SERIE RESEARCH MEMORANDA

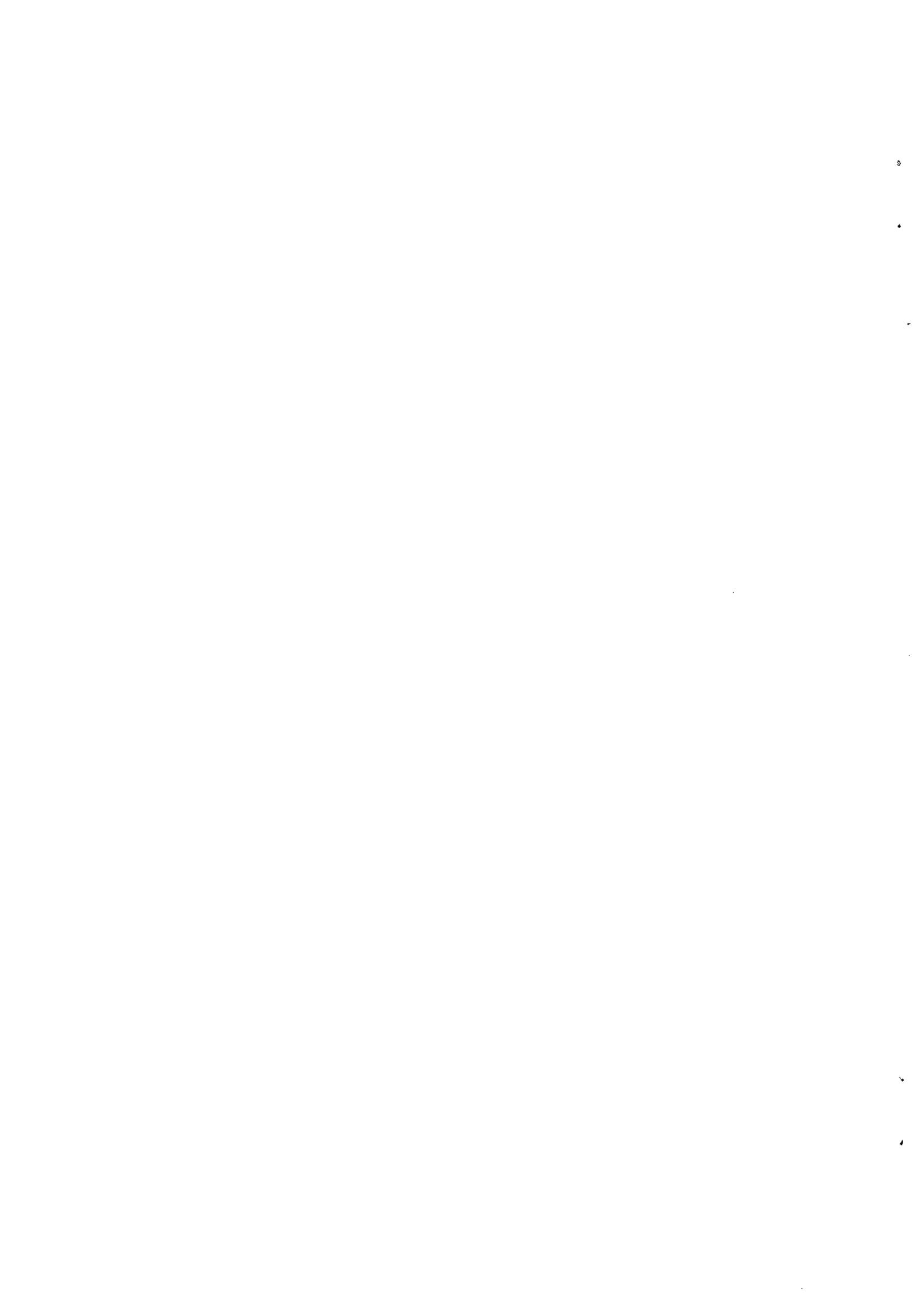CONSISTENT DIFFUSE INITIAL CONDITIONS
IN THE KALMAN FILTER

Ivo J. Steyn

Research Memorandum 1989-15

April 1989

VRIJE UNIVERSITEIT
FACULTEIT DER ECONOMISCHE WETENSCHAPPEN
EN ECONOMETRIE
AMSTERDAM

CONSISTENT DIFFUSE INITIAL CONDITIONS IN THE KALMAN FILTER

Ivo J. Steyn

Abstract : For nonstationary State Space models, diffuse initial conditions are usually represented by choosing the initial variance matrix P_0 equal to k.I, with k some large constant. It is shown that this leads to inappropriate (unconditional) variance structures in some common models, and that it leads to inconsistencies under transformations.
An alternative method of generating P_0 is proposed which does not suffer from these defects, and in a small simulation study appears to have considerable advantages over the traditional choice of k.I.

Keywords : State Space Model, Kalman Filter, nonstationarity, diffuse initial conditions

CONSISTENT DIFFUSE INITIAL CONDITIONS IN THE KALMAN FILTER

Ivo J. Steyn[1]

1. Introduction

The immensely useful State Space model is usually estimated with a Kalman Filter. The Kalman Filter requires the specification of an unconditional density of α_0, an unobserved vector. To express our ignorance about this vector it is desirable to specify "diffuse" initial conditions, specifically a "large" variance matrix.

Most researchers follow Harvey[1981] in using k.I for this initial variance matrix. The approach of Ansley & Kohn[1985] and Kohn & Ansley[1986] is also based on such a variance matrix for nonstationary processes.

In this paper an alternative form for P_0 is examined, a form which depends on the eigenvalues of the transition matrix T – and thus by extension on the autoregressive structure of the model. This matrix is easily calculated and takes the covariance structure of the vector process α_t explicitly into account. It was first suggested by Theo Nijman and has been used in a small simulation study in Gehring[1987].

