Series in Approximations and Decompositions – Vol. 2

APPROXIMATE KALMAN FILTERING

edited by

Guanrong Chen

University of Houston

World Scientific
Singapore • New Jersey • London • Hong Kong

1993

Donald Catlin
Department of Mathematics and Statistics
University of Massachusetts
Amherst, MA 01003
catlin@math.umass.edu

Initializing the Kalman Filter with Incompletely Specified Initial Conditions

Víctor Gómez and Agustín Maravall

Abstract. We review different approaches to Kalman filtering with incompletely specified initial conditions, appropriate for example when dealing with nonstationarity. We compare in detail the transformation approach and modified Kalman Filter (KF) of Ansley and Kohn, the diffuse likelihood and diffuse KF of de Jong, the approach of Bell and Hillmer, whereby the transformation approach applied to an initial stretch of the data yields initial conditions for the KF, and the approach of Gómez and Maravall, which uses a conditional distribution on initial observations to obtain initial conditions for the KF. It is concluded that the latter approach yields a substantially simpler solution to the problem of optimal estimation, forecasting and interpolation for a fairly general class of models.

§1 Introduction

We consider observations generated by a discrete time state space model (SSM) such that the initial state vector $\mathbf{x}_0$ has a distribution which is unspecified. We will further allow for unknown regression type parameters. Examples are non-stationary time series which follow an ARIMA model, regression models with ARIMA disturbances, structural models (as in [9]) and ARIMA component models, among others. In all these cases it is not possible to initialize the Kalman Filter (KF) as usual, by means of the first two moments of the distribution of $\mathbf{x}_0$, because they are not well defined. Therefore, it is necessary to incorporate new assumptions in order to deal with this initialization problem.

Among the different alternatives that have been proposed in the literature, we will focus on the transformation approach of Kohn and Ansley, the diffuse Kalman filter (DKF) of de Jong, the initialization procedure of Bell and Hillmer and the approach of Gómez and Maravall, based on a trivial extension of the KF, to be denoted

Approximate Kalman Filtering
Guanrong Chen (ed.), pp. 39–62.

the Extended Kalman Filter (EKF), with a distribution defined conditionally on the initial observations. There are other approaches as well, like the so-called "big k" method (see, for example, [5] and [7]). This method uses a matrix of the form kI to initialize the state covariance matrix, where k is large to reflect uncertainty regarding the initial state and I is the identity matrix. The big k method is not only numerically dangerous, it is also inexact. An alternative to the big k method is to use the information filter (see [1]). However, as seen in [2], the information filter breaks down in many important cases, including ARMA models.

The paper is structured as follows. In Section 2 we will define the SSM and consider some illustrative examples. In Section 3 we suppose that the initial state vector x_0 is fixed, define the likelihood and show how the EKF and the DKF can be used to evaluate it. In Section 4 we will deal with the different approaches to define and evaluate the likelihood of the SSM in the case when there are no regression type parameters and the initial state vector has an unspecified distribution. In Section 5 we will extend these results to include regression type parameters.

§2 State space model

Definition 1. A vectorial time series $\mathrm{v} = (\mathrm{v}_1^\top, \cdots, \mathrm{v}_N^\top)^\top$ is said to be generated by the State Space Model (SSM) if, for $k = 1, \cdots, N$,

$$\begin{cases} \mathrm{v}_k = X_k\underline{\beta} + C_k\mathrm{x}_k + Z_k\underline{\xi}_k\,, \\ \mathrm{x}_k = W_{k-1}\underline{\beta} + A_{k-1}\mathrm{x}_{k-1} + H_{k-1}\underline{\xi}_{k-1}\,, \end{cases} \tag{1}$$

where $\mathrm{x}_0 = B\underline{\delta}, \underline{\xi}_k \sim N_{iid}(0, \sigma^2 I), k = 0, \cdots, N, \underline{\delta} \sim N(\mathrm{c}, \sigma^2 C)$, with either C nonsingular or $C = 0$, $\underline{\delta}$ and $\underline{\xi} = (\underline{\xi}_0^\top, \cdots, \underline{\xi}_N^\top)^\top$ are independent, B is of full column rank and $\underline{\beta}$ is a vector of fixed regression parameters. Also, $Var(\mathrm{v})$ is nonsingular if $C = 0$.

This definition is similar to the one in [13]; the vector $\underline{\delta}$ models uncertainty with respect to the initial conditions. Following [13], we will say that $\underline{\delta}$ is diffuse if C^{-1} is arbitrarily close to 0 in the Euclidean norm, denoted $C \to \infty$. Contrary to de Jong, we will always suppose that $\underline{\beta}$, the vector of regression parameters, is fixed; considering $\underline{\beta}$ diffuse introduces confusion as to what likelihood should be used and it affects neither the equations nor the computations with the DKF, to be defined below.

The formulation we use for the SSM has the virtue of explicitly separating the time-invariant "mean" effect $\underline{\beta}$ from the state vector x_k, keeping its dimension to a minimum. Choosing adequately the matrices X_k, W_k, H_k and Z_k, appropriate components of $\underline{\beta}$ and $\underline{\xi}_k$ can be excluded from or included in each equation. Thus, the specification covers the case where the mean and disturbance effects in each equation are distinct. Two simple examples will illustrate the definition.

Example 1. Suppose a regression model with random walk disturbance and scalar v_k,

$$\nabla(\mathrm{v}_k - \mathrm{y}_k^\top \underline{\beta}) = \mathrm{a}_k, \tag{2}$$

where $\nabla = 1 - L$, L is the lag operator $(L(\mathrm{v}_k) = \mathrm{v}_{k-1})$, and all the $\mathrm{a}_k \sim N(0, \sigma^2)$ are independent. Model (2) can be put into a state space form by defining $X_k = \mathrm{y}_k^\top, C_k = 1, Z_k = 0, W_k = 0, A_k = 1, H_k = 1, \mathrm{x}_k = \mathrm{v}_k - \mathrm{y}_k^\top \underline{\beta}$ and $\underline{\xi}_{k-1} = \mathrm{a}_k$. That is,

$$\mathrm{x}_k = \mathrm{x}_{k-1} + \mathrm{a}_k\,, \tag{3a}$$
$$\mathrm{v}_k = \mathrm{y}_k^\top \underline{\beta} + \mathrm{x}_k\,. \tag{3b}$$

For initialization, we make $A_0 = 1, H_0 = 1, W_0 = 0, B = 1$ and $\mathrm{x}_0 = \underline{\delta}$. Therefore, the first state is $\mathrm{x}_1 = \underline{\delta} + \mathrm{a}_1$ and $\underline{\delta}$ is in this case equal to the initial state. Because $\{\mathrm{x}_k\}$ follows the non-stationary model ($3a$), the distribution of $\underline{\delta}$ is unspecified.

Example 2. Suppose Example 1, but with ∇ replaced by $1 - \rho L$, where $|\rho| < 1$. Then, we have a regression model with AR(1) disturbances. The SSM is

$$\mathrm{x}_k = \rho \mathrm{x}_{k-1} + \mathrm{a}_k \tag{4}$$

and ($3b$). For initialization, we make $A_0 = 1, H_0 = 1/\sqrt{1-\rho^2}, W_0 = 0, B = 1$ and $\mathrm{x}_0 = 0$ ($\mathrm{c} = 0, C = 0$). In this case $\{\mathrm{x}_k\}$, follows the stationary model (4) and we can use the first two moments of x_k, namely $E\{\mathrm{x}_k\} = 0$ and $Var(\mathrm{x}_k) = \sigma^2/(1-\rho^2)$, to set up the initial conditions. The first state is $\mathrm{x}_1 = (1/\sqrt{1-\rho^2})\mathrm{a}_1$.

A representation which will be very useful in what follows is given by the next theorem.

Theorem 1. *If* $\mathrm{v} = (\mathrm{v}_1^\top, \cdots, \mathrm{v}_N^\top)^\top$ *is generated by the SSM (1), then* $\mathrm{v} = R\underline{\delta} + S\underline{\beta} + \underline{\varepsilon}$, *where the rows of* S *are*

$$\begin{aligned} S_1 &= X_1 + C_1 W_0\,, \\ S_2 &= X_2 + C_2(W_1 + A_1 W_0)\,, \\ &\vdots \\ S_N &= X_N + C_N\{W_{N-1} + A_{N-1}W_{N-2} + \cdots + (A_{N-1} \cdots A_1)W_0\}\,, \end{aligned}$$

and those of R *are*

$$R_i = C_i A_{i-1} \cdots A_0 B, \quad i = 1, \cdots, N.$$

Besides, $\underline{\varepsilon} \sim N(0, \sigma^2\Sigma)$ *with* Σ *nonsingular and* $Cov(\underline{\delta}, \underline{\varepsilon}) = 0$.

Proof: The expressions for S_i and R_i are obtained by repeated substitutions using (1). The vectors $\underline{\varepsilon}_i$ are linear combinations of $\underline{\xi}_0, \underline{\xi}_1, \cdots, \underline{\xi}_i$, $i = 1, \cdots, N$.

§3 Fixed initial state

If $\underline{\delta}$ is fixed ($C = 0$), then $\underline{\delta} = \mathrm{c}$ and the representation of Theorem 1, $\mathrm{v} = R\underline{\delta} + S\underline{\beta} + \underline{\varepsilon}$, constitutes a regression model where the distribution of $\underline{\varepsilon}$ is known. If we define $X = (R, S)$ and $\underline{\gamma} = (\underline{\delta}^\top, \underline{\beta}^\top)^\top$, then the log-likelihood of this model, based on v, is (throughout the paper all log-likelihoods will be defined up to an additive constant)

$$\lambda(\mathrm{v}) = -\frac{1}{2}\{M\ln(\sigma^2) + \ln|\Sigma| + (\mathrm{v} - X\underline{\gamma})^\top\Sigma^{-1}(\mathrm{v} - X\underline{\gamma})/\sigma^2\},$$

where $Var(\mathrm{v}) = \sigma^2\Sigma$ and M denotes the number of components in v, the vector of stacked observations. The maximum likelihood estimator of σ^2 is

$$\hat{\sigma}^2 = (\mathrm{v} - X\underline{\gamma})^\top\Sigma^{-1}(\mathrm{v} - X\underline{\gamma})/M\,. \tag{5}$$

Substituting $\hat{\sigma}^2$ back in $\lambda(\mathrm{v})$ yields the σ^2-maximized log-likelihood:

$$l(\mathrm{v}) = -\frac{1}{2}\{M\ln(\hat{\sigma}^2) + \ln|\Sigma|\}\,.$$

It turns out that we can evaluate $l(\mathrm{v})$ efficiently using the KF.

Definition 2. The Kalman Filter (KF) is the set of recursions

$$\begin{cases} \mathrm{e}_k = \mathrm{v}_k - X_k\underline{\beta} - C_k\hat{\mathrm{x}}_{k,k-1}, \\ D_k = C_kP_{k,k-1}C_k^\top + Z_kZ_k^\top, \\ G_k = (A_kP_{k,k-1}C_k^\top + H_kZ_k^\top)D_k^{-1}, \\ \hat{\mathrm{x}}_{k+1,k} = W_k\underline{\beta} + A_k\hat{\mathrm{x}}_{k,k-1} + G_k\mathrm{e}_k, \\ P_{k+1,k} = (A_k - G_kC_k)P_{k,k-1}A_k^\top + (H_k - G_kZ_k)H_k^\top, \end{cases} \tag{6}$$

with starting conditions $\hat{\mathrm{x}}_{1,0} = W_0\underline{\beta} + A_0B\underline{\delta}$ and $P_{1,0} = H_0H_0^\top$.

