kI_r) \tag{3·6}$$

with k large. More formally, for k fixed, let $\hat{f}(t; k) = E\{z(t)'\beta \mid Y_n; k\}$. Then

$$\hat{f}(t) = z(t)'\hat{\beta} = \lim_{k\to\infty} \hat{f}(t; k). \tag{3·7}$$

It is clear from § 2·3 how to write (3·1) in state space form.

3·3. *Leverage*

We now generalize to signal plus noise models the idea of leverage, which is in common use for regression models. See, for example, Cook & Weisberg (1982, p. 14). Consider first the regression model (3·1) with signal estimate $\hat{f} = Z\hat{\beta} = AY_n$ as in (3·2). Writing the t, jth element of A as $A(t, j)$ we have

$$\hat{f}(t) = A(t, t)y(t) + \textstyle\sum_{j\neq t} A(t, j)y(j), \tag{3·8}$$

which shows that the larger is $A(t, t)$ the larger is the influence of $y(t)$ on the size of $\hat{f}(t)$. The diagonal element $A(t, t)$ is thus said to represent the leverage of $y(t)$ on $\hat{f}(t)$.

Furthermore, as Cook & Weisberg (1982, p. 33) show,

$$y(t) - \hat{f}^{(-t)}(t) = \frac{y(t) - \hat{f}(t)}{1 - A(t, t)}, \tag{3·9}$$

where $\hat{f}^{(-t)}(t)$ is the estimate of $f(t)$ when the tth observation is omitted. This shows that the closer $A(t, t)$ is to 1, the greater is the error when $y(t)$ is estimated from the remainder of the data.

A Bayesian interpretation can be given to the $A(t, t)$ by noting that for the regression model $\sigma^2 A = \text{var}(f \mid Y_n)$, so that, from (3·7), under the prior (3·6)

$$\lim_{k\to\infty} \text{var}\{f(t) \mid Y_n\} = \sigma^2 A(t, t).$$

Hence, up to the scale factor σ^2, $A(t, t)$ is a measure of the posterior uncertainty about $f(t)$. This provides us with an alternative way of looking at leverage.

To generalize the idea of leverage to signal plus noise models we need only show that (3·8) and (3·9) hold for the model (2·1). It is clear from the discussion preceding Lemma 3·1 that (3·8) holds. If we let $Y_n^{(-t)}$ be Y_n with the tth observation $y(t)$ removed, then from (3·8)

$$\hat{f}^{(-t)}(t) = E\{\hat{f}(t) \mid Y_n^{(-t)}\} = \textstyle\sum_{j\neq t} A(t, j)y(j) + A(t, t)E\{y(t) \mid Y_n^{(-t)}\} \tag{3·10}$$

and (3·9) follows by simple algebra. As we have noted already, the $A(t, t)$'s can be computed efficiently by Theorem 2·1 in $O(n)$ operations. Thus we have an efficient way of computing leverage points in the signal plus noise model (2·1).

3·4. *An example*

Consider the data of Durbin & Watson (1951) on annual consumption of spirits. These data consist of 69 observations from 1870 to 1938 of log consumption of spirits per head (y), log real income per head (z_1) and log relative price of spirits (z_2). We used the first 60 observations for model fitting and reserved the last 9 for model validation. Three models were fitted, the first a pure linear regression with intercept, the second a linear regression with first-order autoregressive errors, and the third the random intercept model

$$y(t) = \alpha(t) + \beta_1 z_1(t) + \beta_2 z_2(t) + e(t), \quad \alpha(t) = \alpha(t-1) + \xi(t), \tag{3·11}$$

where $\{\xi(t)\}$ is an independent Gaussian sequence. Model (3·11) is a signal plus noise model which can be written in state space form as in § 2·3.

The residuals in the pure regression model are strongly correlated, with Durbin–Watson statistic 0·25. The fitted value of the autoregressive coefficient in the second model was very close to 1, suggesting a random walk error term. The third model proved best in terms of single- and multi-step predictions on the validation sample.

