The Annals of Statistics
1991, Vol. 19, No. 2, 1073–1083

THE DIFFUSE KALMAN FILTER[1]

BY PIET DE JONG

University of British Columbia

The Kalman recursion for state space models is extended to allow for likelihood evaluation and minimum mean square estimation given states with an arbitrarily large covariance matrix. The extension is computationally minor. Application is made to likelihood evaluation, state estimation, prediction and smoothing.

1. Introduction. This paper deals with likelihood evaluation and minimum mean square error prediction given observations generated by a state space model with a diffuse initial state. A state is said to be diffuse if its covariance matrix is arbitrarily large. Diffuse initial states arise in the context of parameter uncertainty and model nonstationarity as illustrated later.

To deal with a diffuse initial state in the state space model, Schweppe (1973) and Harvey and Phillips (1979) propose initiating the Kalman filter with a very large covariance matrix. This poses numerical problems and does not answer the question of the existence of diffuse constructs. A variant of the Kalman filter, called the information filter, has also been proposed. However, as Ansley and Kohn [(1985), page 1298] point out, the information filter breaks down in many important cases and can be numerically inefficient. Harvey and Pierse (1984) and Harvey (1990) propose initiating the Kalman filter with regression type estimates based on an initial stretch of the data. These estimates may be cumbersome to construct and questions of existence are not dealt with. Pole and West (1989) deal with diffuse initial states in a Bayesian setting. In a sequence of papers, Ansley and Kohn (1985), Kohn and Ansley (1986, 1987a, 1987b) develop, discuss and advocate the use of a modified form of the Kalman filter to allow for diffuse states. This paper presents alternate modifications with the following merits:

1. *Computational*. The modified filter is a computationally trivial extension of the ordinary Kalman filter, turning the two existing vector recursions into matrix recursions and the addition of a matrix recursion. No extra matrix inversions are required.
2. *Analytic*. Proofs are general and direct. Necessary and sufficient conditions for the existence of diffuse constructs are in terms of extended filter quantities.

Received August 1987; revised March 1990.

[1] Research supported by a grant from ZWO, the Dutch agency for the support of pure scientific research and the National Science and Engineering Research Council of Canada.

AMS 1980 *subject classifications*. Primary 62M15; secondary 62M20, 60G35,

Key words and phrases. State space, Kalman filter, smoothing, diffuse, nonstationarity, likelihood.

Special cases of the methods used in this paper have been alluded to in Rosenberg (1973), Wecker and Ansley (1983) and Kohn and Ansley (1985, 1987a), as discussed later.

The program of this paper is as follows. The next section introduces the state space model (SSM), examples and the Kalman filter (KF). Section 3 discusses likelihood evaluation and introduces the diffuse Kalman filter (DKF). Section 4 deals with the diffuse likelihood. Section 5 goes on to consider diffuse prediction. Diffuse smoothing is dealt with in Section 6.

2. The state space model and the Kalman filter. Throughout this article capital letters denote nonrandom matrices. Lower case letters denote column vectors. The notation $\gamma \sim (c, \sigma^2 C)$ indicates γ is a random vector with expectation $E(\gamma) = c$ and covariance matrix $\mathrm{Cov}(\gamma) = \sigma^2 C$. Expectations and covariances are always unconditional.

If A and B have identical column dimension, then $(A; B) \equiv (A', B')'$. Thus $(y_1; y_2; \dots; y_n)$ is the stack of the column vectors $y_1, y_2, \dots, y_n$. The notation $y^{\#}$ indicates the number of components in y while A^- denotes the Moore–Penrose generalized inverse of A.

DEFINITION 2.1. The state space model (SSM). Random vector $y = (y_1; y_2; \dots; y_n)$ is said to be generated by a state space model, denoted $y \leftarrow$ SSM, if for $1 \le t \le n$, $y_t = X_t\beta + Z_t\alpha_t + G_t u_t$, where for $0 \le t \le n$, $\alpha_{t+1} = W_t\beta + T_t\alpha_t + H_t u_t$ and

(i) $(u_0; u_1; \dots; u_n) \sim (0, \sigma^2 I)$ with $\sigma^2 > 0$,

(ii) $\alpha_0 = 0$ and $\beta = b + B\gamma$, where $\gamma \sim (c, \sigma^2 C)$, b is fixed and B has full column rank,

(iii) γ and $(u_0; u_1; \dots; u_n)$ are uncorrelated,

(iv) C is nonsingular unless $C = 0$; $\mathrm{Cov}(y)$ is nonsingular if $C = 0$.

This article deals with state space methods when γ is diffuse. Random vector γ is said to be diffuse, denoted $C \to \infty$, if C^{-1} converges to a zero matrix in the Euclidean norm. Diffuse random vectors arise in two ways. First, to model parameter uncertainty and second, diffuseness arises when a nonstationary model is assumed to have applied since time immemorial. The next example illustrates matters.