Here $\hat{\mathrm{x}}_{k,k-1}$ is the predictor of x_k using $(\mathrm{v}_1^\top, \cdots, \mathrm{v}_{k-1}^\top)^\top$ and $Var(\hat{\mathrm{x}}_{k,k-1} - \mathrm{x}_{k,k-1}) = P_{k,k-1}$. The e_k are the errors of predicting v_k using $(\mathrm{v}_1^\top, \cdots, \mathrm{v}_{k-1}^\top)^\top$. They constitute an orhogonal sequence with $E\{\mathrm{e}_k\} = 0$ and $Var(\mathrm{e}_k) = D_k$, as given in (6). Note that we have supposed $\sigma^2 = 1$ in the equations (6) because we will estimate σ^2 using (5). It can be shown (see, for example, [13]) that if $\mathrm{e} = (\mathrm{e}_1^\top, \cdots, \mathrm{e}_N^\top)^\top$, then there exists a lower triangular matrix K with ones in the main diagonal such that $\mathrm{e} = K(\mathrm{v} - X\underline{\gamma})$ and $K\Sigma K^\top = D = \mathrm{diag}(D_1, D_2, \cdots, D_N)$. Therefore, $\Sigma^{-1} = K^\top D^{-1}K$ and $\hat{\sigma}^2 = \mathrm{e}^\top D^{-1}\mathrm{e}/M$, $\ln|\Sigma| = \ln|D_1| + \ln|D_2| + \cdots + \ln|D_N|$. In the case of scalar v_k, the D_k are also scalar and we can obtain a "square root" of Σ^{-1} by putting $\Sigma^{-1/2} = D^{-1/2}K$. Then, we can use a vector of standardized residuals $\tilde{\mathrm{e}} = D^{-1/2}\mathrm{e}$ such that $\tilde{\sigma}^2 = \tilde{\mathrm{e}}^{\top}\tilde{\mathrm{e}}/M$ and maximizing $l(\mathrm{v})$ becomes equivalent to minimizing the non-linear sum of squares

$$S(\mathrm{v},\underline{\gamma}) = \left(\prod_{k=1}^{N} |D_k^{1/2}|\right)^{1/M} \tilde{\mathrm{e}}^{\top}\tilde{\mathrm{e}} \left(\prod_{k=1}^{N} |D_k^{1/2}|\right)^{1/M} .$$

For vectorial v_k, suppose that in each step of the KF, in addition to D_k, we also obtain, by means of its Cholesky decomposition, a "square root" $D_k^{1/2}$ of $D_k, k = 1, \cdots, N$. Then, the matrix $D^{1/2} = \mathrm{diag}(D_1^{1/2}, D_2^{1/2}, \cdots, D_N^{1/2})$ is a "square root" of D and we can proceed as in the scalar case to evaluate $S(\mathrm{v},\underline{\gamma})$.

Example 2. (Continued) In this case $\underline{\delta} = 0$, so that, by Theorem 1, we have $\mathrm{v} = S\underline{\beta} + \underline{\varepsilon}$. The initial conditions for the KF are $\hat{\mathrm{x}}_{1,0} = 0$ and $P_{1,0} = 1/(1-\rho^2)$. Then, the KF gives

$$\begin{cases} \mathrm{e}_k = \mathrm{v}_k - X_k\underline{\beta} - \hat{\mathrm{x}}_{k,k-1}, \quad D_1 = 1/(1-\rho^2)\,, \quad D_k = 1, \quad k > 1\,, \\ G_k = \rho, \quad \hat{\mathrm{x}}_{k+1,k} = \rho\hat{\mathrm{x}}_{k,k-1} + \rho\mathrm{e}_k\,, \\ P_{1,0} = 1/(1-\rho^2), \quad P_{k+1,k} = 1, \quad k > 0\,. \end{cases}$$

The vector of residuals is $\mathbf{e} = K(\mathrm{v} - S\underline{\beta})$, where

$$K = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ -\rho & 1 & 0 & \cdots & 0 & 0 \\ 0 & -\rho & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -\rho & 1 \end{bmatrix},$$

and the vector of standardized residuals is $\tilde{\mathrm{e}} = D^{-1/2}\mathrm{e}$, with $\tilde{\mathrm{e}}_1 = (\mathrm{v}_1 - \mathrm{y}_1^{\top}\underline{\beta})\sqrt{1-\rho^2}$ and $\tilde{\mathrm{e}}_k = (\mathrm{v}_k - \mathrm{y}_k^{\top}\underline{\beta}) - \rho(\mathrm{v}_{k-1} - \mathrm{y}_{k-1}^{\top}\underline{\beta}),\ k > 1$. The nonlinear sum of squares is

$$S(\mathrm{v},\underline{\gamma}) = (1/\sqrt{1-\rho^2})^{1/N}\tilde{\mathrm{e}}^{\top}\tilde{\mathrm{e}}(1/\sqrt{1-\rho^2})^{1/N} .$$

We now return to the general discussion of the term $S(\mathrm{v},\underline{\gamma})$. We can concentrate $\underline{\gamma}$ out of $S(\mathrm{v},\underline{\gamma})$ if we replace $\underline{\gamma}$ in $S(\mathrm{v},\underline{\gamma})$ by its maximum likelihood estimator $\hat{\underline{\gamma}}$, which is the generalized least squares (GLS) estimator of the model

$$\mathrm{v} = X\underline{\gamma} + \underline{\varepsilon}. \tag{7}$$

We next show how to obtain $\hat{\underline{\gamma}}$ by means of the EKF. From what we have just seen, it is clear that the KF can be seen as an algorithm that, applied to a vector $\underline{\nu}$ of the same dimension as v, yields $K\underline{\nu}$ and D. The algorithm can be trivially extended to compute also $D^{1/2}$. Therefore, if we apply this extended algorithm in model (7) to the data v and to the columns of the X matrix, we obtain $D^{-1/2}K\mathrm{v} = D^{-1/2}KX\underline{\gamma}+D^{-1/2}K\underline{\varepsilon}$, where $Var(D^{-1/2}K\underline{\varepsilon}) = \sigma^2 I_M$, and we have transformed a GLS regression model (7) into an ordinary least squares (OLS) one. The estimator $\hat{\underline{\gamma}}$ can now be efficiently and accurately obtained using the QR algorithm. Supposing X is of full column rank, if p is the number of components in $\underline{\gamma}$, this last algorithm premultiplies both $D^{-1/2}K\mathrm{v}$ and $D^{-1/2}KX$ by an orthogonal matrix Q to obtain $\underline{\omega} = QD^{-1/2}K\mathrm{v}$ and $(U^\top, 0^\top)^\top = QD^{-1/2}KX$, where U is a nonsingular $p \times p$ upper triangular matrix. Then, $\hat{\underline{\gamma}} = U^{-1}\underline{\omega}_1$ where $\underline{\omega}_1$ consists of the first p elements of $\underline{\omega}$, and we can evaluate

$$S(\mathrm{v},\hat{\underline{\gamma}}) = \left(\prod_{k=1}^{N} |D_k^{1/2}|\right)^{1/M} \underline{\omega}_2^\top\underline{\omega}_2 \left(\prod_{k=1}^{N} |D_k^{1/2}|\right)^{1/M},$$

where $\underline{\omega}_2$ consists of the last $M-p$ elements of $\underline{\omega}$.

Definition 3. The Extended Kalman Filter (EKF) is the KF (6) with the equations for e_k and $\hat{\mathrm{x}}_{k+1,k}$, respectively, replaced by

$$E_k = (\mathrm{v}_k, 0, X_k) - C_k\hat{\mathrm{X}}_{k,k-1}, \quad \hat{\mathrm{X}}_{k+1,k} = (0,0,-W_k) + A_k\hat{\mathrm{X}}_{k,k-1} + G_k E_k\,,$$

with the starting condition $\hat{\mathrm{X}}_{1,0} = (0, -A_0 B, -W_0)$. Also, $D_k^{1/2}$ is computed along with D_k.

The columns of the matrix $\hat{\mathrm{X}}_{k,k-1}$ contain the state estimates, and those of E_k the prediction errors, corresponding to the data and to the columns of the X matrix, respectively. The EKF has been suggested in [15] and [19]; it has also been generalized to the case of a rank deficient X matrix in [6].

Example 2. (Continued) Applying the EKF with the starting condition $\hat{\mathrm{X}}_{1,0} = (0,0)$, we get

$$E_k = (\mathrm{v}_k, \mathrm{y}_k^\top) - \hat{\mathrm{X}}_{k,k-1}, \quad \tilde{E}_k = D_k^{-1/2}E_k, \quad \hat{\mathrm{X}}_{k+1,k} = \rho\hat{\mathrm{X}}_{k,k-1} + \rho E_k\,.$$

This implies $E_k = (\mathrm{v}_k - \rho\mathrm{v}_{k-1}, \mathrm{y}_k^\top - \rho\mathrm{y}_{k-1}^\top)$, $k > 1$, and $E_1 = (\mathrm{v}_1, \mathrm{y}_1^\top)$. The GLS model $\mathrm{v} = S\underline{\beta} + \underline{\varepsilon}$ has been transformed into the OLS model

$$\begin{bmatrix} \mathrm{v}_1\sqrt{1-\rho^2} \\ \mathrm{v}_2 - \rho\mathrm{v}_1 \\ \vdots \\ \mathrm{v}_N - \rho\mathrm{v}_{N-1} \end{bmatrix} = \begin{bmatrix} \mathrm{y}_1^\top\sqrt{1-\rho^2} \\ \mathrm{y}_2^\top - \rho\mathrm{y}_1^\top \\ \vdots \\ \mathrm{y}_N^\top - \rho\mathrm{y}_{N-1}^\top \end{bmatrix} \underline{\beta} + \tilde{\underline{\varepsilon}}, \quad \tilde{\underline{\varepsilon}} \sim N(0, \sigma^2 I)\,.$$

Consider next predicting the state x_k using $(\mathrm{v}_1^\top, \mathrm{v}_2^\top, \cdots, \mathrm{v}_{k-1}^\top)^\top$. This is equivalent to first predicting x_k using $(\underline{\gamma}^\top, \mathrm{v}_1^\top, \mathrm{v}_2^\top, \cdots, \mathrm{v}_{k-1}^\top)^\top$, and then predicting this predictor using $(\mathrm{v}_1^\top, \mathrm{v}_2^\top, \cdots, \mathrm{v}_{k-1}^\top)^\top$. The first mentioned predictor is $\hat{\mathrm{x}}_{k,k-1}$, as given in (6). It is easy to check that $\hat{\mathrm{X}}_{k,k-1}(1, -\underline{\gamma}^\top)^\top$, where $\hat{\mathrm{X}}_{k,k-1}$ is given by the EKF, verifies the same recursion and starting condition as $\hat{\mathrm{x}}_{k,k-1}$, and hence $\hat{\mathrm{x}}_{k,k-1} = \hat{\mathrm{X}}_{k,k-1}(1, -\underline{\gamma}^\top)^\top$. Thus, $\tilde{\mathrm{x}}_{k,k-1} = \hat{\mathrm{X}}_{k,k-1}(1, -\hat{\underline{\gamma}}^\top)^\top$ is the predictor we are looking for. Its mean squared error (Mse) is

$$\begin{aligned}
\mathrm{Mse}(\tilde{\mathrm{x}}_{k,k-1}) &= Var(\mathrm{x}_k - \hat{\mathrm{x}}_{k,k-1} + \hat{\mathrm{x}}_{k,k-1} - \tilde{\mathrm{x}}_{k,k-1} \\
&= Var(\mathrm{x}_k - \hat{\mathrm{x}}_{k,k-1}) + Var(\hat{\mathrm{x}}_{k,k-1} - \tilde{\mathrm{x}}_{k,k-1}) \\
&= \sigma^2 P_{k,k-1} + Var\left(\hat{\mathrm{X}}_{k,k-1}(0, (\hat{\underline{\gamma}} - \underline{\gamma})^\top)^\top\right) \\
&= \sigma^2 P_{k,k-1} + \hat{\mathrm{X}}(\underline{\gamma})_{k,k-1}\mathrm{Mse}(\hat{\underline{\gamma}})\hat{\mathrm{X}}(\underline{\gamma})_{k,k-1}^\top ,
\end{aligned}$$

where $\hat{\mathrm{X}}(\underline{\gamma})_{k,k-1}$ is the submatrix of $\hat{\mathrm{X}}_{k,k-1}$ formed by all its columns except the first, and $\mathrm{Mse}(\hat{\underline{\gamma}}) = \sigma^2(U^\top U)^{-1}$. If $\tilde{\mathrm{v}}_k$ is the predictor of v_k using $(\mathrm{v}_1^\top, \mathrm{v}_2^\top, \cdots, \mathrm{v}_{k-1}^\top)^\top$, then it can be shown analogously that $\mathrm{v}_k - \tilde{\mathrm{v}}_k = E_k(1, -\hat{\underline{\gamma}}^\top)^\top$, $\mathrm{Mse}(\tilde{\mathrm{v}}_k) = \sigma^2 D_k + E(\underline{\gamma})_k \mathrm{Mse}(\hat{\underline{\gamma}}) E(\underline{\gamma})_k^\top$, where $E(\underline{\gamma})_k$ is the submatrix of E_k formed by all its columns except the first.

