Biometrika (1988), 75, 1, *pp.* 165-9
Printed in Great Britain

The likelihood for a state space model

BY PIET DE JONG

Faculty of Commerce and Business Administration, University of British Columbia, Vancouver, British Columbia, Canada V6T 1Y8

SUMMARY

This paper derives an expression for the likelihood for a state space model. The expression can be evaluated with the Kalman filter initialized at a starting state estimate of zero and associated estimation error covariance matrix of zero. Adjustment for initial conditions can be made after filtering. Accordingly, initial conditions can be modelled without filtering implications. In particular initial conditions can be modelled as 'diffuse'. The connection between the 'diffuse' and concentrated likelihood is also displayed.

Some key words: Kalman filtering; Maximum likelihood; State space model; Time series.

1. INTRODUCTION

A state space model asserts that observation vectors y_t $(t = 1, \ldots, n)$ are generated by

$$y_t = F_t x_t + u_t, \quad x_t = H_t x_{t-1} + G_t v_t, \tag{1}$$

where the u_t and v_t are zero mean, serially and mutually uncorrelated random vectors with covariance matrices U_t and V_t $(t = 1, \ldots, n)$.

In applications, the F_t, H_t, G_t, U_t and/or V_t contain unknown parameters to be estimated on the basis of observed data $y = (y_1, \ldots, y_n)$. This paper deals with maximum likelihood estimation. Minus twice the log likelihood, apart from an additive and proportionality constant is

$$l(y) = l(y_1) + l(y_2|y_1) + \ldots + l(y_n|y_{n-1}, \ldots, y_1), \tag{2}$$

where the vertical bar denotes conditioning. This is the 'prediction error decomposition' of the likelihood.

Throughout this paper it is assumed that all disturbances u_t and v_t in (1), as well as the initial state x_0 are normally distributed. The likelihood contributions in (2) then only depend on the conditional means and covariance matrices and the Kalman filter can be used for calculating these quantities, given the mean vector μ and covariance matrix C of the initial state x_0. This is the Kalman filter method of evaluating the likelihood due to Schweppe (1965). The method has been propounded by Akaike (1978), Harvey & Phillips (1979), Jones (1980), Gardner, Harvey & Phillips (1980) and Harvey (1981).

To use the Kalman filter method for the evaluation of the likelihood one has to specify the initial conditions μ and C. If (1) embodies an autoregressive moving average, ARMA, model then μ and C follow from the ARMA parameters; see, for example, Gardner et al. (1980). In more general situations μ and C enter the likelihood independently from other parameters. If C is set to zero then the initial state is regarded as fixed but unknown. Rosenberg (1973) has shown that if $C = 0$ then the maximum likelihood estimator of μ can be explicitly displayed and concentrated out of the likelihood.

In many situations, for example if the model embodied in (1) is nonstationary, it is natural to consider the likelihood (2) as C becomes large. A number of suggestions have been made to compute (2) under this scenario. The ordinary Kalman filter can be initiated with an arbitrarily chosen large C; see, for example, Harvey & Phillips (1979). Alternatively, Schweppe (1973) has

suggested the use of an alternative form of the Kalman filter, involving C^{-1}, and C large is accomodated by setting C^{-1} small or to zero. Ansley & Kohn (1985) point out, however, that the filtering implications of having C large or C^{-1} zero may be undesirable. Accordingly they develop and suggest the application of substantially modified Kalman filter equations. The covariance matrix C enters these modified equations in such a way that it is safe to consider, and numerically compute, the limit of filtered quantities as C tends to infinity.

The present paper presents an expression for the likelihood that is easily evaluated with the ordinary Kalman filter. However the filter is invariably initiated with a starting state estimate of zero and associated estimation error covariance matrix of zero. Hence insofar as filtering is concerned, no decisions about starting conditions need to be made. The possibility that μ and C are fully or partially tied to other unknown parameters is not ruled out. Quantities calculated with the filter are accumulated. After running through all the data $y_1, \ldots, y_n$, a final adjustment is made to allow for the uncertainty associated with x_0. Since the filtering involves neither μ nor C, both may be specified arbitrarily without filtering implications. The limit likelihood as C tends to infinity can hence be straightforwardly computed with the ordinary Kalman filter leading to a likelihood which is invariant to x_0.