Figure 1 displays the values of $A(t, t)$ for both the first and third models. For the pure regression model, the $A(t, t)$'s are reasonably constant, but for the random intercept model there are peaks at each end of the data, and also around the years 1915–1919, observations 46–50. It was also found that studentized residuals for years 1909, observation 40, and 1915, observation 46, were large and negative, −3·9 and −3·3 respectively. The estimates of β_1 and β_2 for the third model using all 60 observations were $\hat{\beta}_1 = 0{\cdot}69$ (0·13) and $\hat{\beta}_2 = -0{\cdot}97$ (0·07), where the numbers in brackets are standard errors. When observations 46–50 were omitted, the corresponding estimates were $\hat{\beta}_1 = 0{\cdot}60$ (0·13) and $\hat{\beta}_2 = -0{\cdot}53$ (0·12), showing the parameter estimates to be sensitive to these observations. Deleting observation 40 did not change the estimates appreciably. The plot of $A(t, t)$ for the model (3·11) detected points 46–50 which may be influential, whereas the plot for the pure regression model detected no such points.

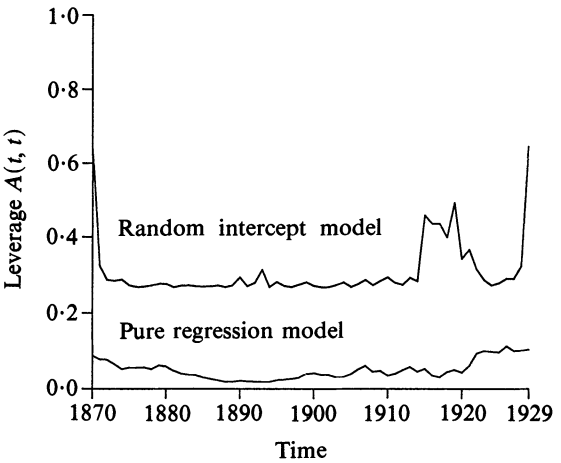

Fig. 1. Annual consumption of spirits: plot of $A(t, t)$ against time.

Observations 46–50 can be shown to be influential from another point of view also. If ∇ is the first differencing operator, so that $\nabla y(t) = y(t) - y(t-1)$, then we can rewrite (3·11) as

$$\nabla y(t) = \beta_1 \nabla z_1(t) + \beta_2 \nabla z_2(t) + u(t), \tag{3·12}$$

where $u(t) = \xi(t) + e(t) - e(t-1)$ is a first-order moving average process. Estimating β_1 and β_2 in (3·12) using ordinary least-squares gives estimates which are consistent but not efficient, as the autocorrelation in $u(t)$ is ignored. However, for this example the autocorrelation is small, and using all 60 observations we obtained parameter estimates similar to those obtained by maximum likelihood in (3·11). Further, points 45, 47 and 49 were found to be influential judged by the $A(t, t)$'s for model (3·12), confirming the results obtained for (3·11). In comparing the equivalent representations (3·11) and (3·12), note that it is more efficient to work with the structural form (3·11) when observations

are missing or deleted because of the difficulty in defining differences in such cases. Moreover, in many cases, ordinary least-squares applied to the differenced model may give significantly different parameter estimates due to autocorrelation in the $u(t)$'s, and influential observations may not be detected.

A further example is discussed by Ansley & Kohn (1987b).

3·5. *Groups of successive observations*

We now consider outlier detection and points of high leverage for groups of successive observations. Suppose each observation in a group of observations is an outlier or a point of high leverage relative to the rest of the data, but not relative to the group. Diagnostics that leave out one observation at a time are sometimes unable to detect that any of the observations in the group is unusual. This is well known for regression diagnostics (Cook & Weisberg, 1982, p. 135). Unfortunately, computing studentized residuals and influence measures for all subgroups of m $(m \geqslant 2)$ observations is very expensive, except in small samples. In time series, however, we have a natural ordering that enables us to limit the number of subgroups that need be considered.

Define $y_{t,m} = \{y(t), \ldots, y(t+m-1)\}'$, and $e_{t,m}$ and $f_{t,m}$ similarly. Let $\hat{f}_{t,m} = E(f_{t,m} \mid Y_n\}$ and $\hat{e}_{t,m} = y_{t,m} - \hat{f}_{t,m}$. Then $\text{var}\,(\hat{e}_{t,m}) = \sigma^2 C_{t,m} = \sigma^2(I_m - A_{t,m})$, where $A_{t,m}$ is the $m \times m$ submatrix of A with i,jth element $A(t+i-1, t+j-1)$, for $i, j = 1, \ldots, m$. We define the tth m-dimensional vector studentized residual as $C_{t,m}^{-1/2}\hat{e}_{t,m}/\hat{\sigma}$, and we can judge whether $\hat{e}_{t,m}$ is unusually large by comparing $\hat{e}'_{t,m}C_{t,m}^{-1}\hat{e}_{t,m}/\hat{\sigma}^2$ to a chi-squared distribution with m degrees of freedom.