EXAMPLE 2.1. Suppose for scalar y_t, $(y_{t+1} - x'_{t+1}\delta) = a(y_t - x'_t\delta) + u_t$, where the $u_t \sim (0, \sigma^2)$ are serially uncorrelated. Special cases of this model are the regression model ($a = 0$), the autoregressive model of order 1 ($x_t = 1$), the random walk ($\delta = 0$, $a = 1$), white noise ($\delta = 0$, $a = 0$), the regression model with random walk disturbances ($a = 1$) and the random walk with drift ($a = 1$, $x_{t+1} - x_t = 1$). A SSM formulation is to take for $t \ge 1$, $Z_t = 1$, $G_t = 0$, $W_t = 0$, $T_t = a$ and $H_t = 1$. Table 1 indicates alternative specifications for the remaining quantities. For the last three cases reported in Table 1, γ is identified with, respectively, δ, $y_0 - x_0'\delta$ and $(\delta; y_0 - x_0'\delta)$. The entries $1/\sqrt{1-a^2}$ in the column for H_0 are derived assuming the model has applied since time immemorial. This is a standard assumption in practical time series modelling leading to parameter parsimony since no new parameters have to be introduced to model initial conditions.

THEOREM 2.1. *If $y \leftarrow$ SSM, then $y = X\beta + \varepsilon$, where*

$$X = [X_1 + Z_1W_0; X_2 + Z_2(W_1 + T_1W_0); \dots;$$
$$X_n + Z_n\{W_{n-1} + T_{n-1}W_{n-2} + \cdots + (T_{n-1} \cdots T_1)W_0\}]$$

and $\varepsilon \sim (0, \sigma^2\Sigma)$ with Σ nonsingular and $\mathrm{Cov}(\beta, \varepsilon) = 0$.

The proof is direct. In terms of this notation, $\mathrm{Cov}(y) = \sigma^2\{(XB)C(XB)' + \Sigma\}$ and in particular, $\sigma^2\Sigma$ is the covariance matrix of y assuming $C = 0$, or in other words, the residual covariance matrix of y given γ. Note that Σ has a complicated but specialized structure as induced by the SSM.

Well-known constructs and results used in this article include the following. Suppose x is a random vector. Then a predictor of x using y is defined as $a + Ay$, where a and A are chosen such that the diagonal entries of $\mathrm{Cov}(x - a - Ay)$ are minimum. If $\hat{x}$ is the predictor of x using y, then define $\mathrm{mse}(\hat{x}) \equiv \mathrm{Cov}(x - \hat{x})$.

The Kalman filter (KF) computes predictors in the context of the SSM when $C = 0$. In particular the KF is the recursion

$$\begin{aligned} e_t &= y_t - X_t\beta - Z_ta_t, & D_t &= Z_tP_tZ_t' + G_tG_t', \\ K_t &= (T_tP_tZ_t' + H_tG_t')D_t^{-1}, & a_{t+1} &= W_t\beta + T_ta_t + K_te_t, \\ P_{t+1} &= (T_t - K_tZ_t)P_tT_t' + (H_t - K_tG_t)H_t', \end{aligned} \tag{2.1}$$

with starting conditions $a_1 = W_0\beta$, $P_1 = H_0H_0'$. Here a_t is the predictor of α_t using $(y_1; y_2; \dots; y_{t-1})$ and $\mathrm{mse}(a_t) = \sigma^2P_t$. Also e_t is the error of predicting y_t using $(y_1; y_2; \dots; y_{t-1})$, $E(e_t) = 0$, $\mathrm{Cov}(e_t) = \sigma^2D_t$ and for $t \neq s$, $\mathrm{Cov}(e_t, e_s) = 0$. It is assumed the D_t are nonsingular, a condition guaranteed by Definition 2.1 (iv); proofs are in Anderson and Moore (1979).

3. Evaluation of the likelihood with the diffuse Kalman filter. Suppose $y \leftarrow$ SSM, γ is fixed ($C = 0$) and y is normally distributed. From Theorem 2.1, $y = X(b + B\gamma) + \varepsilon$ and hence $\mathrm{Cov}(y) = \mathrm{Cov}(\varepsilon) = \sigma^2\Sigma$. The log of the likelihood based on y is, apart from a constant,

$$\begin{aligned} \lambda(y) &= -\tfrac{1}{2}\{\ln|\sigma^2\Sigma| + (y - X\beta)'\Sigma^{-1}(y - X\beta)/\sigma^2\} \\ &= -\tfrac{1}{2}\{y^{\#}\ln(\sigma^2) + \ln|\Sigma| + (q - 2s'\gamma + \gamma'S\gamma)/\sigma^2\}, \end{aligned}$$

where

$$S = (XB)'\Sigma^{-1}(XB), \qquad s = (XB)'\Sigma^{-1}(y - Xb),$$
$$q = (y - Xb)'\Sigma^{-1}(y - Xb).$$

The maximum likelihood estimators (mle's) of γ and σ^2 are, respectively,

$$\hat{\gamma} = S^{-1}s, \qquad \hat{\sigma}^2 = (q - s'S^{-1}s)/y^{\#},$$

where it is assumed that S is nonsingular. Substituting $\hat{\gamma}$ and $\hat{\sigma}^2$ back into $\lambda(y)$ yields the (γ, σ^2)-maximized log-likelihood $-\frac{1}{2}[y^{\#}\{\ln(\hat{\sigma}^2)\} + \ln|\Sigma|]$.

