

Minimally Conditioned Likelihood for a Nonstationary State Space Model

José Casals[†]

Sonia Sotoca^{††}

Miguel Jerez^{†††}

Universidad Complutense de Madrid

Abstract: Computing the gaussian likelihood for a nonstationary state-space model is a difficult problem which has been tackled by the literature using two main strategies: data transformation and diffuse likelihood. The data transformation approach is cumbersome, as it requires nonstandard filtering. On the other hand, in some nontrivial cases the diffuse likelihood value depends on the scale of the diffuse states, so one can obtain different likelihood values corresponding to different observationally equivalent models. In this paper we discuss the properties of the minimally-conditioned likelihood function, as well as two efficient methods to compute its terms with computational advantages for specific models. Three convenient features of the minimally-conditioned likelihood are: (a) it can be computed with standard Kalman filters, (b) it is scale-free, and (c) its values are coherent with those resulting from differencing, being this the most popular approach to deal with nonstationary data.

Keywords: State-space models; Conditional likelihood; Diffuse likelihood; Diffuse initial conditions; Kalman filter; Nonstationarity

[†] Departamento de Fundamentos del Análisis Económico II. Facultad de Ciencias Económicas. Campus de Somosaguas. 28223 Madrid (SPAIN). Email: jcasalsc@cajamadrid.es

^{††} Departamento de Fundamentos del Análisis Económico II. Facultad de Ciencias Económicas. Campus de Somosaguas. 28223 Madrid (SPAIN). Email: sotoca@ccee.ucm.es

^{†††} **Corresponding author.** Departamento de Fundamentos del Análisis Económico II. Facultad de Ciencias Económicas. Campus de Somosaguas. 28223 Madrid (SPAIN). Email: mjerez@ccee.ucm.es, tel: (+34) 91 394 23 61, fax: (+34) 91 394 25 91.

1. Introduction

The most popular approach to deal with nonstationary data consists of differencing the data to induce stationarity, being this transformation useful both, to specify a model and to compute its gaussian likelihood. This approach is simple and suitable in many cases. Not so much in many others such as, e.g., when one wants to estimate non-multiplicative models, such as time-varying parameter regressions or structural time series models (Harvey, 1989). Also, it results in unnecessary data losses when the sample includes missing values or if the model has cointegration constraints (Mauricio, 2006). Finally, for many practical purposes such as, e.g., forecasting or signal extraction, it is more convenient working with original instead of differenced data. In all these cases, it would be interesting to estimate the nonstationary model.

Computing the likelihood for a model with unit roots is a difficult problem which has been tackled by the state-space literature using two main strategies: data transformation and diffuse initialization.

The most representative work in the data transformation approach is Ansley and Kohn (1985), hereafter AK, who proposed a sophisticated data transformation that cancels the nonstationary components of the model. As AK recognize, their approach has two shortcomings: it needs a complex and nonstandard filtering and requires the data transformation to be independent of the parameter values. This requirement is not fulfilled, for example, when one wants to estimate structural time series models (Harvey, 1989). The AK approach has been further developed in many relevant works (Kohn and Ansley, 1986; Ansley and Kohn, 1990, in the univariate case; Bell and Hillmer, 1991; Gomez and Maravall 1994), but none of them addressed the two preciously mentioned issues.

The diffuse likelihood approach considers an initial state where some components could have an arbitrarily large covariance. Building on this idea, De Jong

(1991) defined the diffuse likelihood function and proved that it is a proper likelihood, as it is based in the data transformation that makes the data invariant to the initial diffuse state. In comparison with the AK algorithm, the main advantage of De Jong (1991) proposal was that it used a standard filter, augmented with the propagation of a vector and a matrix, having each as many rows as the diffuse state vector.

Following also the diffuse initialization strategy, Koopman (1997) proposed decomposing the initial state, \mathbf{x}_1 , as:

$$\mathbf{x}_1 = \mathbf{A}\boldsymbol{\eta} + \mathbf{B}\boldsymbol{\delta} \tag{1.1}$$

where the term $\mathbf{A}\boldsymbol{\eta}$ corresponds to the stationary structure, where \mathbf{A} is a fixed-coefficients matrix and $\boldsymbol{\eta} \sim N(\mathbf{0}, \mathbf{I})$. On the other hand, $\mathbf{B}\boldsymbol{\delta}$ corresponds to the diffuse states, with $\boldsymbol{\delta} \sim N(\mathbf{0}, \kappa\mathbf{I})$ and $\kappa \rightarrow \infty$. Finally \mathbf{B} is a coefficient matrix that must be determined heuristically in each case.