The DKF of de Jong can also be used for likelihood evaluation when $C = 0$, although, as we will see, it has other uses as well.

Definition 4. The Diffuse Kalman Filter (DKF) is the EKF without the computation of $D_k^{1/2}$ and with the added recursion $Q_{k+1} = Q_k + E_k^\top D_k^{-1} E_k$, where $Q_1 = 0$.

Given that $(\mathrm{v} - X\hat{\underline{\gamma}})^\top \Sigma^{-1}(\mathrm{v} - X\hat{\underline{\gamma}}) = q - \mathrm{s}^\top S^{-1}\mathrm{s}$, where $q = \mathrm{v}^\top \Sigma^{-1}\mathrm{v}, \mathrm{s} = X^\top \Sigma^{-1}\mathrm{v}$, and $S = X^\top \Sigma^{-1} X$, the Q_k matrix accumulates the partial squares and cross products and

$$Q_{N+1} = \begin{bmatrix} q & \mathrm{s}^\top \\ \mathrm{s} & S \end{bmatrix} . \tag{8}$$

Therefore, the DKF allows us to evaluate the $(\sigma^2, \underline{\gamma})$-maximized log-likelihood, given by

$$-\frac{1}{2}\left\{ M \ln((q - \mathrm{s}^\top S^{-1}\mathrm{s})/M) + \sum_{k=1}^{N} \ln|D_k| \right\} .$$

Example 2. (Continued) The DKF gives, besides E_k and $\hat{\mathrm{X}}_{k+1,k}$, computed as in the EKF, the matrices Q_k. In this case,

$$Q_{N+1} = \begin{bmatrix} (1-\rho^2)\mathrm{v}_1^2 + \sum_{k=2}^N (\mathrm{v}_k - \rho \mathrm{v}_{k-1})^2 & * \\ (1-\rho^2)\mathrm{v}_1 \mathrm{y}_1 + \sum_{k=2}^N (\mathrm{v}_k - \rho \mathrm{v}_{k-1})(\mathrm{y}_k - \rho \mathrm{y}_{k-1}) & (1-\rho^2)\mathrm{y}_1 \mathrm{y}_1^\top + \sum_{k=2}^N (\mathrm{y}_k - \rho \mathrm{y}_{k-1})(\mathrm{y}_k - \rho \mathrm{y}_{k-1})^\top \end{bmatrix} .$$

Finally in this section, we remark that the estimator $\hat{\underline{\gamma}}$ is obtained by solving the normal equations of the regression, $S\underline{\gamma} = s\mathrm{v}$. However, solving the normal equations this way can lead to numerical difficulties because what we are doing is basically squaring a number and then taking its square root. It is numerically more efficient to use a device such as the QR algorithm or the singular value decomposition, once the EKF has been applied. Another alternative, but computationally more expensive, is to use a square root filter version of the DKF.

§4 Initial state with an unspecified distribution: no regression parameters

In this Section, we suppose that $\underline{\delta}$ in $\mathrm{x}_0 = B\underline{\delta}$ has an unspecified distribution, that is $\underline{\delta} \sim N(\mathrm{c}, \sigma^2 C)$ with C nonsingular. We also suppose that there are no regression parameters and, therefore, $W_k = 0$, and $X_k = 0$. Then, Theorem 1 implies

$$\mathrm{v} = R\underline{\delta} + \underline{\varepsilon}. \tag{9}$$

Ansley and Kohn [2], hereafter AK, define the likelihood of (9) by means of a transformation of the data that eliminates dependence on initial conditions. Let J be a matrix with $|J| = 1$ such that JR has exactly rank(R) rows different from zero. Such a matrix always exists. Let J_1 consist of those rank(R) rows of J corresponding to the nonzero rows of JR and let J_2 consist of the other rows of J so that $J_2 R = 0$. AK define the likelihood of (9) as the density of $J_2\mathrm{v}$. We will show later that, under an extra assumption, this definition does not depend on the matrix J. To evaluate the likelihood, however, and merely for algorithmic purposes, given that the transformation usually destroys the covariance structure of the data, they use an equivalent definition of the likelihood and develop what they call "modified Kalman Filter" and "modified Fixed Point Smoother" algorithms. The modified Kalman Filter is of considerable complexity, difficult to program and is less computationally efficient than the procedure in [6], when applicable, or the DKF. Also, it does not explicitly handle fixed effects and requires specialized assumptions regarding the SSM (see [14]).

Another approach to defining the likelihood of (9) is that of de Jong [13], where $\underline{\delta}$ is considered diffuse by letting $C \to \infty$. In order to take this limit we need the following theorem.

Theorem 2. *Let $\underline{\delta} \sim N(\mathrm{c}, \sigma^2 C)$ with C nonsingular. Then, the log-likelihood of* v *is*

$$\begin{aligned}\lambda(\mathrm{v}) = -\frac{1}{2}\Big\{&\ln|C| + \ln|\sigma^2\Sigma| + \ln|C^{-1} + R^\top\Sigma^{-1}R| \\ &+ \{(\tilde{\underline{\delta}} - \mathrm{c})^\top C^{-1}(\tilde{\underline{\delta}} - \mathrm{c}) + (\mathrm{v} - R\tilde{\underline{\delta}})^\top\Sigma^{-1}(\mathrm{v} - R\tilde{\underline{\delta}})\}/\sigma^2\Big\},\end{aligned}$$

where

$$\tilde{\underline{\delta}} = (C^{-1} + R^\top\Sigma^{-1}R)^{-1}(C^{-1}\mathrm{c} + R^\top\Sigma^{-1}\mathrm{v})$$

and $\tilde{\underline{\delta}}$ coincides with the conditional expectation $E\{\underline{\delta}|\mathrm{v}\}$. Also,

$$\mathrm{Mse}(\tilde{\underline{\delta}}) = Var(\underline{\delta}|\mathrm{v}) = \sigma^2(C^{-1} + R^{\top}\Sigma^{-1}R)^{-1}.$$

Proof: The density $p(\mathrm{v})$ verifies $p(\underline{\delta}|\mathrm{v})p(\mathrm{v}) = p(\mathrm{v}|\underline{\delta})p(\underline{\delta})$, where the vertical bar denotes conditional distribution. The maximum likelihood estimator $\tilde{\underline{\delta}}$ of $\underline{\delta}$ on the left hand side of this equation must be equal to the one on the right hand side. Given that the equality between densities implies

$$\begin{aligned}&(\underline{\delta} - E\{\underline{\delta}|\mathrm{v}\})^{\top}\Omega_{\underline{\delta}|\mathrm{v}}^{-1}(\underline{\delta} - E\{\underline{\delta}|\mathrm{v}\}) + (\mathrm{v} - R\mathrm{c})^{\top}\Omega_{\mathrm{v}}^{-1}(\mathrm{v} - R\mathrm{c})\\ &= (\underline{\delta} - \mathrm{c})^{\top}C^{-1}(\underline{\delta} - \mathrm{c}) + (\mathrm{v} - R\underline{\delta})^{\top}\Sigma^{-1}(\mathrm{v} - R\underline{\delta}),\end{aligned}$$

where $\Omega_{\underline{\delta}|\mathrm{v}}$ and Ω_{v} are the covariance matrices, divided by σ^2, of $p(\underline{\delta}|\mathrm{v})$ and $p(\mathrm{v})$, respectively, the left hand side is minimized for $\tilde{\underline{\delta}} = E\{\underline{\delta}|\mathrm{v}\}$. To minimize the right hand side, consider the regression model

$$(\mathrm{c}^{\top}, \mathrm{v}^{\top})^{\top} = (I, R^{\top})^{\top}\underline{\delta} + \underline{\nu}, \quad \underline{\nu} \sim N(0, \mathrm{diag}(C, \Sigma)).$$

Then $\tilde{\underline{\delta}}$ is as asserted and

$$Var(\underline{\delta}|\mathrm{v}) = Var(\tilde{\underline{\delta}}) = \sigma^2(C^{-1} + R^{\top}\Sigma^{-1}R)^{-1}.$$

Theorem 3. *With the notation and assumption of Theorem 2, if $R^{\top}\Sigma^{-1}R$ is nonsingular, then, letting $C \to \infty$, we have*

$$\begin{aligned}\lambda(\mathrm{v}) + \frac{1}{2}\ln|C| &\to -\frac{1}{2}\{\ln|\sigma^2\Sigma| + \ln|R^{\top}\Sigma^{-1}R| + (\mathrm{v} - R\hat{\underline{\delta}})^{\top}\Sigma^{-1}(\mathrm{v} - R\hat{\underline{\delta}})/\sigma^2\},\\ \tilde{\underline{\delta}} \to \hat{\underline{\delta}} &= (R^{\top}\Sigma^{-1}R)^{-1}R^{\top}\Sigma^{-1}\mathrm{v}, \quad \mathrm{Mse}(\tilde{\underline{\delta}}) \to \mathrm{Mse}(\hat{\underline{\delta}}) = \sigma^2(R^{\top}\Sigma^{-1}R)^{-1}.\end{aligned}$$

Proof: It is an immediate consequence of Theorem 2.

It is shown in [13] that $\lambda(\mathrm{v}) + (1/2)\ln|C|$ tends to a proper log-likelihood, called the diffuse log-likelihood. By Theorem 3, in order to compute it, all we have to do is to consider $\underline{\delta}$ fixed in (9) and apply the methodology of Section 3. If the EKF is used, then, with the notation of Section 3, the results of Theorem 3 can be rewritten

$$\begin{aligned}\lambda(\mathrm{v}) + \frac{1}{2}\ln|C| &\to -\frac{1}{2}\left[M\ln(\sigma^2) + 2\left\{\sum_{k=1}^{N}\ln|D_k^{1/2}| + \ln|U|\right\} + \underline{\omega}_2^{\top}\underline{\omega}_2/\sigma^2\right],\\ \tilde{\underline{\delta}} \to \hat{\underline{\delta}} &= U^{-1}\underline{\omega}_1, \quad \mathrm{Mse}(\tilde{\underline{\delta}}) \to \mathrm{Mse}(\hat{\underline{\delta}}) = \sigma^2(U^{\top}U)^{-1},\end{aligned}$$

whereas if the DKF is used, then, with the notation of (8), we obtain

$$\begin{aligned}\lambda(\mathrm{v}) + \frac{1}{2}\ln|C| &\to -\frac{1}{2}\left[M\ln(\sigma^2) + \sum_{k=1}^{N}\ln|D_k| + \ln|S| + (q - \mathrm{s}^{\top}S^{-1}\mathrm{s})/\sigma^2\right],\\ \tilde{\underline{\delta}} \to \hat{\underline{\delta}} &= S^{-1}\mathrm{s}, \quad \mathrm{Mse}(\tilde{\underline{\delta}}) \to \mathrm{Mse}(\hat{\underline{\delta}}) = \sigma^2 S^{-1}.\end{aligned}$$

If S in (8) is singular, de Jong leaves the diffuse log-likelihood undefined. In order to define the limiting expressions of Theorem 3 when S is singular, we have to consider model (9) with an R matrix that is not of full column rank. Let K be a selector matrix formed by zeros and ones such that $KSK^{\top}$ has a rank equal to rank(R) and replace model (9) by

$$\mathrm{v} = RK^{\top}\underline{\delta}_1 + \underline{\varepsilon}, \tag{10}$$

where $\underline{\delta}_1 \sim N(\mathrm{c}, \sigma^2 C)$, with C nonsingular and $\underline{\delta}_1$ is the vector formed by choosing those components in $\underline{\delta}$ corresponding to the selected columns $RK^{\top}$. This amounts to making the assumption that the other components in $\underline{\delta}$ cannot be estimated from the data without further information and are assigned value zero with probability one. The next theorem generalizes the results of Theorem 2 to the case of a possibly singular S matrix.