The expressions for S, s, q and $|\Sigma|$ as indicated before are not computationally practical. However, a viable approach to calculating these quantities is to employ the KF (2.1). This method was first proposed by Schweppe (1965) and extended upon by Rosenberg (1973). The next development indicates further extensions and is stated in terms of the following notation.

Definition 3.1. The diffuse Kalman filter (DKF). The DKF is the KF (2.1) with the equations for e_t and a_{t+1}, respectively, replaced by

$$E_t = (X_t B, y_t - X_t b) - Z_t A_t, \qquad A_{t+1} = W_t(-B, b) + T_t A_t + K_t E_t,$$

with starting condition $A_1 = W_0(-B, b)$. Also the following recursion is added: $Q_{t+1} = Q_t + E_t' D_t^{-1} E_t$, where $Q_1 = 0$.

Theorem 3.1. *Suppose $y \leftarrow SSM$, γ is fixed, y is normally distributed and the DKF is applied. Then $Q_{n+1} = \{(S, s); (s', q)\}$. If S is nonsingular, then the mle's of γ and σ^2 are $\hat{\gamma} = S^{-1}s$ and $\hat{\sigma}^2 = (q - s'S^{-1}s)/y^{\#}$, respectively. The (γ, σ^2)-maximized log-likelihood is $-\frac{1}{2}[y^{\#}\{\ln(\hat{\sigma}^2)\} + \sum_{t=1}^{n} \ln|D_t|]$.*

Proof. Suppose $e = (e_1; e_2; \ldots; e_n)$, where the e_t are as defined in (2.1). Then for some matrices K and L, $e = Ky - L\beta$. In particular, K is zero above the main diagonal and has all ones on the diagonal implying $|K| = 1$. Since $E(e) = 0$, it follows that $KX\beta = L\beta$ for all β and hence $KX = L$ and $e = K(y - X\beta)$. Furthermore, $\mathrm{Cov}(e)$ is block-diagonal with blocks $D_1, D_2, \ldots, D_n$ as given in (2.1) and hence $K\Sigma K' = D = \mathrm{diag}(D_1, D_2, \ldots, D_n)$, $\Sigma^{-1} = K'D^{-1}K$ and

$$\ln|\Sigma| = \ln|D_1| + \ln|D_2| + \cdots + \ln|D_n|,$$
$$S = (KXB)'D^{-1}(KXB), \qquad s = (KXB)'D^{-1}K(y - Xb),$$
$$q = \{K(y - Xb)\}'D^{-1}K(y - Xb).$$

Now suppose β in (2.1) is replaced by b to yield $f = (f_1; f_2; \ldots; f_n)$ instead of e. Then $f = K(y - Xb)$. If $F = KXB$, then the columns of F can be computed in the same way as f, except that y is replaced by zero and b by the corresponding columns of $-B$. In terms of this notation, $S = F'D^{-1}F$, $s = F'D^{-1}f$ and $q = f'D^{-1}f$. It is clear that $E_t = (F_t, f_t)$, where $(F_1; F_2; \ldots; F_n) = F$. Hence Q_{n+1} is as asserted. □

The DKF can thus be used for likelihood evaluation when $C = 0$. Further uses are outlined later. These are similar to the uses of the Ansley and Kohn [(1985), page 1297] algorithm. However, the DKF is simple and efficient compared to the Ansley and Kohn (1985) algorithm: no factorizations are

TABLE 1
Alternative initial conditions

Alternative	X_t	W_0	b	B	c	C	H_0
δ known, $\|a\| < 1$	x_t'	0	δ	I	0	0	$1/\sqrt{1-a^2}$
δ diffuse, $\|a\| < 1$	x_t'	0	0	I	0	∞	$1/\sqrt{1-a^2}$
δ known, $\|a\| \geq 1$	$(x_t', 0)$	$(0, a)$	$(\delta; 0)$	$(0; 1)$	0	∞	1
δ diffuse, $\|a\| \geq 1$	$(x_t', 0)$	$(0, a)$	$(0; 0)$	I	0	∞	1

required and attendant proofs are short, direct and more general [e.g., Assumption 2.5(i) of Ansley and Kohn (1985) is not needed]. Rosenberg [(1973), page 410] has suggested a special case of the DKF for likelihood maximization when $C = 0$. Wecker and Ansley (1983) and Kohn and Ansley (1985) also make Rosenberg's (1973) suggestion. A square root version of the DKF is discussed in de Jong (1990).