Let

$$\eta_{t,m} = y_{t,m} - E\{f_{t,m} \mid y(1), \ldots, y(t-1), y(t+m), \ldots, y(n)\},$$

so that $\eta_{t,m}$ is the tth m-dimensional vector residual when $y(t), \ldots, y(t+m-1)$ are left out. Then, using an argument similar to (3·10), it can be shown that analogously to (3·9)

$$\eta_{t,m} = C_{t,m}^{-1}(y_{t,m} - \hat{f}_{t,m}), \tag{3·13}$$

and the size of $A_{t,m}$ is a measure of the leverage of $y_{t,m}$ in the same way that $A(t, t) = A_{t,1}$ is a measure of the leverage of the observation $y(t) = y_{t,1}$. Possible measures of the size of $A_{t,m}$ are $\det(A_{t,m})$ and the maximum eigenvalue of $A_{t,m}$.

For a given $m > 0$, the elements $S_f(t+i-1, t+j-1)$, for $i, j = 1, \ldots, m$ and $t = 1, \ldots, n$, and hence the matrices $A_{t,m}$ and $C_{t,m}$, can be computed in $O(n)$ operations using Theorem 2·1. Thus we have an efficient algorithm for the vector studentized residuals and leverage matrices.

Finally, we obtain some useful and interesting identities between the residuals $y(t) - \hat{f}^{(-t)}(t)$ obtained by leaving out one observation and the $\eta_{t,m}$ obtained by leaving out m successive observations. Write

$$\zeta(t) = y(t) - \hat{f}^{(-t)}(t) = y(t) - E\{y(t) \mid Y_n^{(-t)}\}, \quad \zeta_{t,m} = \{\zeta(t), \ldots, \zeta(t+m-1)\}'.$$

Let $D_{t,m}$ be the diagonal matrix whose diagonal elements are the same as those of $C_{t,m}$.

LEMMA 3·2. *We have*

$$\eta_{t,m} = C_{t,m}^{-1}D_{t,m}\zeta_{t,m}, \quad \text{var}\,(\eta_{t,m}) = \sigma^2 C_{t,m}, \quad \text{var}\,(\zeta_{t,m}) = \sigma^2 D_{t,m}^{-1}C_{t,m}D_{t,m}^{-1}.$$

Proof. From (3·9) and (3·13), $y_{t,m} - \hat{f}_{t,m} = C_{t,m}\eta_{t,m} = D_{t,m}\zeta_{t,m}$, and, from Lemma 3·1, $\text{var}\,(y_{t,m} - \hat{f}_{t,m}) = \sigma^2(I_m - A_{t,m})$. The lemma follows immediately. □

3·6. *Time-dependent noise variance*

So far we have assumed for simplicity that var $(e) = \sigma^2 I_n$. More generally, let var $(e) = \sigma^2\Lambda$, where Λ is a diagonal matrix with tth diagonal element $\lambda(t) > 0$. Writing $V = \text{var}(f)$, then $A = V(\Lambda + V)^{-1}$, and it is easy to check that (3·9) holds and (3·4) is replaced by

$$S_f = \sigma^{-2} \text{var}(f \mid Y_n) = A\Lambda$$

so that, in particular, $S_f(t, t) = \lambda(t)A(t, t)$. Thus the results in §§ 3·1 and 3·2 require little modification. It is straightforward to generalize the results in § 3·4 similarly. Moreover, Theorem 2·1 also generalizes in a simple way, so that the appropriately modified studentized residuals and measures of leverage can be computed efficiently with $O(n)$ operations.