The paper proceeds as follows : in section 2 the choice of P_0 = k.I is examined and shown to have nonsensical implications in a common model. Section 3 introduces a new method of calculating a sensible P_0, while section 4 gives examples and a small simulation study in which the new option appears to have favourable properties. The last section attempts to summarise.

2. P_0 as unconditional variance matrix

We will consider the following parametrisation of the State Space Model:

(1a) $\alpha_t = T\alpha_{t-1} + R\eta_t \qquad \eta_t \approx N(0,\sigma^2 I_n)$

(1b) $y_t = Z_t\alpha_t + S\epsilon_t \qquad \epsilon_t \approx N(0,\sigma^2 I_g)$

When calculating the likelihood of this model, it is convenient to use the prediction error decomposition:

(2) $f(y_t \ \ldots \ y_1) = \prod_{i=2}^{t} f(y_i|y_{i-1} \ \ldots \ y_1) \cdot f(y_1)$

in which the conditional likelihoods are evaluated easily with the Kalman Filter while the likelihood of the first observation is a simple transformation of the density of α_0. This leaves us with the problem of specifying the unconditional density of α_0, or rather, its unconditional mean a_0 and its unconditional variance P_0 (where we have divided the scale factor σ^2 out of the variance matrix).

It is a common mistake to suppose that stationary processes have an unconditional mean of 0. In reality, a stationary process can have a nonzero mean, such as:

(3) $y_t = \mu + \rho y_{t-1} + \epsilon_t$

However, this is only useful when the parameter μ is actually known. When it is unknown, the appropriate model would be:

(4a)
$$\begin{bmatrix} y_t \\ \mu \end{bmatrix} = \begin{bmatrix} \rho & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ \mu \end{bmatrix} + \begin{bmatrix} \epsilon_t \\ 0 \end{bmatrix}$$

(4b) $y_t = (1 \ 0)(y_t \ \mu)'$

and in this State Space model, one component follows a degenerate nonstationary process. The problem of specifying an unconditional mean for α_0 in (4a)-(4b) is not easily solved.

For the unconditional variance matrix P_0 similar problems exist. When the mean of a stationary process is known, the unconditional variance matrix is a – somewhat complicated– function of the various ARMA coefficients. For nonstationary processes P_0 is of course undefined.