EXAMPLE 2.1 (continued). Suppose $a = 0$ and suppose δ is regarded as diffuse. Then Q_t accumulates the squares and cross products and $Q_{n+1} = (X, y)'(X, y)$. Theorem 3.1 in this case specializes to the usual regression results. Alternatively, supposed $|a| \geq 1$ and the specification of the last row of Table 1. Detailed calculations show for $t \geq 1$, $E_t = (x_t' - ax_{t-1}', 0, y_t - ay_{t-1})$, $D_t = 1$ except that $E_1 = (x_1', a, y_1)$. Thus the symmetric matrix Q is

$$Q = \begin{bmatrix} x_1x_1' + \sum_{t=2}^{n} (x_t - ax_{t-1})(x_t - ax_{t-1})' & \cdot & \cdot \\ ax_1' & a^2 & \cdot \\ y_1x_1' + \sum_{t=2}^{n} (y_t - ay_{t-1})(x_t - ax_{t-1})' & ay_1 & y_1^2 + \sum_{t=2}^{n} (y_t - ay_{t-1})^2 \end{bmatrix}$$

and the mle of δ is

$$\hat{\delta} = \left\{ \sum_{t=2}^{n} (x_t - ax_{t-1})(x_t - ax_{t-1})' \right\}^{-1} \left\{ \sum_{t=2}^{n} (x_t - ax_{t-1})(y_t - ay_{t-1}) \right\},$$

while the mle of $y_0 - x_0'\delta$ is $(y_1 - x_1'\hat{\delta})/a$. Further, the mle of σ^2 is

$$\hat{\sigma}^2 = n^{-1} \sum_{t=2}^{n} \left\{ (y_t - ay_{t-1}) - (x_t - ax_{t-1})'\hat{\delta} \right\}^2$$

and the (δ, σ^2)-maximized log-likelihood is $-\frac{1}{2}n\{\ln(\hat{\sigma}^2)\}$. In the special case of a random walk with drift ($a = 1, x_{t+1} - x_t = 1$), the mle's of δ, $y_0 - x_0'\delta$ and

σ^2 reduce to

$$\frac{y_n - y_1}{n-1}, \quad \frac{ny_1 - y_n}{n-1}, \quad \frac{1}{n}\left\{\sum_{t=2}^{n}(y_t - y_{t-1})^2 - \frac{(y_n - y_1)^2}{n-1}\right\}.$$

4. The diffuse likelihood. The next result generalizes Theorem 3.1 to the case where γ is random. The result sets the stage for a consideration of the diffuse likelihood.

THEOREM 4.1. *Suppose $y \leftarrow$ SSM and $(\gamma; y)$ is normally distributed. Further suppose the DKF is applied and S is nonsingular. Then the mle of c is $\hat{\gamma} = S^{-1}s$, with covariance matrix $\sigma^2 S^{-1}$. The mle of σ^2 is $\hat{\sigma}^2 = (q - s'S^{-1}s)/y^{\#}$, while the mle of C is zero. The log-likelihood $\lambda(y)$ maximized with respect to c, σ^2 and C is* $-\frac{1}{2}[y^{\#}\{\ln(\hat{\sigma}^2)\} + \sum_{t=1}^{n}\ln|D_t|]$.

PROOF. Write, for example, $\lambda(y|\gamma)$ as the conditional log-likelihood of y given γ. Then $\gamma(y) = \lambda(\gamma) + \lambda(y|\gamma) - \lambda(\gamma|y)$ and $-2\lambda(y)$ thus equals

$$\begin{aligned}
&\ln|\sigma^2 C| + (\gamma - c)'C^{-1}(\gamma - c)/\sigma^2 + \ln|\sigma^2\Sigma| \\
&\quad + (y - Xb - XB\gamma)'\Sigma^{-1}(y - Xb - XB\gamma)/\sigma^2 \\
&\quad - \ln\left|\sigma^2(C^{-1} + S)^{-1}\right| - \left\{\gamma - (C^{-1} + S)^{-1}(s + C^{-1}c)\right\}' \\
&\qquad\qquad \times (C^{-1} + S)\left\{\gamma - (C^{-1} + S)^{-1}(s + C^{-1}c)\right\}/\sigma^2.
\end{aligned}$$

Expanding the various terms and simplifying, shows $\lambda(y)$ equals

$$\begin{aligned}
&-\frac{1}{2}\Big[\ln|C| + \ln|C^{-1} + S| + \ln|\sigma^2\Sigma| \\
(4.1)\qquad &\quad + \left\{q + c'C^{-1}c - (C^{-1}c + s)'(S + C^{-1})^{-1}(C^{-1}c + s)\right\}/\sigma^2\Big].
\end{aligned}$$

Maximizing with respect to c yields the first assertion of the theorem. Substituting $c = S^{-1}s$ into (4.1) shows that the c-maximized log-likelihood is $-\frac{1}{2}\{\ln|I + CS| + \ln|\sigma^2\Sigma| + (q - s'S^{-1}s)/\sigma^2\}$ and hence the mle's of σ^2 and C are as asserted. Substituting the mle's of σ^2 and C into c-maximized log-likelihood yields the (c, σ^2, C)-maximized log-likelihood as asserted. □

Thus $\hat{\gamma} = S^{-1}s$ is the mle of c for every σ^2 and C, while the mle's of σ^2 and C are maximizers in the c-maximized likelihood. Furthermore, as $C \to \infty$, both $\hat{\gamma}$ and $\hat{\sigma}^2$ stay fixed. The next theorem considers the behaviour of $\lambda(y)$ as $C \to \infty$. The theorem expands a result of de Jong (1988) and shows the role of the DKF.