There are two further sections to this paper. The next section presents the main result and discusses its implications. Section 3 gives the proof.

2. THE LIKELIHOOD

To conserve space write KF(μ, C) for the Kalman filter applied with starting state estimate μ and associated estimation error covariance matrix C.

THEOREM. *Suppose $y_1, \ldots, y_n$ are generated by a model of the form* (1) *where all disturbances u_t and v_t $(t = 1, \ldots, n)$ and the initial state x_0 are normally distributed. Suppose x_0 has unknown mean μ and unknown nonsingular covariance matrix C, and that, for each $t = 1, \ldots, n$, y_t cannot be perfectly predicted given $y_{t-1}, \ldots, y_1$ and x_0. Then minus twice the log likelihood is, apart from constants,*

$$l(y_1, \ldots, y_n) = \log|C| + \mu' C^{-1}\mu + \sum_{t=1}^{n} \log|D_t| + \sum_{t=1}^{n} e_t' D_t^{-1} e_t$$
$$+ \log|C^{-1} + S| - (C^{-1}\mu + s)'(C^{-1} + S)^{-1}(C^{-1}\mu + s). \qquad (3)$$

Here the e_t and D_t $(t = 1, \ldots, n)$ are the usual innovations and innovation covariance matrices calculated with KF(0, 0). *The vector s and matrix S are calculated in parallel with the e_t and D_t, for $t = 1, \ldots, n$, as follows*

$$s = s + Z_{t-1}' F_t' D_t^{-1} e_t, \quad S = S + Z_{t-1}' F_t' D_t^{-1} F_t Z_{t-1}, \quad Z_t = H_{t+1}(I - K_t F_t) Z_{t-1},$$

with s and S initialized at 0, $Z_0 = I$, and K_t is the Kalman gain matrix from KF(0, 0).

Before proceeding to the proof of this theorem the following points are worth noting. Here and below, e is the vector of all the e_t, X is the matrix with row blocks $F_t Z_{t-1}$, and D is the block diagonal matrix with diagonal blocks D_t $(t = 1, \ldots, n)$. In this notation $s = X'D^{-1}e$ and $S = X'D^{-1}X$.

(i) The unknown parameters μ and C manifest themselves in the likelihood exactly as displayed: they do not occur in e_t, D_t, s or S.

(ii) It follows from the proof below, that $(C^{-1} + S)^{-1}(C^{-1}\mu + s)$ and $(C^{-1} + S)^{-1}$ are respectively the conditional mean vector and covariance matrix of x_0 given y. Hence the equations for s and S serve to perform the fixed point smoothing algorithm (Anderson & Moore, 1979, p. 170) applied to the fixed point $t = 0$.

(iii) On rearrangement of terms the limit of $l(y)$ as C tends to zero is

$$\log|D|+(e-X\mu)'D^{-1}(e-X\mu)=\log|D|+e'D^{-1}e+\mu'S\mu-2s'\mu.$$

This is the likelihood considering $x_0=\mu$ as fixed but unknown.

(iv) The assumption that C is nonsingular is not restrictive. If C is singular such that, for some vector α, $\alpha'C\alpha=0$ and $\alpha'\mu=0$, then x_0 contains redundant information and model (1) can be reorganized such that C is nonsingular by deleting appropriate components of x_0. If C is singular such that for some vector α, $\alpha'C\alpha=0$ but $\alpha'\mu\neq 0$ then x_0 contains 'fixed effects'. This situation is handled as in (iii) above.

(v) The condition that the y_t cannot be perfectly predicted given previous observations and x_0 ensures that the D_t are nonsingular. This assumption is satisfied in most practical situations. If y_t is perfectly predictable, but this perfect predictability can be avoided by excluding predictor random variables other than those contained in x_0, then y_t contains information already observed, and this redundant information in y_t can be ignored.