Theorem 4. *Suppose model (10) with the convention that if R is of full column rank, then matrix K is the identity matrix and $\underline{\delta}_1 = \underline{\delta}$. Then, with the notation and assumptions of Theorem 3, letting $C \to \infty$, we have*

$$\lambda(\mathrm{v}) + \frac{1}{2}\ln|C| \to -\frac{1}{2}\{\ln|\sigma^2\Sigma| + \ln|KR^{\top}\Sigma^{-1}RK^{\top}| + (\mathrm{v} - R\hat{\underline{\delta}})^{\top}\Sigma^{-1}(\mathrm{v} - R\hat{\underline{\delta}})/\sigma^2\},$$
$$\bar{\underline{\delta}} \to \hat{\underline{\delta}} = (R^{\top}\Sigma^{-1}R)^{-}R^{\top}\Sigma^{-1}\mathrm{v}, \quad \mathrm{Mse}(\bar{\underline{\delta}}) \to \mathrm{Mse}(\hat{\underline{\delta}}) = \sigma^2(R^{\top}\Sigma^{-1}R)^{-},$$

where $(R^{\top}\Sigma^{-1}R)^{-} = K^{\top}(K(R^{\top}\Sigma^{-1}R)K^{\top})^{-1}K$ and $\bar{\underline{\delta}}$ and $\hat{\underline{\delta}}$ are interpreted as the particular maximizers obtained by making zero the elements not in $\bar{\underline{\delta}}_1$ and $\hat{\underline{\delta}}_1$, respectively.

Proof: The only thing that needs to be proved is that $|KR^{\top}\Sigma^{-1}RK^{\top}|$ does not depend on K. This can be seen in [18, page 527].

The next theorem shows the relationship between the likelihood of AK and the diffuse likelihood of de Jong. When S is singular, we take as diffuse log-likelihood the one given by Theorem 4.

Theorem 5. *Let J be a matrix with $|J| = 1$ like those used by AK to define their likelihood, and let J_1 and J_2 be the corresponding submatrices such that $J_1R \neq 0$ and $J_2R = 0$. If $p(\mathrm{v})$ is the density of v when C is nonsingular, as given by Theorem 2, and $p(J_2\mathrm{v})$ is the AK likelihood, then, letting $C \to \infty$, we have*

$$|\sigma^2 C|^{1/2} p(\mathrm{v}) \to \left\{\prod\nolimits_J \Big/ (2\pi)^{d/2}\right\} p(J_2\mathrm{v}),$$

where $\prod_J$ is the product of the nonzero eigenvalues of the matrix $R^{\top}J_1^{\top}J_1R$ and d is the number of columns of R, rank(R)$\leq d$.

Proof: Let J be as specified in the theorem. Then, $p(\mathrm{v}) = P(J\mathrm{v})$ because $|J| = 1$. Permuting the rows of JR if necessary, we can always suppose that J_1R are the first rows of JR. This amounts to premultiplying JR by a matrix P obtained from the unit matrix by performing the same permutations. Given that P is orthogonal, we can take PJ instead of J. Let K be a selector $r \times d$ matrix, where $r = \mathrm{rank}(R)$, and consider model (10). If R is of full column rank, then $K = I_d$ and $r = d$. That the determinant $|K^\top R^\top J_1^\top J_1 RK^\top|$ is equal to the product of the non.cro eigenvalues of $R^\top J_1^\top J_1 R$ can be seen, for example, in [18, page 527]. Let $J_1RK^\top = M$. If $\Sigma_J = J\Sigma J^\top$ and we partition $\Sigma_J = (\Omega_{ij}), \Sigma_J^{-1} = (\Omega^{ij})$, $i, j = 1, 2$, conforming to $J = (J_1^\top, J_2^\top)^\top$, then, by Theorem 4, the log-likelihood of v verifies

$$\lambda(\mathrm{v}) + \frac{1}{2}\ln|C| \to -\frac{1}{2}\{\ln|\sigma^2\Omega_{22}(\Omega^{11})^{-1}| + \ln|M^\top\Omega^{11}M| + (J_2\mathrm{v})^\top\Omega_{22}^{-1}(J_2\mathrm{v})/\sigma^2\}.$$

Ansley and Kohn [2], make the following assumption.

Assumption A. Matrix R in (9) and (10) does not depend on the model parameters.

This assumption holds in many practical situations, including the examples of Section 2.

Corollary 1. *If Assumption A holds, then the AK likelihood does not depend on the matrix J.*

Proof: It is an immediate consequence of Theorem 5.

Even if Assumption A holds, Theorem 5 shows that the diffuse log-likelihood and the AK log-likelihood, when maximized with respect to σ^2, do not give the same results. The difference lies in the term $M\ln(\hat{\sigma}^2)$ in the σ^2-maximized diffuse log-likelihood versus $(M-d)\ln(\bar{\sigma}^2)$ in the AK log-likelihood, where $\hat{\sigma}^2 = (1/M)(q - \mathrm{s}^\top S^- \mathrm{s})$ and $\bar{\sigma}^2 = (1/(M-d))(q - \mathrm{s}^\top S^- \mathrm{s})$. This is a consequence of de Jong taking the limit of $\lambda(\mathrm{v}) + (1/2)\ln|C|$ instead of $\lambda(\mathrm{v}) + (1/2)\ln|\sigma^2 C|$ in Theorem 4. We think that it would have been more appropriate to do the latter than the former. For instance, when dealing with an ARIMA model, the AK likelihood coincides with the usual Box-Jenkins likelihood (see [4]), whereas the diffuse likelihood does not.

We now consider predicting the state x_k and v_k using $(\mathrm{v}_1^\top, \mathrm{v}_2^\top, \cdots, \mathrm{v}_{k-1}^\top)^\top$. The next two theorems give the details.

Theorem 6. *Let $\underline{\delta} \sim N(\mathrm{c}, \sigma^2 C)$ with C nonsingular and let $\tilde{\underline{\delta}}_k, \tilde{\mathrm{x}}_k$ and $\tilde{\mathrm{v}}_k$ be the predictors of $\underline{\delta}, \mathrm{x}_k$ and v_k using $(\mathrm{v}_1^\top, \mathrm{v}_2^\top, \cdots, \mathrm{v}_{k-1}^\top)^\top$, respectively. Suppose the EKF or the DKF is applied and let $\hat{\mathrm{X}}(\underline{\delta})_{k,k-1}$ and $E(\underline{\delta})_k$ be the submatrices formed by all but the first columns of $\hat{\mathrm{X}}_{k,k-1}$ and E_k, respectively. Then*

$$\tilde{\underline{\delta}}_k = (C^{-1} + R_k^\top\Sigma_k^{-1}R_k)^{-1}(C^{-1}\mathrm{c} + R_k^\top\Sigma_K^{-1}\mathrm{v}_k),$$

$$\mathrm{Mse}(\tilde{\underline{\delta}}_k) = \sigma^2(C^{-1} + R_k^\top\Sigma_k^{-1}R_k)^{-1}, \quad \tilde{\mathrm{x}}_{k,k-1} = \hat{\mathrm{X}}_{k,k-1}(1, -\tilde{\underline{\delta}}_k^\top)^\top,$$

$$\mathrm{Mse}(\tilde{\mathrm{x}}_{k,k-1}) = \sigma^2 P_{k,k-1} + \hat{\mathrm{X}}(\underline{\delta})_{k,k-1}\mathrm{Mse}(\tilde{\underline{\delta}}_k)\hat{\mathrm{X}}(\underline{\delta})_{k,k-1}^\top,$$

$$\mathrm{v}_k - \tilde{\mathrm{v}}_k = E_k(1, -\tilde{\underline{\delta}}_k^\top)^\top, \quad \mathrm{Mse}(\tilde{\mathrm{v}}_k) = \sigma^2 D_k + E(\underline{\delta})_k\mathrm{Mse}(\tilde{\underline{\delta}}_k)E(\underline{\delta})_k^\top,$$

where R_k is the submatrix formed by the first k rows of R and $\Sigma_k = \sigma^{-2} Var((\underline{\varepsilon}_1^\top, \cdots, \underline{\varepsilon}_k^\top)^\top)$.

Proof: The first two equalities are a consequence of Theorem 2. The other expressions can be proved as the corresponding ones for the case $C=0$ (Section 3).

Theorem 7. *With the notation and assumption of Theorems 6 and 4, if the rows of $\hat{X}(\underline{\delta})_{k,k-1}$ and $E(\underline{\delta})_k$ are in the space generated by the rows of $(R_k^\top \Sigma_k^{-1} R_k)^-$, then, letting $C \to \infty$,*

$$\tilde{\underline{\delta}}_k \to \hat{\underline{\delta}}_k = (R_k^\top \Sigma_k^{-1} R_k)^- R_k^\top \Sigma_k^{-1} \mathrm{v}_k, \quad \mathrm{Mse}(\tilde{\underline{\delta}}_k) \to \mathrm{Mse}(\hat{\underline{\delta}}_k) = \sigma^2 (R_k^\top \Sigma_k^{-1} R_k)^- ,$$

$$\tilde{\mathrm{x}}_{k,k-1} \to \hat{\mathrm{X}}_{k,k-1}(1, -\hat{\underline{\delta}}_k^\top)^\top ,$$

$$\mathrm{Mse}(\tilde{\mathrm{x}}_{k,k-1}) \to \sigma^2 P_{k,k-1} + \hat{\mathrm{X}}(\underline{\delta})_{k,k-1} \mathrm{Mse}(\hat{\underline{\delta}}_k) \hat{\mathrm{X}}(\underline{\delta})_{k,k-1}^\top ,$$

$$\tilde{\mathrm{v}}_k \to C_k \hat{X}_{k,k-1}(1, -\hat{\underline{\delta}}_k^\top)^\top , \quad \mathrm{Mse}(\tilde{\mathrm{v}}_k) \to \sigma^2 D_k + E(\underline{\delta})_k \mathrm{Mse}(\hat{\underline{\delta}}_k) E(\underline{\delta})_k^\top .$$

Proof: The first two limits are consequences of Theorems 2 and 4. The other expressions are a direct consequence of Theorem 6.

By Theorem 7, in order to get the desired predictors, we must consider the regression model (9) with $\underline{\delta}$ fixed and apply the GLS theory. We can use the results of Section 3 and, in order to get an efficient algorithm, we can apply the EKF or the DKF for likelihood evaluation or prediction. Note that the difficulties that may arise stem from the fact that the matrix R may be rank deficient. In this last case, we have to use generalized inverses throughout the process and neither all observations will be predictable, nor will all states be estimatable.

The next theorem states that the predictors obtained with the modified Kalman Filter coincide with those obtained by means of the EKF or the DKF.

Theorem 8. *Let Assumption A hold. Then, the predictors of x_k and v_k obtained with the modified Kalman Filter and those obtained with the EKF or the DKF coincide. If the same estimator of σ^2 is used in both procedures, then the Mse errors also coincide.*

Proof: Theorem 5.2 in [2] states that the AK predictors coincide with the diffuse predictors and the statement about the Mse follows trivially.