Koopman (1997) computes then the likelihood by running two different filters, which propagate the covariances resulting from the diffuse and stationary subsystems respectively. Both filters collapse to a unique standard Kalman Filter (hereafter, KF) when the number of recursions is sufficient to eliminate the dependence on κ . This algorithm has two weak points. First, the size of the sample required to eliminate this dependence is known only when the model is univariate and there are no missing values; in other cases it must be determined heuristically. Second, the double filtering procedure requires using generalized inverses, being these inverses complex, unstable and computationally expensive.

In this work we present the computation and theoretical advantages of the minimally conditioned likelihood for a state-space model. This approach has three clear benefits. First, in comparison with the data transformation alternatives, it only requires standard filtering. Among other advantages, this means that our procedures can cope with missing data and cointegration constraints. Second it is scale-invariant, while in

some cases the diffuse likelihood depends on the scale of the diffuse states. This is illustrated by the examples in sub-sections 2.1 and 5.2, which show that there could be different diffuse likelihood values corresponding to observationally equivalent models. Third, our method provides likelihood values identical to those resulting from differencing when both approaches can be compared.

The minimally conditioned likelihood function can be efficiently computed by two different but equivalent methods that we call: “State Decomposition” (SD) and “Column Deletion” (CD), respectively.

Section 3 describes the SD method, which is based on some ideas due to De Jong (1988). It builds on a decomposition of the conditional likelihood which separates the effects of both, the diffuse and non-diffuse states. Under these conditions, one can compute the likelihood by applying a KF with null initial conditions to the sample and then correcting the effect of the arbitrary initialization. When the model matrices are time-invariant and there are no missing values in the sample, one can apply the filter simplification proposed by Casals *et al.* (1999) to improve the stability and computational efficiency of the algorithm.

The CD algorithm, described in Section 4, is structurally similar to that of Koopman (1997), as it uses an augmented filter to evaluate recursively the likelihood. Its main advantage in comparison with Koopman’s method is that the columns corresponding to the augmented variables are automatically eliminated as the sample is processed. Therefore, the recursion collapses to a standard KF in the minimum number of iterations and there is no need to set this number heuristically. Second, the augmented equations are efficiently computed using the QR algorithm, thus avoiding the use of generalized inverses.

Section 5 presents two examples illustrating the properties of our methods and Section 6 discusses in detail the relative advantages of both algorithms, provides some

concluding remarks and indicates how to obtain a free MATLAB toolbox which implements the methods described in this paper.

All the proofs for the formal results are given in the Appendices.

2. Different forms of the likelihood function

2.1. Diffuse likelihood

Consider the $m \times 1$ random vector z_t , which is the output of the state-space model:

$$\mathbf{x}_{t+1} = \mathbf{\Phi} \mathbf{x}_t + \mathbf{E} \mathbf{w}_t \quad (2.1)$$

$$z_t = \mathbf{H} \mathbf{x}_t + \mathbf{C} \mathbf{v}_t \quad (2.2)$$

where $\mathbf{\Phi}$, \mathbf{E} , \mathbf{H} , and \mathbf{C} are fixed coefficient matrices, \mathbf{x}_t is a $n \times 1$ vector of state variables and \mathbf{w}_t , \mathbf{v}_t are zero-mean uncorrelated vectors of errors, such that the dimensions of $\mathbf{E} \mathbf{w}_t$ and $\mathbf{C} \mathbf{v}_t$ are $n \times 1$ and $m \times 1$ respectively, with $\text{cov}(\mathbf{w}_t) = \mathbf{Q}$, $\text{cov}(\mathbf{v}_t) = \mathbf{R}$, and $\text{cov}(\mathbf{w}_t, \mathbf{v}_t) = \mathbf{S}$

Note that model (2.1)-(2.2) assumes without loss of generality that: (a) the parameter matrices are time-invariant and (b) there are no exogenous inputs. Assuming the immemorial time hypothesis (De Jong, 1991) the initial state of a nonstationary system includes a diffuse component with infinite uncertainty. It is then easy to isolate this component by applying a similar transformation to the initial state, which yields:

$$\mathbf{M} \mathbf{x}_1 = \begin{bmatrix} \mathbf{x}_1^{\text{D}} \\ \mathbf{x}_1^{\text{ND}} \end{bmatrix} \quad (2.3)$$

where \mathbf{M} is the matrix characterizing the transformation, \mathbf{x}_1^{D} is a $d \times 1$ vector that includes the diffuse states, such that $\text{cov}(\mathbf{x}_1^{\text{D}}) \rightarrow \infty$, and \mathbf{x}_1^{ND} is a $(n-d) \times 1$ vector of stationary components. Denoting $\mathbf{M}^{-1} = [\mathbf{T} \quad \mathbf{G}]$ we can write (2.3) as:

$$\mathbf{x}_1 = \mathbf{T} \mathbf{x}_1^{\text{D}} + \mathbf{G} \mathbf{x}_1^{\text{ND}} \quad (2.4)$$

which is equivalent to the decompositions of De Jong (1991) and Koopman (1997). On this basis, both works discuss the evaluation of the diffuse log-likelihood defined as:

$$\log L_{\infty}(\mathbf{Z}) = \log L(\mathbf{Z}) - \frac{1}{2} \log |\text{cov}(\mathbf{x}_1^{\text{D}})| \quad (2.5)$$

where $\log L_{\infty}(\mathbf{Z})$ denotes the diffuse log-likelihood of model (2.1)-(2.2), $\log L(\mathbf{Z})$ is the corresponding gaussian log-likelihood and \mathbf{Z} is the sample.