THEOREM 4.2. *Suppose $y \leftarrow$ SSM and $(\gamma; y)$ is normally distributed. Further, suppose the DKF is applied and S is nonsingular. Then as $C \to \infty$, $\lambda(y) + \frac{1}{2}\ln|C|$ converges to*

$$-\frac{1}{2}\left[y^{\#}\{\ln(\sigma^2)\} + \ln|S| + \sum_{t=1}^{n}\ln|D_t| + (q - s'S^{-1}s)/\sigma^2\right]. \tag{4.2}$$

Moreover (4.2) *is a log-likelihood based on Ny, where N has rank $y^{\#} - \gamma^{\#}$,* $\operatorname{Cov}(Ny, \gamma) = 0$ *and* $\ln|\operatorname{Cov}(Ny)|$ *equals the first three terms in* (4.2). *The log-likelihood* (4.2) *maximized with respect to σ^2 equals*

$$-\frac{1}{2}\left[y^{\#}\{\ln(\hat\sigma^2)\} + \ln|S| + \sum_{t=1}^{n}\ln|D_t|\right]. \tag{4.3}$$

PROOF. Consider $\lambda(y)$ as given in (4.1). Add $\frac{1}{2}\ln|C|$ and let $C \to \infty$ to show $\lambda(y) + \frac{1}{2}\ln|C|$ converges to (4.2).

For the second part of the theorem, let N be any matrix of full row rank such that the row space of N coincides with the row space of M, where $M = I - (XB)\{(XB)'\Sigma^{-1}(XB)\}^{-1}(XB)'\Sigma^{-1}$. Then $NXB = 0$, $E(Ny) = NXb$, $\operatorname{Cov}(Ny) = \sigma^2 N\Sigma N'$ and

$$\begin{aligned} q - s'S^{-1}s &= (y - Xb)'\Sigma^{-1}M(y - Xb) = (y - Xb)'M'\Sigma^{-1}M(y - Xb) \\ &= \{M(y - Xb)\}'M'\Sigma^{-1}M\{M(y - Xb)\} \\ &= \{N(y - Xb)\}'(N\Sigma N')^{-1}\{N(y - Xb)\}. \end{aligned}$$

Thus the likelihood based on Ny agrees with (4.2) if $\ln|\operatorname{Cov}(Ny)|$ is as specified. Expression (4.3) is arrived at by substituting $\hat\sigma^2$ for σ^2 in (4.2). □

Theorem 4.2 implies that (4.2) is a proper log-likelihood, called the diffuse log-likelihood. Since $\operatorname{Cov}(Ny, \gamma) = 0$, the diffuse log-likelihood is a likelihood based on those aspects of y invariant to γ. Note that the diffuse log-likelihood differs from the (c, C, σ^2)-maximized log-likelihood only with respect to the term $\frac{1}{2}\ln|S|$.

EXAMPLE 2.1 (continued). Assume the specification given in the last line of Table 1. Then $|S| = a^2|\Sigma_{t=2}^{n}(x_t - ax_{t-1})(x_t - ax_{t-1})'|$ and the diffuse log-likelihood maximized with respect to σ^2 is

$$n\{\ln(\hat\sigma^2)\} + a^2\left|\sum_{t=2}^{n}(x_t - ax_{t-1})(x_t - ax_{t-1})'\right|.$$

5. Diffuse prediction. The DKF can be used to compute diffuse predictors of α_t and y_t in the context of the SSM. These are predictors constructed under the assumption that $y \leftarrow$ SSM with $C \to \infty$. For $1 \le t \le n + 1$, let $\hat\gamma_t$, $\hat\alpha_t$ and $\hat y_t$ denote the predictors of γ, α_t and y_t using $(y_1; y_2; \ldots; y_{t-1})$. Note that if $C = 0$, then $\hat\gamma_t = c$, $\operatorname{mse}(\hat\gamma_t) = 0$, $\hat\alpha_t = a_t$ and $\hat y_t = X_t\beta + Z_t a_t$, where a_t is as given in the KF (2.1). The case where $C \neq 0$ is treated in the next theorem which uses the notation $Q_t \equiv \{(S_t, s_t); (s_t', q_t)\}$, where q_t is scalar.

THEOREM 5.1. *Suppose $y \leftarrow SSM$, where $C \neq 0$ and the DKF is applied. Let $A_{t\gamma}$ and $E_{t\gamma}$ denote all but the last columns of A_t and E_t. Then*

$$\begin{aligned} \hat{\gamma}_t &= \left(S_t + C^{-1}\right)^{-1}\left(C^{-1}c + s_t\right), & \operatorname{mse}(\hat{\gamma}_t) &= \sigma^2(S_t + C)^{-1}, \\ \hat{\alpha}_t &= A_t(-\hat{\gamma}_t; 1), & \operatorname{mse}(\hat{\alpha}_t) &= \sigma^2 P_t + A_{t\gamma} \operatorname{mse}(\hat{\gamma}_t) A_{t\gamma}', \\ y_t - \hat{y}_t &= E_t(-\hat{\gamma}_t; 1), & \operatorname{mse}(\hat{y}_t) &= \sigma^2 D_t + E_{t\gamma} \operatorname{mse}(\hat{\gamma}_t) E_{t\gamma}'. \end{aligned}$$