We have seen that, in order to evaluate the AK log-likelihood, we can use the modified Kalman Filter of AK, although it is not the best procedure, or we may use the efficient EKF or DKF to evaluate the diffuse log-likelihood, which, by Theorem 5, differs from the AK log-likelihood only in a constant. This constant, under Assumption A, does not depend on model parameters. The EKF or the DKF should be applied to model (9) considering $\underline{\delta}$ fixed ($C = 0$). It would be nice to employ the EKF or the DKF only for an initial stretch of the data, as short as possible, to construct an estimator of $\underline{\delta}$ and, from then on, use the KF. When this occurs, one speaks of a collapse of the EKF or the DKF to the KF. Let rank$(R) = r$ and suppose that the first r rows of R are linearly independent. Let R_I be the submatrix formed by the first r rows and let R_{II} consist of the other rows of R. Partition $\mathrm{v} = (\mathrm{v}_I^\top, \mathrm{v}_{II}^\top)^\top$ and $\underline{\varepsilon} = (\underline{\varepsilon}_I^\top, \underline{\varepsilon}_{II}^\top)^\top$ conforming to $R = (R_I^\top, R_{II}^\top)^\top$. Then, we can write

$$\mathrm{v}_I = R_I \underline{\delta} + \underline{\varepsilon}_I \tag{11a}$$

$$\mathrm{v}_{II} = R_{II} \underline{\delta} + \underline{\varepsilon}_{II} \tag{11b}$$

The next theorem shows how to implement the collapsing of the EKF or DKF to the KF.

Theorem 9. *Under Assumption A, let J with $|J| = 1$ be a matrix like those used by AK to define their likelihood, with corresponding submatrices J_1 and J_2 such that $J_1 R \neq 0$ and $J_2 R = 0$, and let $p(\mathrm{v}_{II}|\mathrm{v}_I, \hat{\underline{\delta}}_I)$ be the density of $\mathrm{v}_{II} - E\{\mathrm{v}_{II}|\mathrm{v}_I, \hat{\underline{\delta}}_I\}$, where $E\{\mathrm{v}_{II}|\mathrm{v}_I, \hat{\underline{\delta}}_I\}$ is the conditional expectation of v_{II} given v_I in model* (11*a*) *and* (11*b*), *considering $\underline{\delta}$ fixed ($C = 0$), and $\underline{\delta}$ replaced by its maximum likelihood estimator $\hat{\underline{\delta}}_I$ in model* (11*a*). *Then,*

$$p(J_2 \mathrm{v}) = p(\mathrm{v}_{II}|\mathrm{v}_I, \hat{\underline{\delta}}_I),$$

where $p(J_2\mathrm{v})$ is the density of $J_2\mathrm{v}$.

Proof: For simplicity, consider that R_I is of full column rank. If not, we would use generalized inverses, but the proof would not be affected. From model (11a) and (11b), we have, considering $\underline{\delta}$ fixed,

$$E\{\mathrm{v}_{II}|\mathrm{v}_I\} = R_{II}\underline{\delta} + \Sigma_{21}\Sigma_{11}^{-1}(\mathrm{v}_I - R_I\underline{\delta}),$$

where $\Sigma_{21} = Cov(\underline{\varepsilon}_{II}, \underline{\varepsilon}_I)$ and $\Sigma_{11} = Var(\underline{\varepsilon}_I)$. Then, $E\{\mathrm{v}_{II}|\mathrm{v}_I, \hat{\underline{\delta}}_I\} = R_{II}R_I^{-1}\mathrm{v}_I$ and $\mathrm{v}_{II} - E\{\mathrm{v}_{II}|\mathrm{v}_I, \hat{\underline{\delta}}_I\} = \mathrm{v}_{II} - R_{II}R_I^{-1}\mathrm{v}_I$. Define the matrix $J = (J_1^\top, J_2^\top)^\top$ with $J_1 = (I, 0)$ and $J_2 = (-R_{II}R_I^{-1}, I)$. Then, J is a matrix of the type used by AK to define their likelihood and $\mathrm{v}_{II} - E\{\mathrm{v}_{II}|\mathrm{v}_I, \hat{\underline{\delta}}_I\} = J_2\mathrm{v}$. This completes the proof of the theorem.

By Theorem 9, to evaluate the log-likelihood of $\mathrm{v}_{II} - E\{\mathrm{v}_{II}|\mathrm{v}_I, \hat{\underline{\delta}}_I\} = J_2\mathrm{v}$, we can proceed as follows. First, use the EKF or the DKF in model (11a) to obtain the maximum likelihood estimator (mle) $\hat{\underline{\delta}}_I$ and initial conditions for the KF

$$E\{\mathrm{x}_s\} = \hat{\mathrm{X}}_{s,s-1}(1, -\hat{\underline{\delta}}_I^\top)^\top, \quad Var(\mathrm{x}_s) = \sigma^2 P_{s,s-1} + \hat{\mathrm{X}}(\underline{\delta})_{s,s-1}\mathrm{Mse}(\hat{\underline{\delta}}_I)\hat{\mathrm{X}}(\underline{\delta})_{s,s-1}^\top,$$

where v_s is the first observation in (11b). Then, proceed with the KF, applied to the second stretch of the data v_{II}, to obtain the log-likelihood of $\mathrm{v}_{II} - E\{\mathrm{v}_{II}|\mathrm{v}_I, \hat{\underline{\delta}}_I\} = J_2\mathrm{v}$ as in Section 3, but with no regression parameters. We have used the initial stretch v_I of the data to construct the mle $\hat{\underline{\delta}}_I$ and the initial conditions for the KF. After that, the effect of $\underline{\delta}$ has been absorbed into the estimator of the state vector and that is the reason why we can collapse the EKF or DKF to the KF.

We now give another interpretation to the result of Theorem 9. With the notation of Theorem 2, we can write

$$\lambda(\mathrm{v}) + \ln|C|/2 = \{\lambda(\mathrm{v}_I) + \ln|C|/2\} + \lambda(\mathrm{v}_{II}|\mathrm{v}_I)\,, \tag{12}$$

where $\lambda(\mathrm{v}_{II}|\mathrm{v}_I)$ is the conditional log-likelihood of v_{II} given v_I. By Theorem 3, letting $C \to \infty$, the term in curly brackets tends to $-\{M_I \ln(\sigma^2) + |R_I^{\top} R_I|\}/2$, where M_I is the number of components in v_I, whereas $\lambda(\mathrm{v}_{II}|\mathrm{v}_I)$ converges to the log-likelihood of $J_2\mathrm{v}$, which, by Theorem 9, is equal to the log-likelihood of $\mathrm{v}_{II} - E\{\mathrm{v}_{II}|\mathrm{v}_I, \hat{\underline{\delta}}_I\}$. Thus, the diffuse log-likelihood of v is the sum of the diffuse log-likelihood of v_I and the log-likelihood of $J_2\mathrm{v}$. Note that the first term does not contribute to either the determinant or the sum of squares of the diffuse log-likelihood.

Bell and Hillmer [3] use a similar idea to construct initial conditions for the KF. Instead of employing the KF for the initial stretch of the data, they use the transformation approach of AK to construct the mle $\hat{\underline{\delta}}_I$ and the initial conditions for the KF directly. Whether this approach is more advantageous than using the KF is something that depends on the pecularities of the problem at hand. If it is easier to obtain the mle and the initial conditions directly, then it can be used. However, the KF approach to construct the mle and the initial conditions has the advantage that it is easy to implement, does not depend on *ad hoc* procedures and it imposes very little computational and/or programming burden.

The case we have been considering, where the submatrix R_I is formed by the first r rows of R is important because it happens often in practice. Examples of this are ARIMA models and ARIMA component models.

If in model (11a) and (11b) we have $\mathrm{v}_I = \underline{\delta}$, then Theorem 9 implies $p(J_2\mathrm{v}) = p(\mathrm{v}_{II}|\mathrm{v}_I)$. Also, if J is a matrix like those used by AK to define their likelihood and J is of the form $J = (J_1^{\top}, J_2^{\top})^{\top}$ with $J_1 = (I, 0)$, then v_I is independent of $J_2\mathrm{v}$. This is the conditional likelihood approach used in [6] in the context of regression models with ARIMA disturbances and generalized to the case when there are missing observations. For ARIMA (p, d, q) models, the situation simplifies still further because it is not necessary to employ the EKF or the DKF for the initial stretch of the data $\mathrm{v}_I = \underline{\delta}$ to obtain initial conditions for the KF. The SSM can be redefined by simply translating forward the initial conditions d units in time, where d is the degree of the differencing operator.

Suppose there are missing observations in v and that in (11a) the vector v_I contains a subvector v_{IM} of missing observations. Let v_{IO}be the subvector of v_I formed by the nonmissing observations and let v_{II} be the subvector of v containing the rest of the nonmissing observations. Then, by analogy with the result of Theorem 9, we can still consider $\mathrm{v}_{II} - E\{\mathrm{v}_{II}|\mathrm{v}_I, \hat{\underline{\delta}}_I\}$, treat v_{IM} as a vector of fixed parameters, and define the likelihood of $(\mathrm{v}_{IO}^{\top}, \mathrm{v}_{II}^{\top})^{\top}$ as that of the regression model

$$\underline{\eta}_{II} = R_{II}U_M \mathrm{v}_{IM} + \underline{\omega}_{II}, \tag{13}$$

where $\underline{\eta}_{II} = \mathrm{v}_{II} - R_{II}U_O\mathrm{v}_{IO}, \underline{\omega}_{II} = \underline{\varepsilon}_{II} - R_{II}R_I^{-1}\underline{\varepsilon}_I$, and U_O and U_M are the submatrices of R_I^{-1} formed by the columns corresponding to v_{IO} and v_{IM}, respectively. Here we have supposed that R_I is of full column rank. If not, we would use generalized inverses, but the main result would not be affected. Note that the vector v_{IM} is considered as a vector of fixed parameters that have to be estimated along with the other parameters of the model. The next theorem shows that this definition of the likelihood is equivalent to the AK definition.

Theorem 10. *Under Assumption A, the $(\sigma^2, \mathrm{v}_{IM})$-maximized log-likelihood corresponding to (13) coincides, up to a constant, with the σ^2-maximized AK log-likelihood.*

Proof: Let $Var(\underline{\omega}_{II}) = \sigma^2\Omega$ and let $\Omega = LL^{\top}$ be the Cholesky decomposition of Ω. If we premultiply (13) by L^{-1}, we obtain the OLS model

$$L^{-1}\underline{\eta}_{II} = L^{-1}R_{II}U_M\mathrm{v}_{IM} + L^{-1}\underline{\omega}_{II}.$$

The QR algorithm, applied to the $L^{-1}R_{II}U_M$ matrix, yields an upper triangular matrix S with nonzero elements in the main diagonal such that $Q^{\top}L^{-1}R_{II}U_M = (S^{\top}, 0^{\top})^{\top}$, where Q is an orthogonal matrix. Then, we can write

$$Q^{\top}L^{-1}\underline{\eta}_{II} = \begin{bmatrix} S \\ 0 \end{bmatrix} \mathrm{v}_{IM} + Q^{\top}L^{-1}\underline{\omega}_{II}.$$

The matrix L^{-1} will not have, in general, unit determinant. If we multiply L^{-1} by $\alpha = |L|^{1/M_{II}}$, where M_{II} is the number of components in v_{II}, then αL^{-1} has unit determinant. Let $K = (K_1^{\top}, K_2^{\top})^{\top}$ with $K_1 = (I, 0)$ and $K_2 = (-R_{II}U_O, I)$ and let $P = (P_1^{\top}, P_2^{\top})^{\top}$ with $P_1 = (I, 0)$ and $P_2 = (0, \alpha Q^{\top}L^{-1})$. Partition $Q^{\top} = (Q_1, Q_2)^{\top}$ conforming to $Q_1^{\top}L^{-1}R_{II}U_M = S$ and $Q_2^{\top}L^{-1}R_{II}U_M = 0$. If $J = PK$, then J has unit determinant and

$$\begin{aligned} &J(\mathrm{v}_{IO}^{\top}, \mathrm{v}_{II}^{\top})^{\top} = (\mathrm{v}_{IO}^{\top}, (\alpha Q^{\top}L^{-1}\underline{\eta}_{II})^{\top})^{\top} \\ &= (R_{IO}^{\top}, (\alpha SR_{IM})^{\top}, 0)^{\top}\underline{\delta} + (\underline{\varepsilon}_{IO}^{\top}, (\underline{\varepsilon}_{IM} + \alpha Q_1^{\top}L^{-1}\underline{\omega}_{II})^{\top}, (\alpha Q_2^{\top}L^{-1}\underline{\omega}_{II})^{\top})^{\top}, \end{aligned}$$

where we have partitioned R_I and $\underline{\varepsilon}_I$ conforming to the partition of v_I into v_{IO} and v_{IM}. Given that $(R_{IO}^{\top}, (\alpha SR_{IM})^{\top})^{\top}$ has rank equal to that of R_I, the matrix J is of the AK type. Therefore, the AK likelihood is the density of $\alpha Q_2^{\top}L^{-1}\underline{\eta}_{II}$ and the AK log-likelihood, maximized with respect to σ^2, is

$$-\frac{1}{2}\left\{(M_{II} - rs)\ln(\hat{\sigma}^2) + \ln|L|^{2(M_{II}-rs)/M_{II}}\right\},$$

where

$$\hat{\sigma}^2 = \underline{\eta}_{II}^\top (L^{-1})^\top Q_2 Q_2^\top L^{-1} \underline{\eta}_{II} / (M_{II} - r_S)$$

and $r_S = rank(S)$. The log-likelihood of (13), maximized with respect to σ^2 and v_{IM} is $-\{M_{II} \ln(\tilde{\sigma}^2) + \ln|L|^2\}/2$, where

$$\tilde{\sigma}^2 = \underline{\eta}_{II}^\top (L^{-1})^\top Q_2 Q_2^\top L^{-1} \underline{\eta}_{II} / M_{II} .$$

This completes the proof of the theorem.