(vi) The limit of $l(y)-\log|C|$ as C tends to infinity in such a way that C^{-1} tends to 0 is

$$\log|D|+e'D^{-1}e+\log|S|-s'S^{-1}s=\log|D|+\log|S|+e'D^{-1}\{I-X(X'D^{-1}X)^{-1}X'D^{-1}\}e. \quad (4)$$

This is the likelihood based on a linear transformation of y making the data invariant to x_0. This likelihood can also be computed with KF(0, 0); compare with Ansley & Kohn (1985, p. 1286). This is possible even though the invariant random vectors on which this likelihood is based, contained in

$$\{I-X(X'D^{-1}X)^{-1}X'D^{-1}\}e=e-XS^{-1}s,$$

can generally be evaluated only with a double pass through the data: first to evaluate the e_t, S and s, and secondly to calculate $e_t-X_tS^{-1}s$ $(t=1,\ldots,n)$.

(vii) The likelihood (3) can be concentrated with respect to μ and C. Differentiating with respect to μ and equating to zero yields $\hat{\mu}=S^{-1}s$ which is independent of C. Substituting this expression back into $l(y)$ yields the concentrated likelihood

$$\log|C|+\log|D|+\log|C^{-1}+S|+e'D^{-1}e-s'S^{-1}s=\log|D|+\log|I+CS|+e'D^{-1}e-s'S^{-1}s.$$

This expression decreases as C tends to zero with a minimum which differs from (4) to the term $\log|S|$. Thus the limit likelihood (4) minus $\log|S|$ is the concentrated likelihood. This concentrated likelihood corresponds to both the cases where $C=0$, and hence x_0 is regarded as fixed but unknown, or concentrated with respect to both μ and C.

(viii) It can be shown that under the conditions of the theorem, S is nonsingular if and only if the matrix with row blocks $F_tH_t,\ldots H_1$ $(t=1,\ldots n)$ is of full column rank. This is an identifiability or observability condition regarding x_0.

3. Proof of the theorem

The following argument is a refinement of a method used by Newbold (1974) for ARMA models coupled with Rosenberg's (1973) observation that the innovations are linear in the starting state estimate.

By assumption, neither the marginal distribution of x_0 nor the conditional distribution of y given x_0 is degenerate. Hence the joint distribution of y and x_0 is nondegenerate and

$$l(y)=l(x_0)+l(y_1|x_0)+\ldots+l(y_n|y_{n-1},\ldots,y_1,x_0)-l(x_0|y).$$

All random vectors are assumed normal. Hence $l(y)$ reduces to

$$\log|C|+(x_0-\mu)'C^{-1}(x_0-\mu)+\sum_t\log|D_t|+\sum_t e_t'(x_0)D_t^{-1}e_t(x_0)-l(x_0|y),$$

where the $e_t(x_0)$ are the conditional random vectors, i.e. innovations, and the D_t are the associated conditional covariance matrices. The argument x_0 of $e_t(x_0)$ emphasizes that the innovations are conditioned with respect x_0.

The innovations $e_t(x_0)$ and associated covariance matrices D_t can be built up with KF$(x_0, 0)$. Hence neither the $e_t(x_0)$ nor D_t depend on C. This is because they are conditional on the first state x_0. Also the D_t do not depend on x_0. However, the innovation vectors $e_t(x_0)$ functionally depend on x_0 as the notation indicates.

The innovations $e_t(x_0)$ are linear in x_0 (Rosenberg, 1973). Hence for fixed matrices X_t, not depending on x_0 or C, $e_t(x_0) = e_t - X_t x_0$ where the $e_t \equiv e_t(0)$ $(t = 1, \ldots, n)$ are the innovations constructed using a starting state estimate of 0. Let e be the vector with vector components e_t and X the matrix with row blocks X_t $(t = 1, \ldots, n)$. Then $l(y)$ can be written as

$$\log|C| + \log|D| + (x_0 - \mu)'C^{-1}(x_0 - \mu) + (e - Xx_0)'D^{-1}(e - Xx_0) - l(x_0|y)$$
$$= \log|C| + \log|D| + \begin{bmatrix} \mu - x_0 \\ e - Xx_0 \end{bmatrix}' \begin{bmatrix} C & 0 \\ 0 & D \end{bmatrix}^{-1} \begin{bmatrix} \mu - x_0 \\ e - Xx_0 \end{bmatrix} - l(x_0|y),$$

where $D = \operatorname{diag}(D_1, \ldots, D_n)$, and where neither e nor D depend functionally on x_0 or C.