4. Spline and partial spline smoothing

An important application of the results in § 2 is to spline and partial spline smoothing. We first discuss spline smoothing and consider the signal plus noise model

$$y(i) = f(t_i) + e(i), \tag{4·1}$$

where $0 \leqslant t_1 \leqslant \ldots \leqslant t_n \leqslant 1$. Suppose the signal $f(t)$ $(0 \leqslant t \leqslant 1)$ is estimated by minimizing the penalized least-squares loss function

$$\sum_{i=1}^{n} \{y(t_i) - g(t_i)\}^2 + \lambda^{-1} \int_0^1 \{g^{(m)}(t)\}^2 \, dt$$

over all functions $g(t)$ having square integrable mth derivative over the interval $[0, 1]$. The resulting estimate $\hat{f}(t)$ of $f(t)$ is called a polynomial smoothing spline of order $2m-1$. In particular, if $m = 2$ we obtain a cubic smoothing spline. The parameter λ is called the smoothness parameter and controls the trade-off between the least-squares fit and the smoothness penalty. See Wecker & Ansley (1983) and Silverman (1985) for examples of spline smoothing.

Wahba (1978) showed that, if $f(t)$ is assumed to follow an integrated Wiener process, then $\hat{f}(t)$ is the conditional mean of $f(t)$ given the data. Thus (4·1) can be thought of as a signal plus noise model and $\hat{f}(t)$ obtained by signal extraction. The integrated Wiener process is often regarded as a prior for the unknown function $f(t)$, so that $\hat{f}(t)$ is the posterior mean.

As Wecker & Ansley (1983) or Kohn & Ansley (1987) we can write the stochastic process for $y(t)$ in state space form and then use the modified Kalman filter of Kohn & Ansley (1986, 1987) together with Theorem 2·1 to compute $\hat{f}(t_i)$ and the posterior variance $\sigma^2 S_f(t_i, t_i)$ for any given λ.

Usually the smoothness parameter λ is unknown and has to be estimated from the data. Craven & Wahba (1979) suggested using generalized cross-validation to estimate λ and this has been popular with practitioners. The generalized cross-validation function is

$$V(\lambda) = \sum_{i=1}^{n} \{y(t_i) - \hat{f}(t_i)\}^2 / \{n - \text{tr}(A)\}^2, \tag{4·2}$$

where tr (A) is the trace of the influence matrix A. By Lemma 3·1, tr $(A) = \Sigma\, S_f(t_i, t_i)$ so that $V(\lambda)$ can be obtained efficiently using Theorem 2·1. Alternatively, $V(\lambda)$ could be obtained by using Theorem 2·1, part (i) to compute $\hat{f}(t_i)$ and the algorithm of Ansley &

Kohn (1987a) to obtain tr (A) directly. As discussed below, computing the $S_f(t_i, t_i)$ using part (ii) of Theorem 2·1 seems to be slightly faster.

Instead of using generalized cross-validation to estimate λ, Wahba & Wold (1975) used cross-validation, where the criterion function is

$$V^*(\lambda) = \sum_{i=1}^{n} \{y(t_i) - \hat{f}(t_i)\}^2 / \{n - A(i, i)\}^2. \tag{4·3}$$

Theorem 2·1 can be used to compute $V^*(\lambda)$ efficiently.

To demonstrate the savings obtained by using Theorem 2·1 in conjunction with the Kalman filter, Table 3 shows for polynomial splines with $m = 1, \ldots, 4$ the number of multiplications or divisions required for each step of the Kalman filter, the fixed interval smoothing algorithm and the new algorithm. If just the signal estimate is required, the column for fixed interval smoothing should be compared with that for the signal computed by (2·7)–(2·8); there are large savings. If both signal estimate and conditional variance are required, the fixed interval smoothing column should be compared with the total column for recursions (2·7)–(2·10); there are still considerable savings. Note that the diagonal elements $A(t, t)$ required to compute $V(\lambda)$ in (4·2) or $V^*(\lambda)$ in (4·3) are also provided by the fixed interval smoothing algorithm. The final column gives the operation count for the algorithm of Ansley & Kohn (1987a) for computing tr(A). This algorithm, together with the recursions (2·7)–(2·8) for the signal estimate, provides an alternative approach to computing $V(\lambda)$, as discussed above. The numbers in the final column are slightly greater than those for the variances obtained from (2·9)–(2·10), showing that the new algorithm is faster. Also, the algorithm of Ansley & Kohn (1987a) gives only tr(A), not the individual diagonal elements $A(t, t)$, and cannot therefore be used to compute $V^*(\lambda)$ in (4·3). Each of the procedures appearing in Table 3 requires the Kalman filter as an initial step. We emphasize that our algorithms for cross-validation and generalized cross-validation are $O(n)$, where n is the sample size.