Theorem 10 generalizes the result obtained in [6] for ARIMA models with missing data. This approach is useful when the matrix R_I corresponding to the first observations v_I (including the missing ones) is of full column rank.

We now suppose that in model (9) the first r rows of R do not, in general, constitute a submatrix of R of rank r. Let R_I be the first submatrix of R formed adjoining consecutive rows to the first row, such that it has full column rank and let R_{II} consist of the other rows of R. Partition $\mathrm{v} = (\mathrm{v}_I^\top, \mathrm{v}_{II}^\top)^\top$ and $\underline{\varepsilon} = (\underline{\varepsilon}_I^\top, \underline{\varepsilon}_{II}^\top)^\top$ conforming to $R = (R_I^\top, R_{II}^\top)^\top$. In the rest of the section, whenever we refer to models (11a) and (11b), we will refer to this partition. Consider the decomposition given by (12). Then, letting $C \to \infty$ as before, the term in curly brackets tends to

$$-\frac{1}{2}\{M_I \ln(\sigma^2) + |\Sigma_{11}| + |R_I^\top \Sigma_{11}^{-1} R_I| + (\mathrm{v}_I - R_I \hat{\underline{\delta}}_I)^\top \Sigma_{11}^{-1} (\mathrm{v}_I - R_I \hat{\underline{\delta}}_I)/\sigma^2\} ,$$

where $Var(\underline{\varepsilon}_I) = \sigma^2 \Sigma_{11}$ and $\hat{\underline{\delta}}_I = (R^\top \Sigma_{11}^{-1} R_I)^{-1} R_I^\top \Sigma_{11}^{-1} \mathrm{v}_I$. The conditional log-likelihood $\lambda(\mathrm{v}_{II}|\mathrm{v}_I)$ converges to the log-likelihood of $J_2 \mathrm{v}$, where

$$J_2 = (-R_{II} S_I^{-1} T_I - \Sigma_{21} \Sigma_{11}^{-1} (I - R_I S_I^{-1} T_I), I) , \quad \Sigma_{21} = Cov(\underline{\varepsilon}_{II}, \underline{\varepsilon}_I) ,$$

and $T_I = R_I^\top \Sigma_{11}^{-1}$. To see this, define $J = (J_1^\top, J_2^\top)^\top$ with $J_1 = (I, 0)$. Then, $\lambda(\mathrm{v}) = \lambda(J\mathrm{v})$ because J has unit determinant and

$$\lambda(\mathrm{v}) + \frac{1}{2}\ln|C| = \{\lambda(\mathrm{v}_I) + \frac{1}{2}\ln|C|\} + \lambda(J_2 \mathrm{v}|\mathrm{v}_I) .$$

Note that now J is not a matrix of the AK type.

Theorem 11. *Let J be the matrix we have just defined , with the corresponding submatrices J_1 and J_2, and let $p(\mathrm{v}_{II}|\mathrm{v}_I, \hat{\underline{\delta}}_I)$ be the density of $\mathrm{v}_{II} - E\{\mathrm{v}_{II}|\mathrm{v}_I, \hat{\underline{\delta}}_I\}$, where $E\{\mathrm{v}_{II}|\mathrm{v}_I, \hat{\underline{\delta}}_I\}$ is the conditional expectation of v_{II} given v_I in model (11a) and (11b), considering $\underline{\delta}$ fixed ($C = 0$) and $\underline{\delta}$ replaced by its maximum likelihood estimator $\hat{\underline{\delta}}_I$ in model (11a). Then,*

$$p(J_2 \mathrm{v}) = p(\mathrm{v}_{II}|\mathrm{v}_I, \hat{\underline{\delta}}_I) ,$$

where $p(J_2\mathrm{v})$ is the density of $J_2\mathrm{v}$.

Proof: The proof is analogous to that of Theorem 9.

Thus, to evaluate the AK log-likelihood or the diffuse log-likelihood, we can still use the EKF or the DKF as before, until we have processed a stretch of observations such that the corresponding submatrix of R has full column rank, and then collapse to the KF. The likelihood is evaluated as the sum of two terms. One corresponding to the stretch $\mathbf{v}_I$ and the other corresponding to $\mathbf{v}_{II}$. More specifically, the EKF or the DKF applied to model ($11a$) yields

$$|\Sigma_{11}|, |R_I^\top \Sigma_{11}^{-1} R_I| \quad \text{and} \quad (\mathrm{v}_I - R_I\hat{\underline{\delta}}_I)^\top \Sigma_{11}^{-1}(\mathrm{v}_I - R_I\hat{\underline{\delta}}_I), \tag{14}$$

where $\hat{\underline{\delta}}_I = (R_I^\top \Sigma_{11}^{-1} R_I)^{-1} R_I^\top \Sigma_{11}^{-1} \mathrm{v}_I$. These three terms will be needed for the computation of the likelihood because now there will be no cancellation of terms. With the notation of Section 3, if the EKF is used, the expressions in (14) are

$$\ln|\Sigma_{11}| = 2\sum_{k=1}^{N_I} \ln|D_k^{1/2}|, \quad |R_I^\top \Sigma_{11}^{-1} R_I| = |U_I|^2$$

and

$$(\mathrm{v}_I - R_I\hat{\underline{\delta}}_I)^\top \Sigma_{11}^{-1}(\mathrm{v}_I - R_I\hat{\underline{\delta}}_I) = \underline{\omega}_{I,2}^\top \underline{\omega}_{I,2},$$

whereas, if the DKF is used, they are

$$\ln|\Sigma_{11}| = \sum_{k=1}^{N_I} \ln|D_k|, \quad |R_I^\top \Sigma_{11}^{-1} R_I| = |S_I|$$

and

$$(\mathrm{v}_I - R_I\hat{\underline{\delta}}_I)^\top \Sigma_{11}^{-1}(\mathrm{v}_I - R_I\hat{\underline{\delta}}_I) = q_I - \mathrm{s}_I^\top S_I^{-1} \mathrm{s}_I.$$

The initialization for the KF, to be used with the second stretch of the data v_{II}, is

$$E\{\mathrm{x}_s\} = \hat{\mathrm{X}}_{s,s-1}(1, -\hat{\underline{\delta}}_I^\top)^\top, \quad Var(\mathrm{x}_s) = \sigma^2 P_{s,s-1} + \hat{\mathrm{X}}(\underline{\delta})_{s,s-1}\mathrm{Mse}(\hat{\underline{\delta}}_I)\hat{\mathrm{X}}(\underline{\delta})_{s,s-1}^\top,$$

where, as before, $\mathbf{v}_s$ is the first observation in ($11b$). Once the run of the KF is completed, we have to add up the terms in (14) to the corresponding terms obtained with the KF, $|Var(J_2\mathrm{v})|$ and $(J_2\mathrm{v})^\top(Var(J_2\mathrm{v}))^{-1}(J_2\mathrm{v})$.

The fact that we don't know for how long we will have to use the EKF or the DKF before we make the transition to the KF may make collapsing unattractive. There is an alternative procedure to evaluate the AK log-likelihood or the diffuse log-likelihood that might be of interest in some cases. It consists essentially of reshuffling the observations in such a way that again the first r rows of R are linearly independent. An algorithm to achieve this is the following. Apply the EKF or the DKF to model (9) and, at the same time, obtain the row echelon form o the R matrix. Each time a new observation v_k is being incorporated, we check whether its corresponding row vector R_k is a linear combination of the rows already processed. If it is, we skip this observation as if it were missing (see [10]). Otherwise, we process the observation as part of the initial stretch of the data v_I. Proceeding in this way, after some time we will have processed a stretch of the data v_I for which the corresponding submatrix R_I of R will be formed by a maximal set of linearly independent row vectors. Let v_{II} consist of the other observations and let v_s be the first observation that we skip as if it were missing. This will be the first observation of v_{II}. Suppose the Fixed Point Smoother (FPS) corresponding to v_s is applied, along with the EKF or the DKF, to all the columns of $\hat{\mathrm{X}}_{k+1,k}$. Then, after processing v_I, we can set up as initial conditions for the KF, to be applied to v_{II}, the following

$$E\{\mathrm{x}_s\} = \hat{\mathrm{X}}_{s,I}(1, -\hat{\underline{\delta}}_I^{\top})^{\top}, \quad Var(\mathrm{x}_s) = \sigma^2 P_{s,I} + \hat{\mathrm{X}}(\underline{\delta})_{s,I}\mathrm{Mse}(\hat{\underline{\delta}}_I)\hat{\mathrm{X}}(\underline{\delta})_{s,I}^{\top},$$

where $\hat{\mathrm{X}}_{s,I}, P_{s,I}$ and $\hat{\mathrm{X}}(\underline{\delta})_{s,I}$ are the matrices obtained with the FPS, and $\hat{\underline{\delta}}_I$ is the mle corresponding to v_I. Note that the advantage of using only the KF for likelihood evaluation comes at the expense of an increase in the computations.

Example 3. Consider the following ARIMA (1,1,0) model

$$(1 + \phi L)\nabla \mathrm{v}_k = \mathrm{a}_k,$$

where the notation is as in Example 1. To obtain an SSM formulation, we define $X_k = 0, C_k = C = (1, 0), Z_k = 0, W_k = 0,$ $A_k = \begin{bmatrix} 0 & 1 \\ \phi & 1-\phi \end{bmatrix} = A$, $H_k = H$, with $H_1 = 1, H_2 = 1 - \phi, \mathrm{x}_{1,k} = \mathrm{v}_k, \mathrm{x}_{2,k} = \mathrm{v}_{k+1} - \mathrm{a}_{k+1}$ and $\underline{\xi}_{k-1} = \mathrm{a}_k$. Then, we can write

$$\mathrm{x}_k = A\mathrm{x}_{k-1} + H\mathrm{a}_k; \quad \mathrm{v}_k = C\mathrm{x}_k.$$

To initialize, we consider that $(1 - L)\mathrm{v}_k = \mathrm{u}_k$ is stationary and follows the model $(1 + \phi L)\mathrm{u}_k = \mathrm{a}_k$. Then,

$$\mathrm{x}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \mathrm{v}_0 + \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -\phi \end{bmatrix} \mathrm{u}_1,$$

and we can choose $A_0 = I, , B = (1, 1)^{\top}, \mathrm{x}_0 = B\underline{\delta}, \underline{\delta} = \mathrm{v}_0$ and

$$H_0 = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -\phi \end{bmatrix} \frac{1}{\sqrt{1 - \phi^2}}.$$