PROOF. Without loss of generality, assume $t = n + 1$. The predictor of γ not using y is c with mse matrix $\sigma^2 C$. By Theorem 2.1, $y = Xb + XB\gamma + \varepsilon$, where $\varepsilon \sim (0, \sigma^2\Sigma)$. Using a well-known result, the predictor of γ using y is thus

$$\begin{aligned} &\left\{(XB)'\Sigma^{-1}(XB) + C^{-1}\right\}^{-1}\left\{C^{-1}c + (XB)'\Sigma^{-1}(y - Xb)\right\} \\ &\quad = \left(S + C^{-1}\right)^{-1}\left(C^{-1}c + s\right), \end{aligned}$$

with mse matrix $\sigma^2(S + C^{-1})^{-1}$. Hence $\hat{\gamma}_t$ and $\operatorname{mse}(\hat{\gamma}_t)$ are as asserted.

Now consider predicting α_t using $(y_1; y_2; \dots; y_{t-1})$. This is equivalent to first predicting α_t using $(\gamma; y_1; y_2; \dots; y_{t-1})$, and then predicting this predictor using $(y_1; y_2; \dots; y_{t-1})$. The first mentioned predictor is a_t given in (2.1). However from the DKF,

$$\begin{aligned} A_{t+1}(-\gamma; 1) &= \left[W_t(-B, b) + T_t A_t + K_t\{(X_t B, y_t - X_t b) - Z_t A_t\}\right](-\gamma; 1) \\ &= W_t\beta + T_t A_t(-\gamma; 1) + K_t\{y_t - X_t\beta - A_t(-\gamma; 1)\}, \end{aligned}$$

with starting condition $A_1(-\gamma; 1) = W_0\beta$. Thus $A_t(-\gamma; 1)$ satisfies the same recursion and starting condition as a_t and hence $a_t = A_t(-\gamma; 1)$. Thus $\hat{\alpha}_t = A_t(-\hat{\gamma}_t; 1)$ as required. The formula for $\operatorname{mse}(\hat{\alpha}_t)$ follows from

$$\begin{aligned} \operatorname{mse}(\hat{\alpha}_t) &= \operatorname{Cov}(\alpha_t - a_t + a_t - \hat{\alpha}_t) = \operatorname{Cov}(\alpha_t - a_t) + \operatorname{Cov}(a_t - \hat{\alpha}_t) \\ &= \sigma^2 P_t + \operatorname{Cov}\{A_t(\hat{\gamma}_t - \gamma; 0)\}, \end{aligned}$$

where the second equality results from the fact that $\alpha_t - a_t$ is a prediction error using the random vector $(\gamma; y_1; y_2, \dots, y_{t-1})$ and both a_t and $\hat{\alpha}_t$ are based on this random vector. The expressions for $\hat{y}_t$ and $\operatorname{mse}(\hat{y}_t)$ are proved similarly. □

Now suppose in the formula for $\hat{\gamma}_t$, c is replaced by its mle $S_t^{-1}s_t$. Then $\hat{\gamma}_t$ reduces to $S_t^{-1}s_t$. This evidently is also the limit of $\hat{\gamma}_t$ as $C \to \infty$. Thus the diffuse predictor of γ based on $(y_1; y_2, \dots, y_{t-1})$ coincides with the predictor of γ replacing c by its mle. Similar statements apply to $\hat{\alpha}_t$ and $\hat{y}_t$. Thus replacing c by its mle is tantamount to treating γ as random with an arbitarily large covariance matrix. A formal statement is contained in Theorem 5.2, the proof of which follows directly from Theorem 5.1.

THEOREM 5.2. *Suppose $y \leftarrow$ SSM and $1 \le t \le n + 1$. If S_t is nonsingular, then as $C \to \infty$, $\hat{\gamma}_t \to S_t^{-1}s_t$ and* $\operatorname{mse}(\hat{\gamma}_t) \to \sigma^2 S_t^{-1}$. *If the rows of $A_{t\gamma}$ are in the row space of S_t, then as $C \to \infty$, $\hat{\alpha}_t \to A_t(-S_t^- s_t; 1)$ and* $\operatorname{mse}(\hat{\alpha}_t) \to \sigma^2(P_t + A_{t\gamma} S_t^- A'_{t\gamma})$. *If the rows of $E_{t\gamma}$ are in the row space of S_t, then as $C \to \infty$, $\hat{y}_t \to \{X_t(-B, b) + Z_t A_t\}(-S_t^- s_t; 1)$ and* $\operatorname{mse}(\hat{y}_t) \to \sigma^2(D_t + E_{t\gamma} S_t^- E'_{t\gamma})$.