One solution to this problem has been suggested by Harvey[1981] among others. Ignorance about the unconditional mean is expressed as a "large" variance matrix P_0, to wit $P_0 = k.I$ with k some large positive number. Harvey also uses this solution as appropriate for nonstationary processes. The use of such "diffuse" initial conditions has the very great advantage that the choice of a_0 becomes unimportant : it is easy to show that the influence of that choice on subsequent estimates drops to a minimum as P_0 approaches infinity (in the sense that its smallest eigenvalue goes to infinity).

While this method has the advantage of simplicity, it suffers from two very serious defects :

–   the choice of $P_0 = k.I$ automatically imposes independence on the components of α_0, and it also assumes these components have equal variance.

–   it leads to contradictory specifications for models which are related through a non-orthogonal transformation

To realise how inappropriate the assumptions regarding variances and covariances usually are, let us examine the popular Structural Time Series model used by Harvey[1985], Harvey & Todd[1983], Den Butter & Mourik[1986] and many others.

(5a) $d_t = d_{t-1} + \beta_{t-1} + \epsilon_{1t}$

(5b) $\beta_t = \beta_{t-1} + \epsilon_{2t}$

(5c) $S_t = - S_{t-1} - S_{t-2} - S_{t-3} + \epsilon_{3t}$

(5d) $y_t = d_t + S_t + \epsilon_{4t}$

in which d_t is an I(2) trend process and S_t a variable seasonal term. The State Space model for (5a)-(5d) is:

(6a) $\alpha_t = \begin{bmatrix} d_t \\ \beta_t \\ S_t \\ S_{t-1} \\ S_{t-2} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & -1 & -1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} d_{t-1} \\ \beta_{t-1} \\ S_{t-1} \\ S_{t-2} \\ S_{t-3} \end{bmatrix} + \begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \\ \epsilon_{3t} \\ 0 \\ 0 \end{bmatrix}$

(6b) $y_t = (1 \ 0 \ 1 \ 0 \ 0)\alpha_t + \epsilon_{4t}$

Since we usually do not have information about α_0, we would use $P_0 = k.I$ in this case. That leads to the following questionable unspoken assumptions about conditions on t=0 :

- the level and the growth of the trend have the same variance and are independent
- the three seasonal terms are independent
- the seasonal terms have the same variance as the level and growth of the trend

Further implications are revealed when one calculates the unconditonal variance of S_1, the first "observed" seasonal : this turns out to be 3k, and the pattern of every fourth seasonal having triple the (unconditional) variance of the other seasonals repeats itself.

The second drawback of the method is revealed when one considers equivalent models for the same process. The seasonal term S_t above can be modelled as in (6a), but an equivalent model would be:

$$(6a') \quad \beta_t = \begin{bmatrix} S_t \\ -S_{t-1}-S_{t-2} \\ -S_{t-1} \end{bmatrix} = \begin{bmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \\ -1 & 0 & 0 \end{bmatrix} \begin{bmatrix} S_{t-1} \\ -S_{t-2}-S_{t-3} \\ -S_{t-2} \end{bmatrix} + \begin{bmatrix} \epsilon_{3t} \\ 0 \\ 0 \end{bmatrix}$$

where β_t and the corresponding section of α_t are related through a simple transformation that leaves RR' and Z_t unchanged[2]. If we denote this relationship by $\beta_t = H\alpha_t$, then a choice of P_0 as variance of α_0 should lead to β_0 having a variance of HP_0H', which, since H is not orthogonal, does not equal k.I. So giving α_0 rather than β_0 a variance of k.I is an arbitrary choice.

Summarising, the choice of $P_0 = k.I$, while computationally convenient, implies a variance structure for α_0 - and, by extension, all following α_t - that is arbitrary and often inappropriate.

