MONASH
UNIVERSITY

AUSTRALIA

KALMAN FILTERING

THE INITIALIZATION PROBLEM

Ralph D. Snyder and Grant R. Saligari

Working Paper No. 2/92

February 1992

DEPARTMENT OF ECONOMETRICS

ISSN 1032-3813

ISBN 0 7326 0352 8

KALMAN FILTERING

THE INITIALIZATION PROBLEM

Ralph D. Snyder and Grant Saligari

Working Paper No. 2/92

February 1992

DEPARTMENT OF ECONOMETRICS, FACULTY OF ECONOMICS COMMERCE & MANAGEMENT

MONASH UNIVERSITY, CLAYTON, VICTORIA 3168, AUSTRALIA.

Kalman Filtering
The Initialization Problem

by

Ralph D. Snyder
Department of Econometrics
Monash University

and

Grant R. Saligari
Strategic Manufacturing Technology Pty Ltd

February, 1992

gs_rs_1

Summary

The problem of computing estimates of the state vector in a non-stationary dynamic linear model is considered. Such estimates cannot be obtained with the usual Kalman filter because it fails, on finite precision computing machines, when seeded with the infinite variances associated with the required diffuse or partially diffuse prior probability distribution. The response in the literature has been the development of a number of relatively complex hybrid filters specifically designed to avoid the problem. However, it is argued in this paper that this response has been largely unwarranted. Rather, it is established that any square root implementation of the Kalman filter is capable of producing satisfactory results, so long as the required triangular orthogonalisation calculations are undertaken with standard fast Givens transformations, rather than the more usual Householder or Gram-Schmidt procedures.

Key words: dynamic linear model, structural model, Kalman filter, square root filter, Givens transformations.

1. Introduction

The dynamic linear framework used in this paper involves an observable random variable y_t associated with typical period t, which depends linearly on a random r-vector β_{t-1} of so called state variables, together with a random p-vector ε_t of disturbances, as follows:

$$y_t = x'\beta_{t-1} + h'\varepsilon_t, \qquad (1a)$$

x and h being commensurate fixed vectors. The state variables themselves are assumed to evolve over time according to the first order Markovian relationship

$$\beta_t = T\,\beta_{t-1} + A\,\varepsilon_t, \qquad (1b)$$

T and A being fixed rxr and rxp matrices respectively. The disturbances ε_t, which all have a mean of zero and a contemporaneous variance matrix σ^2 V, are intertemporally independent. Slightly different definitions are also conceivable (eg see Harvey and Phillips, 1979) but the general ideas outlined in this paper also apply to them.

The importance of the dynamic linear framework can be gauged from the fact that it encompasses all linear models of series which evolve over time. It has been used as the basis for estimating ARMA models (Akaike 1978, Harvey and Phillips 1979, Gardner et al 1980) and ARIMA processes (Harvey and Pierse, 1984, Ansley and Kohn, 1985). It also forms the basis of the structural approach (Harvey and Todd 1983, Snyder 1985) for the analysis and prediction of economic time series.

In many practical cases of interest, it has been found that the state vector may contain elements with a mixture of zero, bounded or unbounded variances. For example, when seeding a non-stationary time series model at least one of the state variances will be undefined. If linear constraints are also incorporated then some variances will also be zero.

Example: Damped Trend Corrected Exponential Smoothing

Consider the damped trend corrected exponential smoothing model

$$\begin{aligned} y_t &= \lambda_{t-1} + \delta_{t-1} + \varepsilon_t \\ \lambda_t &= \lambda_{t-1} + \delta_{t-1} + \alpha_1 \varepsilon_t \\ \delta_t &= \phi\delta_{t-1} + \alpha_2 \varepsilon_t \end{aligned} \qquad (2)$$

for which the initial conditions can be shown to be

$$\begin{bmatrix} \lambda_0 \\ \delta_0 \end{bmatrix} \sim \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \sigma^2 \begin{bmatrix} \tau & v_{12} \\ v_{12} & v_{22} \end{bmatrix} \right) \qquad (3)$$

where $v_{12} = \dfrac{(\phi\alpha_2^2 + (1-\phi^2)\alpha_1\alpha_2)}{(1-\phi)(1-\phi^2)}$, $v_{22} = \dfrac{\alpha_2^2}{(1-\phi^2)}$

and τ is an arbitrarily large number representing, in effect, an infinite variance. The model (2) is non-stationary and the distribution given by (3) is partially diffuse. If seeded with (3), the conventional Kalman filter breaks down.