AK (1985, Theorem 5.1) and De Jong (1991, Theorem 4.2) proved that (2.5) is a proper log-likelihood, as it is based on the components of \mathbf{Z} which are invariant to \mathbf{x}_1^{D} . That is, it coincides with the log-likelihood of the sample after transforming it to avoid dependence on the diffuse components of the initial state vector. On this basis, De Jong (1991) proposes an evaluation algorithm based on the so-called diffuse KF, while Koopman (1997) suggests using two specialized filters for the diffuse and non-diffuse components, respectively.

The previous approach has a clear shortcoming, as the value of the diffuse likelihood may depend on the scale of the state vector. To see this, consider e.g., the observationally equivalent models:

$$\begin{aligned} x_{t+1} &= x_t + w_t \\ z_t &= \alpha x_t + v_t \end{aligned} \quad (2.6)$$

$$\begin{aligned} x_{t+1}^* &= x_t^* + \alpha w_t \\ z_t &= x_t^* + v_t \end{aligned} \quad (2.7)$$

where α is an arbitrary constant, $\text{var}(w_t) = 1$, $\text{cov}(w_t, v_t) = 0$ and $x_t^* = \alpha x_t$.

According to (2.5), the diffuse likelihood of (2.6) and (2.7) are, respectively:

$$\log L_{\infty}(z|\alpha) = \log L(z) - \frac{1}{2} \log |\text{cov}(x_1)| \quad (2.8)$$

$$\log L_{\infty}^*(z|\alpha) = \log L(z) - \frac{1}{2} \log |\text{cov}(\alpha x_1)| \quad (2.9)$$

and these values do not coincide because $\log L_\infty(\mathbf{z}|\alpha) - \log L_\infty^*(\mathbf{z}|\alpha) = \log(\alpha)$. Note that this problem also affects the first-order derivatives because:

$$\frac{\partial \log L_\infty(\mathbf{z}|\alpha)}{\partial \alpha} = \frac{\partial \log L_\infty^*(\mathbf{z}|\alpha)}{\partial \alpha} + \frac{1}{\alpha} \quad (2.10)$$

Therefore, the values of the diffuse likelihood corresponding to equivalent representations, such as (2.6) and (2.7), can be different.

In general, any linear transformation of the initial diffuse vector such that $\mathbf{x}_1^{\text{D}^*} = \mathbf{L}\mathbf{x}_1^{\text{D}}$ would yield the initial state decomposition $\mathbf{x}_1 = \mathbf{T}\mathbf{L}^{-1}\mathbf{x}_1^{\text{D}^*} + \mathbf{G}\mathbf{x}_1^{\text{ND}}$, see (2.4), with $\text{cov}(\mathbf{x}_1^{\text{D}^*}) \rightarrow \infty$. Under these conditions, the diffuse likelihood would be:

$$\log L_\infty(\mathbf{Z}) = \log L(\mathbf{Z}) - \frac{1}{2} \log |\text{cov}(\mathbf{x}_1^{\text{D}^*})| \quad (2.11)$$

which obviously depends on the transformation matrix $\mathbf{T}\mathbf{L}^{-1}$.

2.2. Conditional likelihood

An alternative to the diffuse likelihood would consist of computing a gaussian likelihood, conditional to the minimum subset of the sample required to eliminate the effect of the diffuse states. As we will see, this strategy is closely related to the diffuse likelihood approach, but is unaffected by the scale of the diffuse states.

It is well known that equation (2.2) can be written in matrix form as:

$$\mathbf{Z} = \mathbf{O}\mathbf{x}_1 + \mathbf{Z}^* \quad (2.12)$$

where \mathbf{Z}^* is the part of the sample that does not depend on \mathbf{x}_1 and \mathbf{O} is the extended observability matrix, defined as:

$$\mathbf{O} = \begin{bmatrix} \mathbf{H} \\ \mathbf{H}\Phi \\ \vdots \\ \mathbf{H}\Phi^{n-1} \end{bmatrix} \quad (2.13)$$