3. An alternative choice for P_0

Let us assume that at some distant time in the past, say at $t = -m$, the vector α_{-m} had an unconditional variance of P_{-m}. What does this imply for α_0?
From (1a) it is obvious that

$$(7) \quad P_0 = T^m P_{-m}(T^m)' + \sum_{k=0}^{m-1} T^k RR'(T^k)'$$

We will now assume that the matrix T is diagonalisable, i.e. that it has n distinct eigenvalues, an assumption we will drop later on. We can now write

$$(8) \quad T = Q\Lambda Q^{-1} \qquad \text{where} \quad \Lambda = \text{diag}(\lambda_1....\lambda_n)$$

Now we can rewrite (7) as:

$$(9) \quad \Psi_0 = \Lambda^m \Psi_{-m}(\Lambda^m)' + \sum_{k=0}^{m-1} \Lambda^k \Sigma(\Lambda^k)'$$

where we have defined

$$(10) \quad Q^{-1}P_t(Q^{-1})' = \Psi_t$$

$$(11) \quad Q^{-1}RR'(Q^{-1})' = \Sigma$$

The relation (9) can also be written in terms of individual matrix elements. If λ_i are the (diagonal) elements of Λ and σ_{ij} the ij-th element of Σ we can write

---

[2] If this model should seem contrived, it is in fact the State Space model we would get if (5c) was written as a SSM using the standard SSM form for ARMA models, see Harvey[1981, p.103].

$$(12) \quad \psi_{ij,o} = (\lambda_i\lambda_j)^t\psi_{ij,-m} + \sum_{k=0}^{m-1}(\lambda_i\lambda_j)^k\sigma_{ij}$$

$$= (\lambda_i\lambda_j)^t\psi_{ij,-m} + [1 - (\lambda_i\lambda_j)^m][1 - \lambda_i\lambda_j]^{-1}\sigma_{ij} \qquad (|\lambda_i\lambda_j| \neq 1)$$

$$= \psi_{ij,-m} + m\sigma_{ij} \qquad (\lambda_i\lambda_j = 1)$$

If we move our starting point $t=-m$ further back to minus infinity, the elements of ψ_0 approach the following limit forms:

$$(13) \quad \psi_{ij,o} = \sigma_{ij}[1 - \lambda_i\lambda_j]^{-1} \qquad (|\lambda_i\lambda_j| < 1)$$

$$= m\sigma_{ij} \qquad (\lambda_i\lambda_j = 1)$$

with no limit form available when $|\lambda_i\lambda_j| > 1$, or when $|\lambda_i\lambda_j| = 1$ and $\lambda_i\lambda_j \neq 1$. (although it can be proven that (12) produces bounded functions of m in the latter case) So any assumptions we make about P_{-m} and its tranformation ψ_{-m} will become irrelevant as m increases. The only assumption we need is that P_{-m} exists and that its elements are o(m).
By constructing a pseudolimit form as in (13) for ψ_0 and inverting (10) we obtain P_0. It is this P_0 that we suggest should be used as initial variance matrix.

The eigenvalues λ_i of T play an important part in determining the form of this P_0. It is easy to show that these eigenvalues are equal to the roots of the vector autoregression corresponding to (1a). Hence, a unit eigenvalue means the process α_t (and by extension, y_t) has a unit root. If we disregard explosive processes for the moment, we see that if all the eigenvalues are less than 1 in absolute value, the matrix P_0 as constructed above is finite. It can be shown that this matrix is just the solution to

$$(14) \quad P_0 = TP_0T' + RR' \qquad \text{or} \qquad vec(P_0) = (I-T\otimes T)^{-1}vec(RR')$$

A second point is that not only do unit eigenvalues cause elements of P_0 to go to infinity, so do two eigenvalue of which the product equals 1. This means that any eigenvalue which lies on the unit circle causes elements of P_0 to go to infinity, since its conjugate is also an eigenvalue and their product is 1.
Finally, if we disregard the approximate solution (13) and stay with (12) it is worth noting that the expression for $\psi_{ij,o}$ is a (left-)continuous function of λ_i. Therefore, the dividing line between stationary and nonstationary models does not lead to discontinuities in $\psi_{ij,0}$.