Conventional covariance filters require the covariance matrix to be defined and, therefore, are unable to cope with the non-stationary case. Information filters, on the other hand, which are based on the inverse covariance matrix, are unable to cope with zero variances. Square root versions of these filters, although able to offer improved numerical properties, do not address this fundamental problem. This poses somewhat of a dilemma for the practitioner wishing to compute reliable estimates of the state vector and its covariance.

To address this problem a number of hybrid filters have been proposed in the literature - see, for example, Ansley and Kohn (1985), De Jong (1988) and Snyder (1988). For a mixed covariance-information approach to the problem also refer to Paige (1985). These approaches are complex and, in the case of Paige's approach, require substantial modifications to the form of the conventional state space model. As such, practitioners are

likely to be slow to adopt these methods, relying instead on the computationally inferior standard covariance filter.

In section two the covariance and information approaches to the filtering problem, together with some hybrid filters, are reviewed. In section three a covariance square root filter is described. It is shown that Stirling's (1983), now standard approach for performing the necessary matrix factorisations, automatically caters for the infinite and zero variance cases, without making any special provisions for them in the filter itself. As such, it is suggested that the proposed method is not only more elegant than the available alternatives, but is preferable in that it entails lower computational loads.

2. Covariance and Information Filtering Techniques

Assume that the distribution of β_{t-1} given $y_1, \ldots, y_t$ is known and has a mean b_{t-1} and a covariance matrix $\sigma^2 B_{t-1}$. The covariance filter for obtaining b_t and $\sigma^2 B_t$, conditional on $y_1, \ldots, y_{t-1}$ and a new observation y_t, for a problem of this general form was derived by Jazwinski (1970 pp 210-212) and involves the relationships

$$b_t = Tb_{t-1} + (TB_{t-1}x + AVh)(x'B_{t-1}x + h'Vh)^{-1}(y_t - x'b_{t-1}) \qquad (4a)$$

and

$$B_t = TB_{t-1}T' + AVA' - (TB_{t-1}x + AVh)(TB_{t-1}x + AVh)'(x'B_{t-1}x + h'Vh)^{-1}. \qquad (4b)$$

Such recursions must be initialized with b_0 and $\sigma^2 B_0$, where b_0 and B_0 are assumed to be known.

For the case where all of the inverses exist standard matrix inversion techniques can be used to derive, from the original covariance equations, formulae for propagating the inverse covariance. Because of the relationship of the inverse covariance to the Fisher Information matrix, such filters are termed

information filters (see, for example, Anderson and Moore 1979, section 6.3). These filters obtain B_t^{-1} and $B_t^{-1}b_t$ from their lagged counterparts and a new observation y_t along similar lines to the covariance equations (Kaminski et al 1976 demonstrates an interesting duality between the two sets of equations). Seeding with a diffuse prior distribution is performed by setting B_0^{-1} equal to zero.

During the course of the calculations on finite precision computing machines it is possible for rounding errors to cause a state covariance matrix (or its inverse) to be computed that is not positive semi-definite - a theoretical impossibility. To circumvent this difficulty various methods for updating the covariance and inverse covariance matrices in square root form have been proposed.

Working with the covariance form, Potter (1964) outlined a method for updating the covariance matrix in square root form in the absence of process noise. Potter's algorithm was extended by Bellantoni and Dodge (1967) to handle vector measurements and subsequently by Andrews (1968) to include process noise as well. Morf and Kailath (1975) combined the ideas of time and measurement updates into a single step.

Square root information filters have also been proposed in the literature. Golub (1965) and Businger and Golub (1965) demonstrated a square root solution to the least squares problem using Householder transformations. Hanson and Lawson (1969) extended this to the case of rank deficient systems, whilst Dyer and McReynolds (1969) applied the Householder algorithm to obtain both measurement and time updates of the information square root. More recently, a regression formulation of the problem has been suggested (Duncan and Horne 1972) which Paige and Saunders (1977) have used as the basis of a square root information filter.