Table 3. *Multiplications and divisions per step for spline smoothing; m, order*

m	Kalman filter	Fixed int. smoothing	Signal (2·7)–(2·8)	Variance (2·9)–(2·10)	Total (2·7)–(2·10)	Trace
1	4	3	1	2	3	3
2	18	27	3	14	17	14
3	42	92	5	36	41	38
4	82	184	10	61	71	80

Partial spline smothing is used for observations $y(i)$ which depend on $r+1$ regressors, r of which enter linearly and the $(r+1)$st nonlinearly, as follows:

$$y(i) = z(i)'\beta + f(t_i) + e(i). \tag{4·4}$$

In (4·4), $z(i)$ is a $r \times 1$ vector of independent regressors, and the $(r+1)$st regressor t_i enters nonlinearly through the function f. Partial spline estimates of the coefficient vector β and the unknown function $f(t)$ are obtained by minimizing

$$\sum_{i=1}^{n} \{y(i) - z(i)'\beta - g(t_i)\}^2 + \lambda^{-1} \int_0^1 \{g^{(m)}(t)\}^2 \, dt$$

over all functions $g(t)$ having square integrable mth derivative on $[0, 1]$. It is shown by Heckman (1986) and more generally by Kohn & Ansley (1988) that the partial spline

estimates are the posterior means obtained when a diffuse prior is placed on β and the same prior as in the spline model placed on $f(t)$. The resulting stochastic model is a signal plus noise model of the type considered in § 2·3, and Theorem 2·1, as modified in § 2·2, can be used to estimate both β and $f(t)$. If λ is unknown it can be estimated by generalized cross-validation as discussed by Ansley & Kohn (1987a). Examples of partial spline smoothing are given by Ansley & Wecker (1981).

The first $O(n)$ algorithm for obtaining tr (A) for generalized cross-validation in polynomial spline smoothing was given by Hutchinson & de Hoog (1985), using an approach based on Reinsch's (1967) algorithm for spline smoothing. A computer program implementing this method for cubic splines is given by Hutchinson (1986). Previously Silverman (1984) had given an $O(n)$ method for approximating tr (A).

The new algorithm in § 2 seems very competitive with that of Hutchinson & de Hoog (1985). For example, the new algorithm requires 35 operations per observation to compute $\hat{f}(t_i)$, $S_f(t_i, t_i)$ and tr (A) for the cubic spline problem, compared to over 40 for that of Hutchinson & de Hoog (1985), based on the code of Hutchinson (1986). Our algorithm will also be faster for quintic splines ($m=3$) and just a little slower for $m=4$.

The state space approach to computing splines and partial splines has a number of advantages over the Reinsch and Hutchinson & de Hoog method. First, the state space algorithm is the same for splines of any order, and also for the more general splines discussed by Kohn & Ansley (1987, 1988). It is essentially the same for partial splines also. The authors have implemented the new algorithm for both polynomial spline and partial spline smoothing of general order. The computer code is particularly simple, as is clear from (2·7)–(2·10). The Reinsch and Hutchinson & de Hoog method seems to have been implemented for cubic splines only, and not at all for partial splines. Secondly, the state space approach gives the posterior variance $S_f(t, t)$ at all points t, including $t \neq t_i$, and also the posterior mean and variance for each of the derivatives $f^{(j)}(t)$ $(j=1, \ldots, m-1)$. None of this is provided by the Reinsch and Hutchinson & de Hoog method. Thirdly, because the state space approach does not require divided differences, it seems to be numerically more stable than the Reinsch and Hutchinson & de Hoog method. Finally, the theoretical derivation of the spline algorithm is simple and direct, and fits into a unified signal plus noise framework.