Applying the decomposition (2.4) to (2.12) we obtain:

$$\mathbf{Z} = \mathbf{O}_D \mathbf{x}_1^D + \mathbf{Z}^{ND} \quad (2.14)$$

where $\mathbf{Z}^{ND} = \mathbf{O}_{ND} \mathbf{x}_1^{ND} + \mathbf{Z}^*$, $\mathbf{O}_D = \mathbf{O} \mathbf{T}$ is the extended observability matrix corresponding to the diffuse initial states and $\mathbf{O}_{ND} = \mathbf{O} \mathbf{G}$ is the analogous matrix affecting the non-diffuse initial states. Therefore, \mathbf{Z}^{ND} is the part of the sample that is not affected by the diffuse initial states. Under these conditions, there always exists a matrix \mathbf{A} such that $\mathbf{A}^T \mathbf{O}_D = \mathbf{I}$ with $\text{rank}(\mathbf{A}) = \text{dimension}(\mathbf{x}_1^D)$. Then, premultiplying both sides of (2.14) by \mathbf{A}^T we obtain:

$$\mathbf{A}^T \mathbf{Z} = \mathbf{x}_1^D + \mathbf{A}^T \mathbf{Z}^{ND} \quad (2.15)$$

Denoting $\mathbf{A}^T \mathbf{Z} \equiv \mathbf{U}$ and taking conditional expectations of both sides of (2.14) we obtain:

$$\mathbb{E}(\mathbf{Z} | \mathbf{U}) = \mathbf{O}_D \mathbb{E}(\mathbf{x}_1^D | \mathbf{U}) \quad (2.16)$$

Note that $\mathbb{E}(\mathbf{Z}^{ND} | \mathbf{U}) = \mathbf{0}$ as this conditional expectation depends on the inverse of $\text{cov}(\mathbf{U})$, which is null. Hence, the conditional covariance of the sample is:

$$\text{cov}(\mathbf{Z} | \mathbf{U}) = (\mathbf{I} - \mathbf{O}_D \mathbf{A}^T) \mathbf{V} (\mathbf{I} - \mathbf{O}_D \mathbf{A}^T)^T \quad (2.17)$$

where $\mathbf{V} \equiv \text{cov}(\mathbf{Z}^{ND} | \mathbf{U})$ and the covariance matrix given by (2.17) is finite and computable, as it only depends on the stationary part of \mathbf{Z} .

Choosing $\mathbf{A} = \mathbf{O}_D (\mathbf{O}_D^T \mathbf{O}_D)^{-1}$ yields the transformation proposed by AK (1985). However, there are other valid and more convenient choices for \mathbf{A} . In this paper, we will use $\mathbf{A} = \begin{pmatrix} \mathbf{O}_1 \\ \mathbf{0} \end{pmatrix} (\mathbf{O}_1^T \mathbf{O}_1)^{-1}$, where \mathbf{O}_1 includes the first columns of \mathbf{O}_D so that $\text{rank}(\mathbf{O}_D) = \text{rank}(\mathbf{O}_1)$ and $\mathbf{A}^T \mathbf{O}_D = (\mathbf{O}_1^T \mathbf{O}_1)^{-1} (\mathbf{O}_1^T \quad \mathbf{0}) \begin{pmatrix} \mathbf{O}_1 \\ \mathbf{O}_2 \end{pmatrix} = \mathbf{I}$, where \mathbf{O}_D has been partitioned as $\mathbf{O}_D = \begin{pmatrix} \mathbf{O}_1 \\ \mathbf{O}_2 \end{pmatrix}$. Note that these expressions are particularized for the first observations in the sample, but any other subsample with $\text{rank}(\mathbf{O}_D) = \text{rank}(\mathbf{O}_1)$ would have been a valid choice.

The likelihood of \mathbf{Z} conditional to \mathbf{U} is given by two main terms: (a) the determinant of $\text{cov}(\mathbf{Z}/\mathbf{U})$, given in (2.17) and (b) a weighted sum of squares of the observations. About the first, expression (2.17) immediately implies:

$$|\text{cov}(\mathbf{Z}/\mathbf{U})| = \left| (\mathbf{I} - \mathbf{O}_d \mathbf{A}^T) \mathbf{V} (\mathbf{I} - \mathbf{O}_d \mathbf{A}^T)^T \right| \quad (2.18)$$

and $(\mathbf{I} - \mathbf{O}_d \mathbf{A}^T)$ can be written as:

$$(\mathbf{I} - \mathbf{O}_d \mathbf{A}^T) = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ -\mathbf{O}_2 (\mathbf{O}_1^T \mathbf{O}_1)^{-1} \mathbf{O}_1^T & \mathbf{I} \end{pmatrix} \quad (2.19)$$

Taking into account the structure of (2.19) and applying some well-known algebraic results, (2.18) can be written as:

$$|\text{cov}(\mathbf{Z}/\mathbf{U})| = \frac{|\mathbf{V}| |\mathbf{O}_d^T \mathbf{V}^{-1} \mathbf{O}_d|}{|\mathbf{O}_1^T \mathbf{O}_1|} \quad (2.20)$$