Although square root filters provide improvements in accuracy a number of problems remain. Both the conventional covariance and square root covariance filters are unable to cope with the diffuse

or partially diffuse prior probability distributions required in the non-stationary case. Conversely, information versions of these filters are unable to cope with the case of zero variances. Harvey and Phillips (1979) have suggested seeding the ordinary covariance filter with B_0 equal to τI, where τ is a large number. This method, however, is inexact and often suffers from numerical instability brought about by the large numbers involved. To circumvent these difficulties a number of hybrid filters have been proposed in the literature.

Ansley and Kohn (1985) demonstrate that when a Kalman filter is seeded with a variance matrix B_0 with the general form

$$B_0 = \tau B_0^{[1]} + B_0^{[0]}$$

then, except for an error term $O(\tau^{-1})$ which converges to zero as τ tends to infinity, successive b_t and B_t generated by a covariance filter are constant and linear functions of τ respectively. More specifically

$$b_t = b_t^{[0]} + O(\tau^{-1}) \qquad (5a)$$

and

$$B_t = \tau B_t^{[1]} + B_t^{[0]} + O(\tau^{-1}). \qquad (5b)$$

By substituting (5a) and (5b), together with their lagged counterparts, into the covariance filter (4a) and (4b), equations for obtaining successive $b_t^{[0]}$, $B_t^{[1]}$ and $B_t^{[0]}$ are obtained. These equations, which will not be reproduced here, define a hybrid filter involving more equations and therefore more calculations. Normally $B_t^{[1]} = 0$ when $t > r$, r being the dimension of the state vector β_t, in which case the hybrid reduces to a conventional covariance filter. The increased computational loads are, therefore, temporary.

De Jong (1988) derives an expression for the likelihood function of a state space model initialized at a starting estimate of zero and an associated estimation error covariance matrix of zero. Adjustment for initial conditions is made after filtering. The

covariance filter can, therefore, be applied without special consideration of start-up effects. In particular, the initial conditions can be modelled as diffuse.

Using our notation, minus two times the log likelihood is, apart from constants, given by

$$l(y) = \log|B_0| + b_0'B_0^{-1}b_0 + \Sigma_{t=1}^{n}\log|\lambda_t| + \Sigma_{t=1}^{n}e_t^2/\lambda_t$$
$$+ \log|B_0^{-1} + S| - (B_0^{-1}b_0 + s)'(B_0^{-1} + S)^{-1}(B_0^{-1}b_0 + s).$$

where e_t and λ_t are the usual one step ahead prediction error and its variance. The vector s and the matrix S are calculated in parallel with the e_t and λ_t as follows

$$s_t = s_{t-1} + Z_{t-1}'x'e_t/\lambda_t$$
$$S_t = S_{t-1} + Z_{t-1}'xx'Z_{t-1}/\lambda_t$$
$$Z_t = T(I - k_tx')Z_{t-1}$$

with s and S initialized at 0, $Z_0 = I$ and with k_t being the Kalman gain vector.

The auxiliary variables, s and S, capture the effect of the starting conditions. De Jong shows how these may be set to reflect the appropriate distribution assumptions at the end. Although this can provide some insight into the nature of the distribution of b_t, because of the need to compute and store these auxiliary variables, it does appear that, as an operational method, it is expensive in terms of requiring extra computations and possessing greater storage needs throughout.

Both Ansley and Kohn and De Jong's approaches are able to handle the diffuse or partially diffuse prior case. As they are essentially covariance filters they also have no difficulty in coping with the zero variance case. An alternative approach that also meets these requirements was proposed in Paige (1985) and involves a combined covariance - information approach to the

problem.