If T should have repeated eigenvalues we can no longer find a matrix Q so that Λ is diagonal. However, we can then make Λ a Jordan Normal Form. (9) remains valid, but the determination of individual elements of ψ_0 becomes difficult. In such a case, P_0 is perhaps more easily calculated by directly evaluating (7) through repeated substitution in

$$(15) \quad P_i = TP_{i-1}T' + RR'$$

A more tractable case occurs when T has two unit eigenvalues. We can then regard Λ as a Jordan Normal Form again, which means the left upper two-by-two block has 2 possible forms:

$$\begin{matrix} 1 & 1 \\ & & \text{or} \\ 0 & 1 \end{matrix} \qquad \begin{matrix} 1 & 0 \\ \\ 0 & 1 \end{matrix}$$

depending on dim Ker(T-I). The second form is rather uninteresting and leads to conclusions very similar to those for the case where there is only one unit eigenvalue. So we will concentrate on the first form and refer to it as L from now on.

Since

(16) $\quad L^i = \begin{bmatrix} 1 & i \\ 0 & 1 \end{bmatrix}$

we get the following expressions for the relevant elements of Ψ_0 :

(17a) $\quad \psi_{11,0} \quad = \psi_{11,-m} + 2m.\psi_{21,-m} + m^2.\psi_{22,-m} +$

$$\sum_{s=0}^{m-1} (\sigma_{11} + 2s.\sigma_{21} + s^2.\sigma_{22})$$

$$\approx \sigma_{22}m^3/3$$

(17b) $\quad \psi_{21,0} \quad = \psi_{21,-m} + m.\psi_{22,-m} + \sum_{s=0}^{m-1} (\sigma_{21} + s.\sigma_{22})$

$$\approx \sigma_{22}m^2/2$$

(17c) $\quad \psi_{22,0} \quad = \psi_{22,-m} + m.\sigma_{22}$

$$\approx \sigma_{22}m$$

So the crucial elements of Ψ_0 approach infinity at different speeds. Of course, the implications of this for the behaviour of our approximation for P_0 depend on the form of Q.
It is again worth noting that the initial matrix P_{-m} plays a neglegible part in the limit expressions (17a)-(17c). The only requirement we need is that it is o(m).
Cases where T has more than 2 unit roots can be treated in the same fashion.

The method outlined above has one important advantage over $P_0 = k.I$: it produces consistent variance matrices in transformed models. If we should transform (1a)-(1b) by putting $\beta_t = H\alpha_t$, we get the following transformed model:

(1a') $\quad \beta_t = T^*\beta_{t-1} + R^*\eta_t \qquad \eta_t \approx N(0,\sigma^2 I_p)$

(1b') $\quad y_t = Z_t^*\beta_t + S\epsilon_t \qquad \epsilon_t \approx N(0,\sigma^2 I_k)$

with

(18) $\quad T^* = HTH^{-1} \qquad R^* = HR \qquad Z_t^* = Z_t H^{-1}$

It is easy to see that the P_0^* calculated for this model with (7) is related to the P_0 calculated for model (1a)-(1b) through

(19) $\quad P_0^* = HP_0H'$

which is just what we want.

4. Examples

As a first example we will examine a linear regression in which the coefficients follow a random walk in time. The relevant matrices are:

$$T = I \qquad\qquad RR' = diag(q_1, q_2, \ldots q_n)$$

where n is the number of exogenous variables and also the dimension of α_t. The eigenvalues

of T are obvious and no diagonalisation is necessary. Our procedure for calculating P_0 leads to:

$$P_o = \text{diag}(q_1 m, \; q_2 m, \; \ldots .. q_n m)$$

Note that different variances for the various coefficients lead to an initial variance matrix which differs from $k.I$.

The second example to be examined is the trend model (5a)-(5b). The relevant matrices are:

$$T = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \qquad RR' = \begin{bmatrix} 1 & 0 \\ 0 & q \end{bmatrix} \qquad \text{where } q = \text{var}(\epsilon_2)/\text{var}(\epsilon_1)$$

T has two unit eigenvalues, but it is already in Jordan Normal Form, so we can proceed directly to (17a)-(17c) and we get:

$$P_o = q \cdot \begin{bmatrix} m^3/3 & m^2/2 \\ m^2/2 & m \end{bmatrix}$$

We note with satisfaction that level and growth of the trend process are now correlated, and that the level has a considerably greater variance than the growth.