The first state is $\mathrm{x}_1 = B\underline{\delta} + H_0\mathrm{a}_1$. Model (9) specializes to $R = (1, 1, \cdots, 1)^{\top}$ and $\underline{\varepsilon}_k = \mathrm{u}_1 + \cdots + \mathrm{u}_k, k = 1, \cdots, N$. The AK likelihood can be obtained as the density of the differenced data. This is equivalent to multiply v by the matrix

$$J=\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{bmatrix},$$

define $J_1=(1,0,\cdots,0)$ and J_2 such that $J=(J_1^\top, J_2^\top)^\top$, and take as AK density the density of $J_2\mathrm{v}$. Note that $JR=(1,0,\cdots,0)^\top$ and $J\underline{\varepsilon}=(\mathrm{u}_1,\mathrm{u}_2,\cdots,\mathrm{u}_N)^\top$. The EKF or the DKF produces $D_1=1/(1-\phi^2), D_k=1, k=2,\cdots,N$,

$$\begin{aligned} &\ln|\Sigma| = \ln|J\Sigma J^\top| = \ln|D_1| + \cdots + \ln|D_N| = \ln(1/(1-\phi^2)), \\ &\ln|R^\top\Sigma^{-1}R| = \ln|R^\top J^\top(J\Sigma J^\top)^{-1}JR| = 0, \quad \hat{\underline{\delta}} = \mathrm{v}_1, \\ &(\mathrm{v}-R\hat{\underline{\delta}})^\top\Sigma^{-1}(\mathrm{v}-R\hat{\underline{\delta}}) = (\mathrm{v}_2+\phi\mathrm{v}_1-\mathrm{v}_1)^2 \\ &\quad + \sum_{k=3}^{N}[(\mathrm{v}_k+\phi\mathrm{v}_{k-1})-(\mathrm{v}_{k-1}+\phi\mathrm{v}_{k-2})]^2 . \end{aligned}$$

In case the DKF is applied,

$$Q_{N+1}=\begin{bmatrix} \begin{array}{c}(1-\phi^2)\mathrm{v}_1^2+(\mathrm{v}_2+\phi\mathrm{v}_1-\mathrm{v}_1)^2 \\ +\sum_{k=3}^{N}[(\mathrm{v}_k+\phi\mathrm{v}_{k-1})-(\mathrm{v}_{k-1}+\phi\mathrm{v}_{k-2})]^2\end{array} & (1-\phi^2)\mathrm{v}_1^\top \\ (1-\phi^2)\mathrm{v}_1 & (1-\phi^2) \end{bmatrix}.$$

Model (11*a*) becomes the first equation of (9), $\mathrm{v}_1=\underline{\delta}+\mathrm{u}_1$, where $\mathrm{v}_I=\mathrm{v}_1, R_I=1$ and $\underline{\varepsilon}_I=\mathrm{u}_1$. Model (11*b*) consists of the rest of the equations. Suppose we use the EKF or the DKF in (11*a*) to obtain $\hat{\underline{\delta}}_I$ and initial conditions for the KF, that we will apply later to model (11*b*). Then,

$$\begin{aligned} &\hat{\mathrm{X}}_{1,0}=\begin{bmatrix} 0 & -1 \\ 0 & -1 \end{bmatrix}, \quad E_1=(\mathrm{v}_1,1), \quad P_{1,0}=\frac{1}{1-\phi^2}\begin{bmatrix} 1 & 1-\phi \\ 1-\phi & (1-\phi)^2 \end{bmatrix}, \\ &G_1=\begin{bmatrix} 1-\phi \\ \phi+(1-\phi)^2 \end{bmatrix}, \quad D_1=\frac{1}{1-\phi^2}, \\ &\hat{\mathrm{X}}_{2,1}=\begin{bmatrix} (1-\phi)\mathrm{v}_1 & -\phi \\ (1-\phi+\phi^2)\mathrm{v}_1 & \phi(\phi-1) \end{bmatrix}, \quad P_{2,1}=\begin{bmatrix} 1 & 1-\phi \\ 1-\phi & (1-\phi)^2 \end{bmatrix}. \end{aligned}$$

Given that $\hat{\underline{\delta}}_I=\mathrm{v}_1$, we have $\mathrm{Mse}(\hat{\underline{\delta}}_I)=1/(1-\phi^2)$ and the initial conditions for the KF are

$$\begin{aligned} &E\{\mathrm{x}_2\}=\hat{\mathrm{X}}_{2,1}\begin{bmatrix} 1 \\ -\hat{\underline{\delta}}_I \end{bmatrix}=\begin{bmatrix} \mathrm{v}_1 \\ \mathrm{v}_1 \end{bmatrix}, \\ &Var(\mathrm{x}_2)=P_{2,1}+\hat{\mathrm{X}}(\underline{\delta})_{2,1}\mathrm{Mse}(\hat{\underline{\delta}}_I)\hat{\mathrm{X}}^\top(\underline{\delta})_{2,1}=\frac{1}{1-\phi^2}\begin{bmatrix} 1 & 1-\phi \\ 1-\phi & (1-\phi)^2 \end{bmatrix}. \end{aligned}$$

Therefore, using the EKF or the DKF in ($11a$) to estimate $\underline{\delta}$ and to compute initial conditions for the KF yields the same starting values, but shifted ahead one period of time. This is an example where we can redefine the SSM, taking $\mathrm{v}_I = \underline{\delta}$ and translate the initial conditions forward one unit of time. This is true for all ARIMA (p, d, q) models (see [6]).

§5 Initial state with an unspecified distribution: the general case

In this Section we consider a more general SSM than that of Section 3. Besides making the assumption that $\underline{\delta}$ in $\mathrm{x}_0 = B\underline{\delta}$ has an unspecified distribution, $\underline{\delta} \sim N(\mathrm{c}, \sigma^2 C)$, with C nonsingular, we allow for regression parameters. That is, we consider $\underline{\beta}$ fixed but unknown. By Theorem 1, we have $\mathrm{v} = R\underline{\delta} + S\underline{\beta} + \underline{\varepsilon}$. Defining $X = (R, S)$ and $\underline{\gamma} = (\underline{\delta}^\top, \underline{\beta}^\top)^\top$ we can write the model more concisely as

$$\mathrm{v} = X\underline{\gamma} + \underline{\varepsilon}. \tag{15}$$

To define the AK likelihood, consider a matrix J of the type used by AK when there are no regression parameters and let J_1 and J_2 be the corresponding submatrices such that $J_1 R \neq 0$ and $J_2 R = 0$. Then, the AK likelihood is the density of $J_2(\mathrm{v} - S\underline{\beta})$. In order to efficiently evaluate the likelihood and predict and interpolate unobserved v_k's, they use their modified KF and modified FPS, applying them also to the columns of the regression matrix, as outlined in Section 3. The reader is referred to [15] and [16] for details. For the reasons mentioned in Section 4, we consider the modified KF and modified FPS computationally less efficient and conceptually more complex than the EKF or the DKF.

To compute the diffuse log-likelihood of (15) we have to consider that $\underline{\delta}$ is diffuse, $C \to \infty$, and $\underline{\beta}$ is fixed. De Jong does not consider explicitly this case, although it is a case that is often encountered in practice. Proceeding as in Theorems 2 and 3 of Section 4, replacing v by $\mathrm{v} - S\underline{\beta}$, and letting $C \to \infty$, we have

$$\begin{aligned}\lambda(\mathrm{v}) + \frac{1}{2}\ln|C| \to -\frac{1}{2}\{&\ln|\sigma^2\Sigma| + \ln|R^\top\Sigma^{-1}R| \\ &+ (\mathrm{v} - S\underline{\beta} - R\hat{\underline{\delta}})^\top\Sigma^{-1}(\mathrm{v} - S\underline{\beta} - R\hat{\underline{\delta}})/\sigma^2\}\,, \\ \bar{\underline{\delta}} \to \hat{\underline{\delta}} = (R^\top\Sigma^{-1}R)^{-1}&R^\top\Sigma^{-1}(\mathrm{v} - S\underline{\beta})\,.\end{aligned}$$

Minimizing this diffuse log-likelihood with respect to $\underline{\beta}$ yields an estimator $\hat{\underline{\beta}}$ which minimizes $(\mathrm{v} - S\underline{\beta})^\top P^\top\Sigma^{-1}P(\mathrm{v} - S\underline{\beta})$, where $P = I - R(R^\top\Sigma^{-1}R)^{-1}R^\top\Sigma^{-1}$. It can be shown that the estimators $\hat{\underline{\delta}}$ and $\hat{\underline{\beta}}$ obtained in this way can be obtained in a single stage as the GLS estimator $\hat{\underline{\gamma}} = (\hat{\underline{\delta}}^\top, \hat{\underline{\beta}}^\top)^\top$ of model (15). Thus, the EKF or the DKF can be used to compute the $(\sigma^2, \underline{\gamma})$-maximized diffuse log-likelihood, given by

$$-\frac{1}{2}\{M\ln(\hat{\sigma}^2) + \ln|\Sigma| + \ln|R^\top\Sigma^{-1}R|\}\,,$$

where M is the number of components in v and $\hat{\sigma}^2 = (1/M)(\mathrm{v}-X\hat{\underline{\gamma}})^\top\Sigma^{-1}(\mathrm{v}-X\hat{\underline{\gamma}})$. Under Assumption A of Section 4, the AK $(\sigma^2,\underline{\gamma})$-maximized log-likelihood differs from the $(\sigma^2,\underline{\gamma})$-maximized diffuse log-likelihood only in a constant. As in Section 4, it is possible to employ the EKF or the DKF for an initial stretch of the data to construct an estimator of $\underline{\delta}$. However, it will not be possible now to collapse to the KF because we will still have to estimate the $\underline{\beta}$ parameters. The most we can do is to collapse to a reduced dimension EKF or DKF. More specifically, let $\mathrm{rank}(R) = r$ and suppose that the first r rows of R are linearly independent. Let R_I be the submatrix formed by the first r rows and let R_{II} consist of the other rows of R. Partition $\mathrm{v} = (\mathrm{v}_I^\top, \mathrm{v}_{II}^\top)^\top, S = (S_I^\top, S_{II}^\top)^\top$, and $\underline{\varepsilon} = (\underline{\varepsilon}_I^\top, \underline{\varepsilon}_{II}^\top)^\top$ conforming to $R = (R_I^\top, R_{II}^\top)^\top$. Then, we can write

$$\mathrm{v}_I = R_I\underline{\delta} + S_I\underline{\beta} + \underline{\varepsilon}_I\,, \tag{16a}$$

$$\mathrm{v}_{II} = R_{II}\underline{\delta} + S_{II}\underline{\beta} + \underline{\varepsilon}_{II}\,. \tag{16b}$$

Suppose that R_I has full column rank. If not, we would use generalized inverses instead of true inverses but the main result would not be affected. As in Section 4, we will apply first the EKF or the DKF to (16a) to obtain a GLS estimator $\hat{\underline{\delta}}_I$ of $\underline{\delta}$. However it will not be possible now to absorb both $\underline{\delta}$ and $\underline{\beta}$ into the state estimator. Only $\underline{\delta}$ will be absorbed. In this way, the EKF or the DKF will only be simplified, not collapsed to the KF, when we apply it to (16b) in the second step of the procedure. The number of states of the EKF will be reduced by a number equal to the number of components in $\underline{\delta}$. Let v_s be the first observation in (16b). We showed in Section 3 that, if $\underline{\delta}$ and $\underline{\beta}$ are known, then the estimator of the state x_s using $(\underline{\gamma}^\top, \mathrm{v}_1^\top, \mathrm{v}_2^\top, \cdots, \mathrm{v}_{s-1}^\top)^\top$ is

$$\hat{\mathrm{x}}_{s,s-1} = \hat{\mathrm{X}}_{s,s-1}(1-\underline{\gamma}^\top)^\top = \hat{\mathrm{X}}(\mathrm{v})_{s,s-1} - \hat{\mathrm{X}}(\underline{\delta})_{s,s-1}\underline{\delta} - \hat{\mathrm{X}}(\underline{\beta})_{s,s-1}\underline{\beta}, \tag{17}$$

where $\hat{\mathrm{X}}(\mathrm{v})_{s,s-1}, \hat{\mathrm{X}}(\underline{\delta})_{s,s-1}$ and $\hat{\mathrm{X}}(\underline{\beta})_{s,s-1}$ are the columns of $\hat{\mathrm{X}}_{s,s-1}$ corresponding to v_s, $\underline{\delta}$ and $\underline{\beta}$, respectively. The GLS estimator $\hat{\underline{\delta}}_I$ of $\underline{\delta}$ obtained from (16a) is