A special case of this result is displayed in Rosenberg (1973). Kohn and Ansley [(1987a), page 45], in the context of spline smoothing, show that their modified filter for computing limiting predictors yields the same results as the approach taken in Wecker and Ansley (1983) which is that of Rosenberg (1973). Kohn and Ansley [(1987a), page 45] go on to conclude that the Rosenberg (1973) approach is numerically inefficient. A comparison of the DKF and the Ansley and Kohn [(1985), page 1297] filter shows that this is not so.

Like the Ansley and Kohn filter, the DKF can be collapsed to the ordinary Kalman filter after a few iterations. This combines the ideas of this section and Section 4. Initially, suppose γ corresponds to purely initial conditions (as in the third row of Table 1) and hence for $1 \le t \le n$, $X_t\beta = 0$ and $W_t\beta = 0$. Suppose m is the first integer such that S_m is nonsingular. Consider

$$\lambda(y) + \tfrac{1}{2}\ln|C| = \{\lambda(y_1, y_2, \dots, y_{m-1}) + \tfrac{1}{2}\ln|C|\}$$
$$+ \lambda(y_m, \dots, y_n | y_1, \dots, y_{m-1}),$$

where $\lambda(y_m, \dots, y_n | y_1, \dots, y_{m-1})$ is the conditional log-likelihood based on $(y_m; \dots; y_n)$ conditioning on $(y_1; \dots; y_{m-1})$. As $C \to \infty$, the term in curly brackets converges to

$$-\frac{1}{2}\left[(y_1; \dots; y_{m-1})^{\#} \ln(\sigma^2) + \ln|S_m| + \sum_{t=1}^{m-1} \ln|D_t| + \left(q_m - s'_m S_m^{-1} s_m\right)/\sigma^2\right],$$

while $\lambda(y_m, \dots, y_n | y_1, \dots, y_{m-1})$ converges to the log-likelihood based on $(y_m; \dots; y_n)$ generated by a SSM with initial conditions

$$E(\alpha_m) = A_m\left(-S_m^{-1}s_m; 1\right), \qquad \operatorname{Cov}(\alpha_m) = \sigma^2\left(P_m + A_{m\gamma} S_m^{-1} A'_{m\gamma}\right).$$

Thus, if KF is initialized at $t = m$ with $E(\alpha_m)$ and $\operatorname{Cov}(\alpha_m)$ as given, then as $C \to \infty$,

$$\lambda(y) + \tfrac{1}{2}\ln|C| \to -\frac{1}{2}\left[y^{\#} \ln(\sigma^2) + \ln|S_m| + \sum_{t=1}^{n} \ln|D_t| + q_n/\sigma^2\right],$$

where $q_n = q_m - s'_m S_m^{-1} s_m + \Sigma_{t=m}^{n} e'_t D_t^{-1} e_t$. The KF as initialized also directly yields the limits of $\hat{\alpha}_t$ and $\operatorname{mse}(\hat{\alpha}_t)$, $m < t \le n$ as $C \to \infty$. If $X_t\beta \ne 0$ or $W_t\beta \ne 0$ for some $1 \le t \le n$, then the state vector for $m \le t \le n$ is taken to be $(\alpha_t; \gamma)$ and the DKF can be collapsed to the KF based on the SSM employing the augmented state vector.

6. Diffuse smoothing. Smoothing refers to predicting the state α_t, $1 \leq t \leq n$, using the entire observation vector y where $y \to$ SSM. Smoothing can be based on the recursion

$$(6.1) \qquad N_{t-1} = Z_t' D_t^{-1} E_t + L_t' N_t, \qquad R_{t-1} = Z_t' D_t^{-1} Z_t + L_t' R_t L_t,$$

where $N_n = 0$, $R_n = 0$ and for $1 \leq t \leq n$, $L_t = T_t - K_t Z_t$, and all other quantities are defined as in the DKF.

THEOREM 6.1. *Suppose $y \leftarrow$ SSM, where $C \neq 0$ and the DKF is applied followed by the recursion (6.1). For $1 \leq t \leq n = 1$, suppose $\tilde{\alpha}_t$ denotes the predictor of α_t using y. Then*

$$\tilde{\alpha}_t = (A_t + P_t N_{t-1})(-\hat{\gamma}_{n+1}; 1),$$

$$\operatorname{mse}(\tilde{\alpha}_t) = \sigma^2(P_t - P_t R_{t-1} P_t) + N_{t-1,\gamma} \operatorname{mse}(\hat{\gamma}_{n+1}) N_{t-1,\gamma}',$$

where $N_{t-1,\gamma}$ denotes all but the last column of $A_t + P_t N_{t-1}$ and $\hat{\gamma}_{n=1}$ and $\operatorname{mse}(\hat{\gamma}_{n+1})$ *are as given in Theorem 5.1 with $t = n + 1$. Furthermore, as $C \to \infty$ and provided $N_{t-1,\gamma}$ is in the row space of S_{n+1}, $\tilde{\alpha}_t$ and* $\operatorname{mse}(\tilde{\alpha}_t)$ *converge to the previous expressions with $\hat{\gamma}_{n+1}$ replaced by $S_{n+1}^- s_{n+1}$ and* $\operatorname{mse}(\hat{\gamma}_{n+1})$ *replaced by $\sigma^2 S_{n+1}^-$.*