As for the quadratic term, its expression is:

$$\begin{aligned} \mathbf{Z}^T [\text{cov}(\mathbf{Z}/\mathbf{U})]^{-1} \mathbf{Z} &= \\ &= \mathbf{Z}^T \mathbf{V}^{-1} \mathbf{Z} - \mathbf{Z}^T \mathbf{V}^{-1} \mathbf{O}_d [\mathbf{O}_d^T \mathbf{V}^{-1} \mathbf{O}_d]^{-1} \mathbf{O}_d^T \mathbf{V}^{-1} \mathbf{Z} \end{aligned} \quad (2.21)$$

The most efficient way to compute (2.20)-(2.21) consists of applying a standard KF to the observations \mathbf{Z} . If we denote by \mathbf{F} the *en-bloc* linear KF reducing the observations to uncorrelated innovations, $\tilde{\mathbf{Z}} = \mathbf{F}\mathbf{Z}$, then $\mathbf{B} = \mathbf{F}\mathbf{V}\mathbf{F}^T$, where \mathbf{B} is a block-diagonal matrix of innovation variances. Therefore:

$$\mathbf{V}^{-1} = \mathbf{F}^T \mathbf{B}^{-1} \mathbf{F} \quad (2.22)$$

and the quadratic term in (2.21) would be:

$$\begin{aligned} \mathbf{Z}^T [\text{cov}(\mathbf{Z}/\mathbf{U})]^{-1} \mathbf{Z} &= \\ &= \tilde{\mathbf{Z}}^T \mathbf{B}^{-1} \tilde{\mathbf{Z}} - \tilde{\mathbf{Z}}^T \mathbf{B}^{-1} \tilde{\mathbf{O}}_d [\tilde{\mathbf{O}}_d^T \mathbf{B}^{-1} \tilde{\mathbf{O}}_d]^{-1} \tilde{\mathbf{O}}_d^T \mathbf{B}^{-1} \tilde{\mathbf{Z}} \end{aligned} \quad (2.23)$$

where $\tilde{\mathbf{O}}_d$ is defined as $\tilde{\mathbf{O}}_d = \mathbf{F} \mathbf{O}_d$, which is the result of applying a KF to the columns of \mathbf{O}_d . The first addend in the right-hand-side of (2.23) corresponds to the sum of squared innovations associated to a KF with the initial conditions $(\mathbf{G} \bar{\mathbf{x}}_1^{\text{ND}}, \mathbf{P}_1)$.

The second addend is a correction that compensates the effect of conditioning to a minimal subsample over the likelihood.

On the other hand, using the result (2.22) the determinant in (2.20) reduces to:

$$|\text{cov}(\mathbf{Z}/\mathbf{U})| = \frac{|F^{-1}\mathbf{B}(F^T)^{-1}|\mathbf{O}_D^T F^T \mathbf{B}^{-1} F \mathbf{O}_D|}{|\mathbf{O}_1^T \mathbf{O}_1|} = \frac{|\mathbf{B}|\tilde{\mathbf{O}}_D^T \mathbf{B}^{-1} \tilde{\mathbf{O}}_D|}{|\mathbf{O}_1^T \mathbf{O}_1|} \quad (2.24)$$

Finally the conditional log likelihood, ignoring constant terms, would be:

$$\begin{aligned} \ell(\mathbf{Z}|\mathbf{U}) = & \frac{1}{2} [\log|\mathbf{B}| + \log|\tilde{\mathbf{O}}_D^T \mathbf{B}^{-1} \tilde{\mathbf{O}}_D| - \log|\mathbf{O}_1^T \mathbf{O}_1| + \\ & + \tilde{\mathbf{Z}}^T \mathbf{B}^{-1} \tilde{\mathbf{Z}} - \tilde{\mathbf{Z}}^T \mathbf{B}^{-1} \tilde{\mathbf{O}}_D (\tilde{\mathbf{O}}_D^T \mathbf{B}^{-1} \tilde{\mathbf{O}}_D)^{-1} \tilde{\mathbf{O}}_D^T \mathbf{B}^{-1} \tilde{\mathbf{Z}}] \end{aligned} \quad (2.25)$$

Comparing (2.25) with the diffuse likelihood of De Jong (1991, Theorem 4.2) it can be seen that:

$$\ell(\mathbf{Z}|\mathbf{U}) = \log L_\infty(\mathbf{Z}) - \log|\mathbf{O}_1^T \mathbf{O}_1| \quad (2.26)$$

where $\log L_\infty(\mathbf{Z})$ denotes the diffuse log-likelihood. Expression (2.26) implies that the conditional and diffuse log-likelihood functions coincide but for the addend $-\log|\mathbf{O}_1^T \mathbf{O}_1|$ which is very important, as it avoids the undesirable scale effect described in sub-section 2.1.