Consider a random vector x with mean $\bar{x}$ and covariance matrix Q. Let U and S be the matrices in the factorisation of Q^{-1} and Q respectively. As Paige points out both

$$x = \bar{x} + Se \qquad E[e] = 0 \text{ and } E[ee'] = I \qquad (6a)$$

and

$$Ux = r + e \qquad E[e] = 0 \text{ and } E[ee'] = I \qquad (6b)$$

are possible representations for x. Only when $S = U^{-1}$ are these equivalent. Interestingly, (6a) allows for some elements of x to be constant by specifying an S matrix with less than full row rank. Conversely (6b) allows for some elements of x to be completely unspecified by specifying a U matrix with less than full row rank. As Paige indicates, by replacing these with Ux = r + Se, with e defined as before, both situations can be represented. In particular the state vector β can be modelled as $\beta_0 = Ub_0 + Se$.

3. Square Root Covariance Filter

In this section we propose a square root covariance filter for the estimation of the problem described in (1). As well as being of interest in its own right, the solution demonstrates the use of Givens transformations for the computation of the necessary matrix factorisations, the main point being that then the diffuse and partially diffuse cases can be handled without special modifications to the filter itself (cf Ansley and Kohn 1985, Snyder 1988).

Suppose that R_{t-1} and Λ_{t-1}, respectively a unit upper triangular matrix and a diagonal matrix from the Cholesky decomposition of B_{t-1}, are known from a previous iteration of the algorithm. It is shown in the Appendix that Givens transformations can be used to process the matrix

$$\begin{bmatrix} R_{t-1}x & R_{t-1}T' \\ h & A' \end{bmatrix} \text{ relative to the weight matrix } \begin{bmatrix} \Lambda_{t-1} & 0 \\ 0 & V \end{bmatrix} \quad (7a)$$

to obtain a unit upper triangular matrix

$$\begin{bmatrix} 1 & k_t \\ 0 & R_t \end{bmatrix} \text{ and a diagonal matrix } \begin{bmatrix} \lambda_t & 0 \\ 0 & \Lambda_t \end{bmatrix}, \quad (7b)$$

where the r-vector k_t is the so-called Kalman gain vector. It is also shown that the revised state estimates are given by

$$b_t = Tb_{t-1} + k_t(y_t - x'b_{t-1}),$$

that, if required, B_t can be computed with

$$B_t = R'_t\Lambda_t R_t$$

and that λ_t is the mean squared error of $(y_t - x'b_{t-1})$. The algorithm is seeded with the Cholesky factorization of B_0 given by

$$B_0 = R'_0\Lambda_0 R_0.$$

When some elements of Λ_0 are infinite and hence the corresponding elements of Λ_0^{-1} are zero, Stirling's (1983) implementation of Givens transformations may be applied without special provisions being made. To see this consider the problem of updating a diagonal matrix C and a unit upper triangular matrix R to allow for the addition of successive rows of data to a typical matrix X, such that

$$X'V^{-1}X = R'C^{-1}R,$$

where V^{-1} is diagonal with elements serving as row weights. Stirling's formula for updating the rows r_i of R corresponding to elements c_i of C upon addition of a new row $\mathbf{x}$ with left-most non-zero element x_i, are as follows.

Case 1 ($0 < c_i < \infty$)

$$v^+ = v + c_i x_i^2$$
$$c_i^+ = c_i v / v^+$$
$$r^+ = (v/v^+) r_i + (c_i x_i / v^+) \mathbf{x}$$
$$x^+ = \mathbf{x} - x_i r_i$$

Case 2 ($c_i = 0$)

$$x^+ = \mathbf{x} - x_i r_i$$

Case 3 ($c_i = \infty.$)

$$r^+ = \mathbf{x}/x_i$$
$$c_i^+ = v/x_i^2$$
$$\mathbf{x}^+ = 0$$

These formulae are applied repetitively until $\mathbf{x}$ is reduced to a null vector. It is apparent that in cases 2 and 3, which are special versions of case 1, the algorithm has a particularly simple form and involves lower than normal computational loads.

When applied to (7) the algorithm is particularly appealing. Initialisation with a diffuse prior probability distribution requires setting the elements of Λ_0 to infinity, corresponding to case 3. Thus processing the first r rows of (7a) involves reduced computational loads. For the remaining rows and upon subsequent iterations of the algorithm, conventional Givens transformations are used. Computational loads are, therefore, reduced during the "run-in". The particularly appealing factor, however, is that all of these cases are transparently handled by Stirling's algorithm without modifying the basic structure of the filter.