The third example is the seasonal process (5c). The relevant matrices are

$$T = \begin{bmatrix} -1 & -1 & -1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \qquad RR' = \begin{bmatrix} q^* & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \qquad \text{where } q^* = \text{var}(\epsilon_3)/\text{var}(\epsilon_1)$$

T has eigenvalues -1, i and $-i$. We get

$$Q = \begin{bmatrix} 1 & 1 & 1 \\ -1 & -i & i \\ 1 & -1 & -1 \end{bmatrix} \qquad 4Q^{-1} = \begin{bmatrix} 2 & 0 & 2 \\ 1+i & 2i & -1+i \\ 1-i & -2i & -1-i \end{bmatrix}$$

$$\Sigma = q^*/8 \begin{bmatrix} 2 & 1+i & 1-i \\ 1+i & i & 1 \\ 1-i & 1 & -i \end{bmatrix} \qquad \Psi_o = q^*/8 \begin{bmatrix} 2m & . & . \\ . & . & m \\ . & m & . \end{bmatrix}$$

where the dots denote finite elements, neglegible compared to m.
Finally, we get

$$P_o = \tfrac{1}{2} q^* \begin{bmatrix} 2m & -m & . \\ -m & 2m & -m \\ . & -m & 2m \end{bmatrix}$$

The proportional constant $\frac{1}{2}q^*$ seems superfluous at first glance. That it is vital is made clear when the whole model (6a)-(6b) is used as an example. Since T is block diagonal and the eigenvalue analysis can thus be performed per block we can simply combine the results of out second and third examples and we get

$$(20) \qquad P_0 = q \begin{bmatrix} m^3/3 & m^2/2 & 0 & 0 & 0 \\ m^2/2 & m & 0 & 0 & 0 \\ 0 & 0 & 2mu & -mu & 0 \\ 0 & 0 & -mu & 2mu & -mu \\ 0 & 0 & 0 & -mu & 2mu \end{bmatrix}$$

where $u = \frac{1}{2}q^*/q$. The proportionality constant q can be neglected, the variance ration u cannot. As usual, the choice of m is made through a compromise between the urge to have m as large as possible, and the upper limit imposed by machine imprecision.

The question is : does it matter what form P_0 has ? To answer this question data were generated with (5a)-(5d) and these data were then subjected to a Kalman Filter for the model (6a)-(6b). Two possibilities for P_0 were used : $P_0 = k.I$, and P_0 as in (20). For clarity we will refer to these two possibilities as P and P^* from now on. To introduce some synchronisation between P and P^*, k and m were chosen so that $|P| = |P^*|$. The means and standard deviations of various criteria were recorded over 200 replications and the results are given in Table 1.

From the simulation results a number of conclusions can be drawn, although the usual disclaimer that a simulation study as limited as this one cannot be used for indisputable statements applies.

In the first place, the log-likelihoods cannot be meaningfully compared, since the two models with different initial conditions are nonnested. Similarly, any comparisons with regard to estimation criteria must be seen as indicative only.

With this disclaimer firmly in mind, the results do seem to support P^* as initial matrix. The differences are particularly striking for the seasonal term, both regarding the estimation of the seasonal term itself, and of the variance of ϵ_{3t}. The model with P as initial matrix performs poorly in the first three quarters, since the assumed independence of the three seasonal terms in (6a) precludes more accurate estimation of S_2 and S_3 when S_1 has been actually "observed". The model with P^* as initial matrix does take covariances of the seasonal terms into account, and is thus able to predict S_2 and S_3 more accurately once S_1 has been estimated. This difference diminishes when the first few observations are dropped while calculating the likelihood (a procedure suggested by Harvey), but the difference did remain. However, the estimation of the other variance parameters also seems to be more accurate when P^* is taken as initial variance matrix. But it must be stressed that this simulation study is too limited for firm conclusions about small-sample characteristics to be drawn.

5. Conclusion

The exact form of P_0 may seem like a trivial problem. The influence of that choice on subsequent estimates decreases rapidly as t increases, so any mistakes made in specifying P_0 - or a_0 for that matter - will be swept under the ergodic carpet.