Appendix

Proof of Theorem 2·1

We need the following preliminary results. Let LRL' be the Cholesky decomposition of σ^{-2} var (Y_n), where L is a lower triangular matrix with ones on the diagonal and R is a diagonal matrix. As given by Kohn & Ansley (1985), $\varepsilon = L^{-1}Y_n = \{\varepsilon(1), \ldots, \varepsilon(n)\}'$ is the vector of innovations, and the tth diagonal element of R is $R(t) = \sigma^{-2}$ var $\{\varepsilon(t)\}$. Write $\iota_t = (0, \ldots, 0, 1, 0, \ldots, 0)'$, where the 1 is in the tth position, and let $\varepsilon_t = \{\varepsilon_t(1), \ldots, \varepsilon_t(n)\}' = L^{-1}\iota_t$. Then

$$\hat{e}(t) = E\{e(t) \mid Y_n\} = \text{cov}\{e(t), Y_n\}\{\text{var}(Y_n)\}^{-1}Y_n$$

$$= \varepsilon(t)/R(t) + \sum_{s=t+1}^{n} \varepsilon_t(s)\varepsilon(s)/R(s). \qquad \text{(A·1)}$$

Let $x_t(s \mid j)$ be the value of $x(s \mid j)$ when $Y_n = \iota_t$. Lemma A·1 follows directly from Lemma 2·1 of Ansley & Kohn (1987a).

Lemma A·1. (i) *For* $1 \leqslant s < t \leqslant n$, $\varepsilon_t(s) = 0$ *and* $x_t(s+1 \mid s) = 0$.
(ii) *For* $1 \leqslant t \leqslant n$, $\varepsilon_t(t) = 1$ *and* $x_t(t+1 \mid t) = K(t)$.

(iii) *For* $1 \leq t < s \leq n$, $\varepsilon_t(s) = -h(s)'x_t(s|s-1)$ *and* $x_t(s+1|s) = M(s)x_t(s|s-1)$.

(iv) *From* (ii) *and* (iii) *above, for* $1 \leq t < n$, $\varepsilon_t(t+1) = -K(t)'h(t+1)$, *and, for* $1 < t+1 < s \leq n$, $\varepsilon_t(s) = -K(t)'M(t+1)' \ldots M(s-1)'h(s)$.

Proof of Theorem 2·1. (i) Let $a_t(t+1) = -h(t+1)\varepsilon(t+1)/R(t+1)$ and, for $s = t+2, \ldots, n$,

$$a_t(s) = -M(t+1)' \ldots M(s-1)'h(s)\varepsilon(s)/R(s). \tag{A·2}$$

Define $a(n) = 0$ and

$$a(t) = a_t(t+1) + \ldots + a_t(n) \quad (t < n). \tag{A·3}$$

The recursions (2·7) and (2·8) follow from Lemma A·1.

(ii) Because $b(t) = \sigma^{-2} \operatorname{var}\{a(t)\}$ the recursion (2·9) follows immediately from (2·8). Now

$$\operatorname{var}\{e(t)|Y_n\} = \operatorname{var}\{e(t)\} - \operatorname{var}\{\hat{e}(t)\},$$

and from (2·8)

$$\operatorname{var}\{\hat{e}(t)\} = 1/R(t) + K(t)'b(t)K(t),$$

giving (2·10).

For $s > 0$, from (A·1) we have

$$\hat{e}(t) = \sum_{j=t}^{t+s-1} \varepsilon_t(j)\varepsilon(j)/R(j) + \varepsilon_t(t+s)\varepsilon(t+s)/R(t+s) + \sum_{j=t+s+1}^{n} \varepsilon_t(j)\varepsilon(j)/R(j),$$

and from (A·2), (A·3) and Lemma A·1

$$\sum_{j=t+s+1}^{n} \varepsilon_t(j)\varepsilon(j)/R(j) = K(t)'M(t+1)' \ldots M(t+s)'a(t+s)$$

so that

$$\operatorname{cov}\{\hat{e}(t), \hat{e}(t+s)\} = \varepsilon_t(t+s)/R(t+s) + K(t)'M(t+1)' \ldots M(t+s)'b(t+s)K(t+s),$$

and (2·11) follows because

$$\sigma^2 S_f(t, s) = \operatorname{cov}\{e(t), e(t+s)|Y_n\} = \operatorname{cov}\{e(t), e(t+s)\} - \operatorname{cov}\{\hat{e}(t), \hat{e}(t+s)\}.$$

Finally, from Lemma A·1, $\varepsilon_t(t+1) = -h(t+1)'K(t)$ and $\varepsilon_t(t+s)$ is given by (2·12). □

REFERENCES

ANDERSON, B. D. O. & MOORE, J. B. (1979). *Optimal Filtering.* Englewood Cliffs, New Jersey: Prentice-Hall.