4. Conclusion

The basic message of this paper is that hybrid procedures for estimating non-stationary time series models that might also incorporate linear constraints are, in fact, unnecessary. Conventional square root filters whether they be based on

covariance or inverse covariance matrices, may be used without any basic structural changes, provided that fast Givens transformations are employed for the required triangular orthogonalisation transformations. This, therefore, considerably simplifies what has hitherto been a daunting computational issue in the area of Kalman filtering.

References

Akaike H., (1974), "Markovian representation of stochastic processes and its application to the analysis of autoregressive moving average processes", *Annals of the Inst. of Stat. Math.*, 26, 363-387.

Akaike H., (1978), "Covariance matrix computation of the state variable of a stationary Gaussian process", *Annals of the Inst. of Stat. Math.*, 30, 499-504.

Anderson and Moore, (1979), *Optimal Filtering*, Prentice-Hall Information and System Sciences Series, Englewood Cliffs N.J.

Andrews A., (1968), "A square root formulation of the Kalman covariance equations", *AIAA J.*, 6, 1165-1166.

Ansley C.F. and Kohn R., (1985), "Estimation, filtering and smoothing in state space models with incompletely specified initial conditions", *The Annals of Statistics*, 13, 1286-1316.

Bellantoni J.F. and Dodge K.W., (1967), "A square root formulation of the Kalman-Schmidt filter", *AIAA J.*, 5, 1309-1314.

Businger P. and Golub G.H., (1965), "Linear least squares solution by Householder transformations", *Numer. Math*, 7, 269-276.

Duncan D.B. and Horne S.D., (1972), "Linear dynamic recursive estimation from the viewpoint of regression analysis", *Journal of the American Statistical Association*, 67, 815-821.

De Jong P., (1988), "The likelihood for a state space model", *Biometrika*, 75, 165-169.

Dyer P. and McReynolds S., (1969), "Extension of square root filtering to include process noise", *Journal Optimiz. Theory Appl.*, 3, 444-459.

Gardner G., Harvey A.C. and Phillips G.D.A, (1980), "An algorithm for exact maximum likelihood estimation by means of Kalman filtering", *Applied Statastics*, 29, 311-322.

Golub G.H., (1965), "Numerical methods for solving linear least squares problems", *Numer. Math*, 7, 206-216.

Hanson R.J. and Lawson C.L., (1969), "Extensions and applications of the Householder algorithm for solving linear least squares problems", *Math Comput.*, 23, 787-812.

Harvey A.C. and Pierse R.G., (1984), "Estimating missing observations in Economic time series", *Journal of the American Statistical Association*, 79, pp 125-131.

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 Todd P.H.J., (1983), "Forecasting economic time series with structural and Box-Jenkins models (with discussion)", *Journal of Business and Economic Statistics*, 1, 299-315.

Jazwinski A.H., (1970), *Stochastic Processes and Filtering Theory*, Academic Press New York.

Kaminski P.G., Bryson A.E. and Schmidt S.F., (1971), "Discrete square root filtering: a survey of current techniques", *IEEE Trans. Automatic Control*, AC-16, 727-734.

Morf M. and Kailath T., "Square root algorithms for least squares estimation", *IEEE Trans. Automatic Control*, AC-20, 487-497.

Potter (1964), in R.H. Battin, *Astronautical Guidance*, McGraw-Hill New York.

Paige C.C, (1985), "Covariance matrix representation in linear

filtering", *Contemporary Mathamatics*, 47, 309-321.

Paige C.C. and Saunders M.A., (1977), "Linear least squares estimation of discrete dynamic systems using orthogonal transformations", *SIAM Journal of Numerical Analysis*, 15, 180-193.

Snyder R.D. ,(1985), "Recursive estimation of dynamic linear statistical models", *Journal of the Royal Statistical Society*, B, 47, 272-276.

Snyder R.D., (1988), "Computational aspects of Kalman filtering with a diffuse prior distribution", *Journal of Statistical Computations and Simulations*, 29, 79-86.

Stirling W.D., (1983), "Least squares subject to linear constraints", *Journal of Applied Statistics*, 30, 204-212.