Result (2.26) can be very useful to implement a likelihood computation procedure. Specifically, if one has the code required to calculate the diffuse log-likelihood, then it would be enough to add the correction $-\log|\mathbf{O}_1^T \mathbf{O}_1|$ to obtain a conditional likelihood algorithm.

On the other hand, expression (2.26) is conceptually important because it characterizes the conditioning set employed. For example, (2.26) is conditional to the beginning of the sample because the correction is computed using \mathbf{O}_1 . It would be easy to compute corrections relying on other sample sub-sets or even to the whole sample, which would require using $-\log|\mathbf{O}_D^T \mathbf{O}_D|$. Note that the conditional likelihood in this case would coincide with the marginal likelihood of Francke, Koopman and de Vos (2010).

Finally, the conditional approach provides likelihood values that are coherent with those obtained by differencing the data; for a formal proof, see Appendix 1.

3. State decomposition (SD) algorithm

The efficiency of the algorithm outlined in Section 2 can be improved by segregating the terms affected by the initial conditions. To this end, consider again the expression (2.14):

$$\mathbf{Z} = \mathbf{O}_D \mathbf{x}_1^D + \mathbf{Z}^{ND} = \begin{bmatrix} \mathbf{O}_D & \mathbf{O}_{ND} \end{bmatrix} \begin{bmatrix} \mathbf{x}_1^D \\ \mathbf{x}_1^{ND} \end{bmatrix} + \mathbf{Z}^* \quad (3.1)$$

where, \mathbf{Z}^* is the part of the sample that does not depend on \mathbf{x}_1 ; $\text{cov}(\mathbf{x}_1^D) \rightarrow \infty$, and $\mathbf{x}_1^{ND} \sim N(\mathbf{G}\bar{\mathbf{x}}_1^{ND}, \mathbf{P}_1)$, with $\mathbf{P}_1 > \mathbf{0}$.

Theorem: Expression (3.1) implies that the determinant (2.20) and the quadratic term (2.21) of the likelihood function can be written, respectively, as:

$$|\text{cov}(\mathbf{Z}|\mathbf{U})| = \frac{|\mathbf{V}^*| \left| \mathbf{\Pi} + \bar{\mathbf{O}}^T (\mathbf{V}^*)^{-1} \bar{\mathbf{O}} \right| |\mathbf{P}_1|}{|\mathbf{O}_1^T \mathbf{O}_1|} \quad (3.2)$$

$$\mathbf{Z}^T \left[\text{cov}(\mathbf{Z}|\mathbf{U}) \right]^{-1} \mathbf{Z} = \mathbf{Z}^T (\mathbf{V}^*)^{-1} \mathbf{Z} - \mathbf{Z}^T (\mathbf{V}^*)^{-1} \bar{\mathbf{O}} \left[\mathbf{\Pi} + \bar{\mathbf{O}}^T (\mathbf{V}^*)^{-1} \bar{\mathbf{O}} \right]^{-1} \bar{\mathbf{O}}^T (\mathbf{V}^*)^{-1} \mathbf{Z} \quad (3.3)$$

where $\mathbf{\Pi} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_1^{-1} \end{bmatrix}$; $\mathbf{V}^* \equiv \text{cov}(\mathbf{Z}^*|\mathbf{U})$ is finite and $\bar{\mathbf{O}} = [\mathbf{O}_D \quad \mathbf{O}_{ND}]$

Proof. See Appendix 2.

The main advantage of using (3.2)-(3.3) instead of (2.20)-(2.21) is that these expressions separate the effects of both, the diffuse and non-diffuse initial states. This allows us to apply an idea due to De Jong (1988), consisting of computing efficiently

the likelihood by propagating a KF with initial conditions $(\mathbf{G} \bar{\mathbf{x}}_1^{\text{ND}}, \mathbf{0})$ and afterwards correcting the effect of this ad-hoc initialization. This approach yields the simplified KF:

$$\tilde{\mathbf{z}}_t = \mathbf{z}_t - \mathbf{H} \mathbf{x}_{t|t-1} \quad (3.4)$$

$$\mathbf{B}_t = \mathbf{H} \mathbf{P}_{t|t-1} \mathbf{H}^T + \mathbf{C} \mathbf{R} \mathbf{C}^T \quad (3.5)$$

$$\mathbf{K}_t = (\Phi \mathbf{P}_{t|t-1} \mathbf{H}^T + \mathbf{E} \mathbf{S} \mathbf{C}^T) \mathbf{B}_t^{-1} \quad (3.6)$$