Econometricians use the Kalman Filter for the analysis of what a systems theorist will consider to be very small datasets. Therefore, we can not afford to waste the information contained in the first few observations. An inappropriate choice for P_0 will handicap the Kalman Filter in processing the first n observations, with distorted estimates the result. Our method of generating a P_0 is based on a reasonable and extremely weak assumption, namely that at some point in the past, α_{-m} had a variance matrix of which the elements were small compared to m. From that assumption, the exact or approximate form of P_0 can be calculated using the procedure outlined in section 3. By comparison, setting P_0 equal to k I conflicts with virtually all statistical properties of the system (1a)-(1b), and apparently the cost of this is a

lower estimation accuracy. The method is not sensitive to the stationary-nonstationary interface, and, in the limit, is equivalent to calculating the exact unconditional variance matrix for stationary processes.

References

Ansley, C.F. & R. Kohn "Estimation, filtering and smoothing in state space models with incompletely specified initial conditions". *The Annals of Statistics*, vol.13, no.4, April 1985, pp.1286-1316

Den Butter, F.A.G. & T.J. Mourik "The application of the Kalman Filter in seasonal correction" (in Dutch) Research report #8612, De Nederlandse bank N.V., August 1986.

Gehring, M. "Seizoencorrectie met behulp van tijdreeksstructuurmodellen" Doctoral thesis, Eindhoven Technical University, 1987. (with Th.E. Nijman)

Harvey, A.C. *"Time Series Models"*. Philip Allan Publishers Ltd., Oxford, 1981.

Harvey, A.C. "Trend and cycles in macroeconomic time series". *Journal of Business and Economics Statistics*, vol.3, no.3, July 1985, pp.216-227

Harvey, A.C. & P.H.J. Todd "Forecasting economic time series with structural and Box-Jenkins models". *Journal of Business and Economics Statistics* vol.1, no.4, August 1983, pp.299-315.

Kohn, R. & C.F. Ansley "Estimation, prediction and interpolation for ARIMA models with missing data". *Journal of the American Statistical Association*, vol.81, no.395, September 1986, pp.751-761.

Table 1 Simulation results for P and P*

Criterion	$P_0 - P$	$P_0 - P^*$	difference
Log(lik)	-199.8124 (17.4968)	-175.1195 (7.2941)	-24.6930 (16.1005)
RMS seasonal error[3]	2.8556 (1.3945)	1.2823 (0.5974)	1.5733 (1.3701)
σ^2 (true value 1.03)[4]	1.1028 (0.2362)	1.0289 (0.1692)	0.0739 (0.1915)
Ditto, squared error	0.0611 (0.1701)	0.0286 (0.0483)	0.0325 (0.1641)
Var(ϵ_{1t}) (true value 0.01)	0.0405 (0.0895)	0.0496 (0.0806)	-0.0091 (0.0896)
Ditto, squared error	0.0089 (0.0331)	0.0081 (0.0197)	0.0008 (0.0342)
Var(ϵ_{2t}) (true value 0.01)	0.0931 (0.1516)	0.0089 (0.0042)	0.0842 (0.1502)
Ditto, squared error	0.0336 (0.1452)	0.0004 (0.0002)	0.0332 (0.1451)
Var(ϵ_{3t}) (true value 0.01)	0.0636 (0.0546)	0.0033 (0.0031)	0.0603 (0.0541)
Ditto, squared error	0.0059 (0.0094)	0.0001 (0.0000)	0.0058 (0.0098)
Var(ϵ_{4t}) (true value 1.00)	0.9056 (0.2058)	0.9670 (0.1862)	-0.0614 (0.1126)
Ditto, squared error	0.0513 (0.0843)	0.0358 (0.0538)	0.0155 (0.0756)

---

[3] We wished to check how well the seasonal component S_t was estimated in the two models. This criterion compares the estimated S_t to the true S_t.

[4] The scale factor σ^2 was defined as the sum of the variances of all noise processes in the model, to ensure that the variance parameters in R were less than one.