Appendix : Derivation of the square root covariance filter

1. Assume that the distribution of β_{t-1} conditional on $y_1,\ldots,$ y_{t-1} is known and given by $N(b_{t-1},\sigma^2B_{t-1})$. Hence we can write $\beta_{t-1} = b_{t-1} + \xi_{t-1}$, where $\xi_{t-1} \sim N(0,\sigma^2B_{t-1})$. Substitute this into (1a) and (1b) to give

$$y_t = x'b_{t-1} + (x'\xi_{t-1} + h'\varepsilon_t)$$

and

$$\beta_t = Tb_{t-1} + (T\xi_{t-1} + A\varepsilon_t),$$

so that

$$\begin{bmatrix} y_t \\ \beta_t \end{bmatrix} \sim N\left(\begin{bmatrix} x'b_{t-1} \\ Tb_{t-1} \end{bmatrix}, \sigma^2 \begin{bmatrix} x'B_{t-1}x + h'Vh & x'B_{t-1}T' + h'VA' \\ TB_{t-1}x + AVh & TB_{t-1}T' + AVA' \end{bmatrix} \right). \quad (8)$$

By Cholesky factorization this can be rewritten as

$$\begin{bmatrix} y_t \\ \beta_t \end{bmatrix} \sim N\left(\begin{bmatrix} x'b_{t-1} \\ Tb_{t-1} \end{bmatrix}, \sigma^2 \begin{bmatrix} 1 & 0 \\ k_t & R_t' \end{bmatrix} \begin{bmatrix} \lambda_t & 0 \\ 0 & \Lambda_t \end{bmatrix} \begin{bmatrix} 1 & k_t' \\ 0 & R_t \end{bmatrix} \right). \quad (9)$$

Hence

$$\begin{bmatrix} 1 & 0 \\ k_t & R_t' \end{bmatrix}^{-1} \begin{bmatrix} y_t \\ \beta_t \end{bmatrix} \sim N\left(\begin{bmatrix} 1 & 0 \\ k_t & R_t' \end{bmatrix}^{-1} \begin{bmatrix} x'b_{t-1} \\ Tb_{t-1} \end{bmatrix}, \sigma^2 \begin{bmatrix} \lambda_t & 0 \\ 0 & \Lambda_t \end{bmatrix} \right).$$

It follows, after explicitly inverting and expanding, that $\beta_t = Tb_{t-1} + k_t(y_t - x'b_{t-1}) + R_t'\eta_t$, where $\eta_t \sim N(0,\Lambda_t)$, and that $y_t \sim N(x'b_{t-1},\lambda_t)$. Accordingly, β_t conditional on $y_1 \ldots y_{t-1}$, together with y_t, has the distribution

$$\beta_t \sim N(Tb_{t-1} + k_t(y_t - x'b_{t-1}),\ R_t'\Lambda_tR_t).$$

In other words the new mean and variance are

$$b_t = b_{t-1} + k_t(y_t - x'b_{t-1})$$

and

$$B_t = R_t'\Lambda_tR_t.$$

2. It is not particularly efficient to form the variance matrix (8) and then use the Cholesky factorization algorithm to obtain the triangular representation in (9). An alternative strategy is to recognize that B_{t-1} will be available in decomposed form as

$$B_{t-1} = R'_{t-1}\Lambda_{t-1}R_{t-1} ,$$

in which case (9) can be rewritten as

$$\begin{bmatrix} y_t \\ \beta_t \end{bmatrix} \sim N\left(\begin{bmatrix} x'b_{t-1} \\ Tb_{t-1} \end{bmatrix}, \sigma^2 \begin{bmatrix} x'R'_{t-1} & h' \\ TR'_{t-1} & A \end{bmatrix} \begin{bmatrix} \Lambda_{t-1} & 0 \\ 0 & V \end{bmatrix} \begin{bmatrix} R_{t-1}x & R_{t-1}T' \\ h & A' \end{bmatrix}\right). \quad (10)$$

Givens transformations can then be used to obtain the triangular decomposition of the variance matrix as described in the paper without explicitly having to compute either B_t or B_{t-1}. Not only does this reduce computational loads but it also leads to a more accurate algorithm because "squaring" operations are avoided.