$$\hat{\mathbf{x}}_{t+1|t} = \Phi \hat{\mathbf{x}}_{t|t-1} + \mathbf{K}_t \tilde{\mathbf{z}}_t \quad (3.7)$$

$$\begin{aligned} \mathbf{P}_{t+1|t} = & \bar{\Phi}_t \mathbf{P}_{t|t-1} (\bar{\Phi}_t)^T + \mathbf{E} \mathbf{Q} \mathbf{E}^T + \mathbf{K}_t \mathbf{C} \mathbf{R} \mathbf{C}^T (\mathbf{K}_t)^T - \\ & - \mathbf{K}_t \mathbf{E} \mathbf{S} \mathbf{C}^T - \mathbf{C} \mathbf{S}^T \mathbf{E}^T (\mathbf{K}_t)^T \end{aligned} \quad (3.8)$$

where $\tilde{\mathbf{z}}_t$ are the innovations, \mathbf{B}_t its covariance matrix, \mathbf{K}_t is the KF gain, $\hat{\mathbf{x}}_{t+1|t}$ is an estimate of the state vector at time $t+1$ conditional to the information available up to time t , $\mathbf{P}_{t+1|t}$ is its covariance and $\bar{\Phi}_t = \Phi - \mathbf{K}_t \mathbf{H}$. This filter is augmented with the additional equations:

$$\mathbf{w}_t = \mathbf{w}_{t-1} + \left(\bar{\bar{\Phi}}_{t-1} \right)^T (\mathbf{H})^T \mathbf{B}_t^{-1} \tilde{\mathbf{z}}_t \text{ with } \mathbf{w}_0 = 0 \quad (3.9)$$

$$\mathbf{W}_t = \mathbf{W}_{t-1} + \left(\bar{\bar{\Phi}}_{t-1} \right)^T (\mathbf{H})^T \mathbf{B}_t^{-1} \mathbf{H} \bar{\bar{\Phi}}_{t-1} \text{ with } \mathbf{W}_0 = 0 \quad (3.10)$$

$$\bar{\bar{\Phi}}_t = \bar{\bar{\Phi}}_t \bar{\bar{\Phi}}_{t-1} \text{ with } \bar{\bar{\Phi}}_0 = \mathbf{I} \quad (3.11)$$

where the terms $\bar{\mathbf{O}}^T (\mathbf{V}^*)^{-1} \mathbf{Z}$ and $\bar{\mathbf{O}}^T (\mathbf{V}^*)^{-1} \bar{\mathbf{O}}$ in (3.3) and (3.2), respectively, are given by \mathbf{w}_N and \mathbf{W}_N , defined in (3.9) and (3.10). The conditional log-likelihood would then be:

$$\begin{aligned} \ell(\mathbf{Z}/\mathbf{U}) = & \frac{1}{2} \left\{ \left(\sum_{t=1}^T [\tilde{\mathbf{z}}_t^T \mathbf{B}_t^{-1} \tilde{\mathbf{z}}_t + \log |\mathbf{B}_t|] - \mathbf{w}_N^T \mathbf{M}^T (\mathbf{\Pi} + \mathbf{M}^T \mathbf{W}_N \mathbf{M})^{-1} \mathbf{M} \mathbf{w}_N + \right. \right. \\ & \left. \left. + \log |\mathbf{P}_1| + \log |\mathbf{\Pi} + \mathbf{M}^T \mathbf{W}_N \mathbf{M}| - \log |\mathbf{O}_1^T \mathbf{O}_1| \right) + [\log(2\pi)](N-d) \right\} \end{aligned} \quad (3.12)$$

where T is the sample size, $N = T m$, being m the dimension of \mathbf{z}_t , and d is the number of diffuse states. If the sample includes some missing values the value of T must be adjusted accordingly.

Initializing the filter (3.4)-(3.11) with $(\mathbf{G}\bar{\mathbf{x}}_1^{\text{ND}}, \mathbf{0})$ simplifies the propagation equations. Specifically, in a time-invariant innovations model (e.g., VARMAX) with no missing values, the solution to the Riccati algebraic equation associated to (3.8) is null and, therefore, the initial condition $\mathbf{P}_1 = \mathbf{0}$ implies that $\mathbf{P}_{t|t-1} = \mathbf{0} \forall t$ (Casals *et al.* 1999). This property simplifies the likelihood computation because if $\mathbf{P}_{t|t-1} = \mathbf{0} \forall t$, it is not necessary to propagate equations (3.5), (3.6) and (3.8). Additional efficiency can be obtained by computing the term $|\mathbf{O}_1^T \mathbf{O}_1|$ and the number of diffuse initial states, d , by applying the QR decomposition to the matrix $\mathbf{H}\mathbf{T}$, where \mathbf{T} denotes the matrix in (2.4), see Appendix 3.