Let $\hat{x}_0$ be the weighted least-squares regression vector estimate after regressing $(\mu', e')'$ on $(I, X')'$ with weighting matrix $\operatorname{diag}(C, D)$. Then

$$\begin{bmatrix} \mu \\ e \end{bmatrix} - \begin{bmatrix} I \\ X \end{bmatrix} x_0 = M \begin{bmatrix} \mu \\ e \end{bmatrix} - \begin{bmatrix} I \\ X \end{bmatrix} (x_0 - \hat{x}_0),$$

where M is the usual error projection matrix associated with the regression

$$I - \begin{bmatrix} I \\ X \end{bmatrix} \left[(I, X') \begin{bmatrix} C & 0 \\ 0 & D \end{bmatrix}^{-1} \begin{bmatrix} I \\ X \end{bmatrix} \right]^{-1} (I, X') \begin{bmatrix} C & 0 \\ 0 & D \end{bmatrix}^{-1}$$
$$= I - \begin{bmatrix} I \\ X \end{bmatrix} (C^{-1} + X'D^{-1}X)^{-1}(C^{-1}, X'D^{-1}).$$

The matrix M has the properties

$$M(I, X')' = 0, \quad MM = M, \quad \operatorname{diag}(C^{-1}, D^{-1})M = M' \operatorname{diag}(C^{-1}, D^{-1}),$$

from which it follows that $l(y)$ is given by

$$\log|C| + \log|D| + \begin{bmatrix} \mu \\ e \end{bmatrix}' \begin{bmatrix} C & 0 \\ 0 & D \end{bmatrix}^{-1} M \begin{bmatrix} \mu \\ e \end{bmatrix} + (x_0 - \hat{x}_0)'(C^{-1} + X'D^{-1}X)(x_0 - \hat{x}_0) - l(x_0|y).$$

Now the penultimate term is the exponent term in the conditional density of x_0 given y. This follows from direct computations using matrix inversion identities. Hence the conditional covariance matrix of x_0 given y is $(C^{-1} + X'D^{-1}X)^{-1}$ from which it follows that $l(y)$ is

$$\log|C| + \log|D| - \log|(C^{-1} + X'D^{-1}X)^{-1}| + \begin{bmatrix} \mu \\ e \end{bmatrix}' \begin{bmatrix} C & 0 \\ 0 & D \end{bmatrix}^{-1} M \begin{bmatrix} \mu \\ e \end{bmatrix}.$$

Substituting the above expression for M yields (3) by putting $S = X'D^{-1}X$ and $s = X'D^{-1}e$. That the $X_t \equiv F_t Z_{t-1}$ $(t = 1, \ldots, n)$ matrices are as asserted in the statement of the theorem follows from the Kalman filter algorithm. This completes the proof.

ACKNOWLEDGEMENTS

This research was supported by a grant from Z.W.O., the Dutch agency for the support of pure scientific research. I am indebted to Jan G. de Gooijer, Andrew C. Harvey and Murray J. Mackinnon for helpful comments.

References

AKAIKE, H. (1978). Covariance matrix computations of the state variable of a stationary Gaussian process. *Ann. Inst. Statist. Math.* B **30**, 499–504.

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.

GARDNER, G., HARVEY, A. C. & PHILLIPS, G. D. A. (1980). An algorithm for exact maximum likelihood estimation by means of Kalman filtering. *Appl. Statist.* **29**, 311–22.

HARVEY, A. C. (1981). *Time Series Models.* New York: Wiley.

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

JONES, R. H. (1980). Maximum likelihood fitting of ARMA models to time series with missing observations. *Technometrics* **22**, 389–95.

NEWBOLD, P. (1974). The exact likelihood function for a mixed autoregressive-moving average process. *Biometrika* **61**, 423–6.

ROSENBERG, B. (1973). Random coefficient models: The analysis of a cross-section of time series by stochastically convergent parameter regression. *Ann. Econ. Social Meas.* **2**, 399–428.

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

SCHWEPPE, F. C. (1973). *Uncertain Dynamic Systems.* Englewood Cliffs, New Jersey: Prentice-Hall.

[*Received March* 1987. *Revised June* 1987]