PROOF. From de Jong (1989), if $r_{t-1} = Z_t' D_t^{-1} e_t + K_t' r_t$ with $r_n = 0$ and e_t as in (2.1), then the predictor of α_t and associated mse matrix using $(\gamma; y)$ are respectively, $a_t + P_t r_{t-1}$ and $\sigma^2(P_t - P_t R_{t-1} P_t)$. Now

$$N_{t\text{-}1}(-\gamma; 1) = \left(Z_t' D_t^{-1} E_t + L_t' N_t\right)(-\gamma; 1) = Z_t' D_t^{-1} e_t + L_t' N_t(-\gamma; 1),$$

with starting condition $N_n(-\gamma; 1) = 0$. Thus $N_{t-1}(-\gamma; 1) = r_{t-1}$ and the predictor of α_t using $(\gamma; y)$ is $(A_t + P_t N_{t-1})(-\gamma; 1)$, which implies $\tilde{\alpha}_t$ and $\operatorname{mse}(\tilde{\alpha}_t)$ are as asserted. The limiting expressions as $C \to \infty$ are arrived at by letting $C \to \infty$ in the expressions for $\hat{\gamma}_{n+1}$ and $\operatorname{mse}(\hat{\gamma}_{n+1})$. □

Using the results in de Jong (1989), expressions analogous to those in Theorem 6.1 can also be derived for predictors of the signal $X_t\beta + Z_t\alpha_t$ using y or $X_t\beta + Z_t\alpha_t$ and α_t using $(y_1; \dots; y_{t-1}; y_{t+1}; \dots; y_n)$. Furthermore, the results can be generalized to derive cross covariances of the form, for example, $\operatorname{Cov}(\alpha_t - \tilde{\alpha}_t, \alpha_s - \tilde{\alpha}_s)$, $t \neq s$ and fixed point and fixed lag smoothing algorithms for the case $C \to \infty$.

EXAMPLE 2.1 (continued). Consider the specification as in the last row of Table 1. Then $N_{t-1} = E_t$ and $R_{t-1} = 1$. Thus as $C \to \infty$, $\tilde{\alpha}_t \to y_t - x_t'\hat{\delta}$, where $\hat{\delta} = S_{n+1}^{-1} s_{n+1}$.

Acknowledgments. I am indebted to Andrew Harvey and anonymous referees for helpful comments.

REFERENCES

ANDERSON, B. D. O. and MOORE, J. B. (1979). *Optimal Filtering*. Prentice-Hall, Englewood Cliffs, N.J.

ANSLEY, C. F. and KOHN, R. (1985). Estimation, filtering and smoothing in state space models with incompletely specified initial conditions. *Ann. Statist.* **13** 1286–1316.

DE JONG, P. (1988). The likelihood for a state space model. *Biometrika* **75** 165–169.

DE JONG, P. (1989). Smoothing and interpolation with the state space model. *J. Amer. Statist. Assoc.* **84** 1085–1088.

DE JONG, P. (1991). Stable algorithms for the state space model. *J. Time Ser. Anal.* To appear.

HARVEY, A. C. (1990). *Forecasting, structural time series models, and the Kalman filter*. Cambridge Univ. Press.

HARVEY, A. C. and PHILLIPS, G. D. A. (1979). Maximum likelihood estimation of regression models with autoregressive-moving average disturbances. *Biometrika* **66** 49–58.

HARVEY, A. C. and PIERSE, R. G. (1984). Estimating missing observations in economic time series. *J. Amer. Statist. Assoc.* **79** 125–131.

KOHN, R. and ANSLEY C. F. (1985). Efficient estimation and prediction in time series regression models. *Biometrika* **72** 694–697.

KOHN, R. and ANSLEY C. F. (1986). Estimation, prediction, and interpolation for ARIMA models with missing data. *J. Amer. Statist. Assoc.* **81** 751–761.

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

KOHN, R. and ANSLEY C. F. (1987b). Signal extraction for finite nonstationary time series. *Biometrika* **74** 411–421.

POLE, A. and WEST M. (1989). Reference analysis for the DLM. *J. Time Ser. Anal.* **10** 131–147.

ROSENBERG, B. (1973). The analysis of a cross section of time series by stochastically convergent parameter regression. *Ann. Social Economic Measurement* **2** 399–428.

SCHWEPPE, F. C. (1965). Evaluation of likelihood functions for Gaussian signals. *IEEE Trans. Inform. Theory* **11** 61–70.

SCHWEPPE, F. C. (1973). *Uncertain Dynamic Systems*. Prentice-Hall, Englewood Cliffs, N.J.

WECKER, W. and ANSLEY, C. F. (1983). The signal extraction approach to nonlinear regression and spline smoothing. *J. Amer. Statist. Assoc.* **78** 81–89.

FACULTY OF COMMERCE AND BUSINESS ADMINISTRATION
UNIVERSITY OF BRITISH COLUMBIA
VANCOUVER, B.C. V6T 1Y8
CANADA