ANSLEY, C. F. & KOHN, R. (1985). Estimation, filtering and smoothing in state space models with incompletely specified initial conditions. *Ann. Statist.* **13**, 1286–316.

ANSLEY, C. F. & KOHN, R. (1987a). Efficient generalized cross-validation for state space models. *Biometrika* **74**, 139–48.

ANSLEY, C. F. & KOHN, R. (1987b). A fast algorithm for smoothing, cross validation and influence in state space models. *Proc. Bus. Econ. Statist. Sect., Am. Statist. Assoc.*, pp. 106–13.

ANSLEY, C. F. & WECKER, W. E. (1981). Extensions and examples of the signal extraction approach to regression. In *Proc. ASA-CENSUS-NBER Conf. Applied Time Series Analysis of Economic Data*, pp. 181–92. Washington, D.C.: U.S. Bureau of the Census.

COOK, R. D. & WEISBERG, S. (1982). *Residuals and Influence in Regression.* New York: Chapman and Hall.

CRAVEN, P. & WAHBA, G. (1979). Smoothing noisy data with spline functions. *Numer. Math.* **31**, 377–403.

DURBIN, J. & WATSON, G. S. (1951). Testing for serial correlation in least squares regression II. *Biometrika* **38**, 159–78.

HARVEY, A. C. & TODD, P. H. J. (1983). Forecasting economic time series with structural and Box–Jenkins models: a case study (with discussion). *J. Bus. Econ. Statist.* **1**, 299–315.

HECKMAN, N. E. (1985). Spline smoothing in a partly linear model. *J.R. Statist. Soc.* B **48**, 244–8.

HUTCHINSON, M. G. (1986). A fast procedure for calculating minimum cross-validation cubic smoothing splines. *Assoc. Comput. Mach. Trans. Math. Soft.* **12**, 150–3.

HUTCHINSON, M. G. & DE HOOG, F. R. (1985). Smoothing data with spline functions. *Numer. Math.* **47**, 99–106.

KITAGAWA, G. & GERSCH, W. (1984). A smoothness priors-state space approach to time series with trend and seasonalities. *J. Am. Statist. Assoc.* **79**, 378–89.

KOHN, R. & ANSLEY, C. F. (1985). Efficient estimation and prediction in time series regression models. *Biometrika* **72**, 694–7.

KOHN, R. & ANSLEY, C. F. (1986). Estimation, prediction and interpolation for ARIMA models with missing data. *J. Am. Statist. Assoc.* **81**, 751–61.

KOHN, R. & ANSLEY, C. F. (1987). A new algorithm for spline smoothing based on smoothing a stochastic process. *SIAM J. Sci. Statist. Comput.* **8**, 33–48.

KOHN, R. & ANSLEY, C. F. (1988). The equivalence between Bayesian smoothness priors and optimal smoothing for function estimation. In *Bayesian Analysis of Time Series and Dynamic Models.* Ed. J. C. Spall, pp. 393–430. New York: Marcel Dekker.

MORF, M. & KAILATH, T. (1975). Square root algorithms for least squares estimation. *IEEE Trans. Aut. Cont.*, AC-**20**, 487–97.

REINSCH, C. H. (1967). Smoothing by spline functions. *Numer. Math.* **10**, 177–83.

SILVERMAN, B. W. (1984). A fast and efficient cross-validation method for smoothing parameter choice in spline regression. *J. Am. Statist. Assoc.* **79**, 584–9.

SILVERMAN, B. W. (1985). Some aspects of the spline smoothing approach to nonparametric curve fitting (with discussion). *J.R. Statist. Soc.* B **47**, 1–52.

WAHBA, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *J.R. Statist. Soc.* B **40**, 364–72.

WAHBA, G. & WOLD, S. (1975). A completely automatic French curve: Fitting spline functions by cross-validation. *Comm. Statist.* **4**, 1–17.

WECKER, W. E. & ANSLEY, C. F. (1983). The signal extraction approach to nonlinear regression and spline smoothing. *J. Am. Statist. Assoc.* **78**, 81–9.

[*Received December* 1987. *Revised June* 1988]