$$\hat{\underline{\delta}}_I = \mathcal{V}^{-1}T(\mathrm{v}_I - S_I\underline{\beta})\,,$$

where $\mathcal{V} = R_I^\top\Sigma_I^{-1}R_I, T = R_I^\top\Sigma_I^{-1}$ and $Var(\underline{\varepsilon}_I) = \sigma^2\Sigma_I$. Substituting $\hat{\underline{\delta}}_I$ back in (17) yields

$$\begin{aligned}\bar{\mathrm{x}}_{s,s-1} &= \hat{\mathrm{X}}(\mathrm{v})_{s,s-1} - \hat{\mathrm{X}}(\underline{\delta})_{s,s-1}\mathcal{V}^{-1}T\mathrm{v}_I - (\hat{\mathrm{X}}(\underline{\beta})_{s,s-1} - \hat{\mathrm{X}}(\underline{\delta})_{s,s-1}\mathcal{V}^{-1}TS_I)\underline{\beta} \\ &= \bar{\mathrm{X}}(\mathrm{v})_{s,s-1} - \bar{\mathrm{X}}(\underline{\beta})_{s,s-1}\underline{\beta}\,,\end{aligned}$$

where $\bar{\mathrm{x}}_{s,s-1}$ is the estimator of x_s using $(\underline{\beta}^\top, \mathrm{v}_1^\top, \mathrm{v}_2^\top, \cdots, \mathrm{v}_{s-1}^\top)^\top$ and $(\bar{\mathrm{X}}(\mathrm{v})_{s,s-1}, \bar{\mathrm{X}}(\underline{\beta})_{s,s-1})$ are the estimators, respectively, of the states corresponding to the data and the $\underline{\beta}$ parameters. Given that

$$\begin{aligned}\mathrm{Mse}(\bar{\mathrm{x}}_{s,s-1}) =& Var(\mathrm{x}_s - \hat{\mathrm{x}}_{s,s-1} + \hat{\mathrm{x}}_{s,s-1} - \bar{\mathrm{x}}_{s,s-1}) \\ =& Var(\mathrm{x}_s - \hat{\mathrm{x}}_{s,s-1}) + Var(\hat{\mathrm{x}}_{s,s-1} - \bar{\mathrm{x}}_{s,s-1})\,,\end{aligned}$$

we have $\mathrm{Mse}(\tilde{\mathrm{x}}_{s,s-1}) = \sigma^2 P_{s,s-1} + \hat{\mathrm{X}}(\underline{\delta})_{s,s-1}\mathrm{Mse}(\hat{\underline{\delta}}_I)\hat{\mathrm{X}}(\underline{\delta})^{\top}_{s,s-1}$. By Theorem 9 of Section 4, the EKF, to be applied to (16b), can then be initialized with $Var(\mathrm{x}_s) = \mathrm{Mse}(\tilde{\mathrm{x}}_{s,s-1})$ and $\hat{\mathrm{X}}_{s,s-1} = (\tilde{\mathrm{X}}(\mathrm{v})_{s,s-1}, \tilde{\mathrm{X}}(\underline{\beta})_{s,s-1})$. If the DKF is to be employed, the initialization for the Q matrix would be

$$\begin{bmatrix} Q_{11} & Q_{13} \\ Q_{31} & Q_{33} \end{bmatrix} - \begin{bmatrix} Q_{12} \\ Q_{32} \end{bmatrix} Q_{22}^{-1} \begin{bmatrix} Q_{21} & Q_{23} \end{bmatrix},$$

where $Q_s = (Q_{ij})$, $i,j = 1,2,3$. This can be seen considering that, after estimating $\underline{\delta}$, the sum of squares is $(\mathrm{v}_I - S_I\underline{\beta})^{\top} P^{\top} \Sigma_I^{-1} P(\mathrm{v}_I - S_I\underline{\beta})$, with $P = I - R_I(R_I^{\top}\Sigma_I^{-1}R_I)^{-1}R_I^{\top}\Sigma_I^{-1}$. If the first r rows of R do not constitute a submatrix of R of rank r, we would proceed as in the last part of Section 4.

Example 1. (Continued) Model (15) specializes to $\underline{\varepsilon}_k = \mathrm{a}_1 + \cdots \mathrm{a}_k$, $k = 1, \cdots, N$, $R = (1, \cdots, 1)^{\top}$ and $S = (\mathrm{y}_1, \cdots, \mathrm{y}_N)^{\top}$. The AK likelihood is, as in Example 3, the density of the differenced data. If J, J_1 and J_2 are the matrices defined in Example 3, with $J = (J_1^{\top}, J_2^{\top})^{\top}$, then the AK likelihood is the density of

$$J_2(\mathrm{v} - S\underline{\beta}) = (\mathrm{v}_2 - \mathrm{v}_1 - (\mathrm{y}_2 - \mathrm{y}_1)^{\top}\underline{\beta}, \cdots, \mathrm{v}_N - \mathrm{v}_{N-1} - (\mathrm{y}_N - \mathrm{y}_{N-1})^{\top}\underline{\beta})^{\top}.$$

The EKF or DKF produces $D_k = 1, k = 1, \cdots, N, E_1 = (\mathrm{v}_1, 1, \mathrm{y}_1^{\top}), E_k = (\mathrm{v}_k, 0, \mathrm{y}_k^{\top})$, $k = 2, \cdots, N, \ln|\Sigma| = 0, \ln|R^{\top}\Sigma^{-1}R| = 0, \hat{\underline{\delta}} = \mathrm{v}_1 - \mathrm{y}_1^{\top}\hat{\underline{\beta}}$,

$$\hat{\underline{\beta}} = \left[\sum_{k=2}^{N}(\mathrm{y}_k - \mathrm{y}_{k-1})(\mathrm{y}_k - \mathrm{y}_{k-1})^{\top}\right]^{-1} \sum_{k=2}^{N}(\mathrm{y}_k - \mathrm{y}_{k-1})(\mathrm{v}_k - \mathrm{v}_{k-1}),$$

and

$$(\mathrm{v} - X\hat{\underline{\gamma}})^{\top}\Sigma^{-1}(\mathrm{v} - X\hat{\underline{\gamma}}) = \sum_{k=2}^{N}\left[\mathrm{v}_k - \mathrm{v}_{k-1} - (\mathrm{y}_k - \mathrm{y}_{k-1})^{\top}\hat{\underline{\beta}}\right]^2.$$

Model (16a) becomes the first equation of (15), $\mathrm{v}_1 = \underline{\delta} + \mathrm{y}_1^{\top}\underline{\beta} + \mathrm{a}_1$, where $\mathrm{v}_I = \mathrm{v}_1, R_I = 1, S_I = \mathrm{y}_1^{\top}$ and $\underline{\varepsilon}_I = \mathrm{a}_1$. Model (16$b$) consists of the rest of the equations. Suppose we use the EKF or the DKF in (16b) to obtain $\hat{\underline{\delta}}_I$ and then collapse to a reduced dimension EKF or DKF, to be applied later to model (16b). Then,

$$\begin{gathered} \hat{\mathrm{X}}_{1,0} = (0, -1, 0)\,,\ E_1 = (\mathrm{v}_1, 1, \mathrm{y}_1^{\top})\,,\ P_{1,0} = 1, \\ G_1 = 1\,,\ D_1 = 1\,,\ \hat{\mathrm{X}}_{2,1} = (\mathrm{v}_1, 0, \mathrm{y}_1^{\top})\,,\ P_{2,1} = 1\,. \end{gathered}$$

Clearly, $\hat{\underline{\delta}}_I = \mathrm{v}_1 - \mathrm{y}_1^{\top}\underline{\beta}$, and, therefore,

$$\begin{gathered} \tilde{\mathrm{x}}_{2,1} = \mathrm{v}_1 - \mathrm{y}_1^{\top}\underline{\beta} = \tilde{\mathrm{X}}(\mathrm{v})_{2,1} - \tilde{\mathrm{X}}(\underline{\beta})_{2,1}\underline{\beta} \\ \mathrm{Mse}(\tilde{\mathrm{x}}_{2,1}) = 1\,. \end{gathered}$$

The collapsed EXF or DKF can be initialized with $\hat{X}_{2,1} = (v_1, y_1^T)$ and $Var(x_2) = 1$. Note that now the dimension is that of the original EXF or DKF minus one. In case that the DKF is applied, we have

$$Q_2 = \begin{bmatrix} v_1^2 & v_1 & v_1 y_1^T \\ v_1 & 1 & y_1^T \\ y_1 v_1 & y_1 & y_1 y_1^T \end{bmatrix},$$

and the matrix of partial squares and cross products to initialize the collapsed DKF is

$$Q_2 = \begin{bmatrix} v_1^2 & v_1 y_1^T \\ y_1 v_1 & y_1 y_1^T \end{bmatrix} - \begin{bmatrix} v_1 \\ y_1 \end{bmatrix} [v_1 \ \ y_1^T] = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

References

1. Anderson, B. and J. Moore, *Optimal Filtering*, Prentice Hall, Englewood Cliffs, N. J., 1979.
2. Ansley, C. F. and R. Kohn, Estimation, filtering and smoothing in state space models with incompletely specified initial conditions, *Ann. Statist.* 13 (1985), 1286-1316.
3. Bell, W. and S. C. Hillmer, Initializing the Kalman filter for nonstationary time series models, *J. of Time Ser. Anal.* 12 (1991), 283-300.
4. Box, G. E. P. and G. M. Jenkins, *Time Series Analysis, Forecasting and Control*, Holden-Day, San Francisco, CA, 1976.
5. Burridge, P. and K. F. Wallis, Calculating the variance of seasonally adjusted series, *J. of Amer. Statist. Assoc.* 80 (1985), 541-552.
6. Gómez, V. and A. Maravall, Estimation, prediction and interpolation for nonstationary series with the Kalman filter, EUI Working Paper ECO 92/80 (under revision for *J. of Amer. Statist. Assoc.*).
7. Harvey, A. C. and G. D. A. Phillips, Maximum likelihood estimation of regression models with autoregressive-moving average disturbances, *Biometrika* 66 (1979), 49-58.
8. Harvey, A. C. and R. G. Pierse, Estimating missing observations in economic time series, *J. of Amer. Statist. Assoc.* 79 (1984), 125-131.
9. Harvey, A. C., *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge University Press, Cambridge, UK, 1989.
10. Jones, R., Maximum likelihood fitting of ARMA models to time series with missing observations, *Technometrics* 22 (1980), 389-395.
11. Jong, Piet de, The likelihood for the state space model, *Biometrika* 75 (1988), 165-169.
12. Jong, Piet de, Smoothing and interpolation with the state space model, *J. of Amer. Statist. Assoc.* 84 (1989), 408-409.

13. Jong, Piet de, The diffuse Kalman filter, *Ann. Statist.* **19** (1991), 1073-1083.
14. Jong, Piet de, Stable algorithms for the state space model, *J. of Time Ser. Anal.* **12** (1991), 143-156.
15. Kohn, R. and C. F. Ansley, Efficient estimation and prediction in time series regression models, *Biometrika* **72** (1985), 694-697.
16. Kohn, R. and C. F. Ansley, Estimation, prediction and interpolation for ARIMA models with missing data, Tech. Rept., Graduate School of Business, University of Chicago, IL, 1984.
17. Kohn, R. and C. F. Ansley, Estimation, prediction and interpolation for ARIMA models with missing data, *J. of Amer. Statist. Assoc.* **81** (1986), 751-761.
18. Rao, C., *Linear Statistical Inference and its Applications*, John Wiley & Sons, New York, 1973.
19. Wecker, W. and C. F. Ansley, The Signal extraction approach to nonlinear regression and spline smoothing, *J. of Amer. Statist. Assoc.* **78** (1983), 81-89.

Víctor Gómez
Instituto Nacional de Estadistica
Paseo de la Castellana 183
28046 Madrid, Spain

Agustín Maravall
European University Institute
Badia Fiesolana, I-50016
S. Domenico di Fiesole (FI), Italy
maravall@bf.iue.it