The simplification described above can be extended to any general time invariant model, see Casals *et al.* (1999), so it provides a very efficient way to evaluate the minimally-conditioned likelihood for many common representations such as, e.g., VARMAX or structural time series models.

4. Column deletion (CD) algorithm

The conditional log-likelihood $\ell(\mathbf{Z}|\mathbf{U})$ given in (2.26) can also be computed by an alternative column deletion algorithm, which is structurally similar to Koopman (1997) method.

Defining $\mathbf{x}_t = \mathbf{x}_t^d + \mathbf{x}_t^{nd}$, with $\mathbf{x}_1^d = \mathbf{T}\mathbf{x}_1^D$, see (2.4), we can write the state equation (2.1) as: $\mathbf{x}_{t+1}^d + \mathbf{x}_{t+1}^{nd} = \mathbf{\Phi}(\mathbf{x}_t^d + \mathbf{x}_t^{nd}) + \mathbf{E}\mathbf{w}_t$, and then, break it into the corresponding diffuse (superindex “d”) and non-diffuse (superindex “nd”) equations:

$$\mathbf{x}_{t+1}^d = \mathbf{\Phi}\mathbf{x}_t^d \tag{4.1}$$

$$\mathbf{x}_{t+1}^{nd} = \mathbf{\Phi}\mathbf{x}_t^{nd} + \mathbf{E}\mathbf{w}_t \tag{4.2}$$

Accordingly, (2.2) can be written as $\mathbf{z}_t = \mathbf{H}(\mathbf{x}_t^d + \mathbf{x}_t^{nd}) + \mathbf{C}\mathbf{v}_t$ or, equivalently, as:

$$z_t = \mathbf{H}x_t^d + z_t^{nd} \quad (4.3)$$

where $z_t^{nd} = \mathbf{H}x_t^{nd} + C v_t$ is the non-diffuse component of the endogenous variables. In $t=1$, the observer (4.3) would be: $z_1 = \mathbf{H}x_1^d + z_1^{nd}$, or, in compare notation:

$$z_1 = \mathbf{H}^* x_1^D + z_1^{nd} \quad (4.4)$$

where $\mathbf{H}^* = \mathbf{H} \mathbf{T}$. This expression shows that z_1 depends on the diffuse vector x_1^D , which uncertainty is infinite. Accordingly, it is not possible to determine the likelihood of z_1 or to apply a standard KF to compute the likelihood of the sample.

This problem can be tackled by:

- 1) Decomposing x_1^D into two components, one formed by the linear combinations of x_1^D which affect z_1 , and another one which components affect the rest of the sample (z_2, z_3, \dots, z_N).
- 2) Estimating the part of x_1^D which depends on z_1 , conditional to this value.
- 3) Repeating step 2) by successively including the values z_2, z_3, \dots, z_N until the dimensions of the term affected by the diffuse conditions collapse to zero.

The following Subsections describe in detail these steps.

4.1 First Step: Decomposition

Consider the matrix \mathbf{Q} , which spans a d -dimensional space, such that it can be partitioned as $\mathbf{Q} = [\mathbf{Q}_H \quad \mathbf{Q}_H^\perp]$, where \mathbf{Q}_H is the $d \times d_1$ matrix that generates the row-subspace of \mathbf{H}^* and the $d \times d_2$ matrix \mathbf{Q}_H^\perp generates the subspace orthogonal to that of \mathbf{H}^* , so that $\mathbf{H}^* \mathbf{Q}_H^\perp = \mathbf{0}$. The dimensions of these matrices are $d_1 + d_2 = d$, with $d_1 \leq m$. Under these conditions x_1^D can be decomposed as follows:

$$x_1^D = \mathbf{Q}_H \alpha_1 + \mathbf{Q}_H^\perp \alpha_1^\perp \quad (4.5)$$

where α_1 and α_1^\perp are $(d_1 \times 1)$ and $(d_2 \times 1)$ vectors of diffuse initial states respectively.

Substituting (4.5) in (4.4) yields:

$$z_1 = \mathbf{H}^{**} \alpha_1 + z_1^{nd} \quad (4.6)$$

where $\mathbf{H}^{**} = \mathbf{H}^* \mathbf{Q}_H$ is a full rank matrix, with $\text{rank}(\mathbf{H}^{**}) = d_1 \leq m$. Note that the number of diffuse states that affect z_1 reduces in this first step from d to d_1 .

In this situation we can estimate of α_1 conditional to the information in z_1 :

$$\hat{\alpha}_1 = \left[(\mathbf{H}^{**})^T \mathbf{B}_1^{-1} \mathbf{H}^{**} \right]^{-1} (\mathbf{H}^{**})^T \mathbf{B}_1^{-1} z_1 \quad (4.7)$$

with the conditional covariance:

$$c\hat{\text{ov}}(\hat{\alpha}_1 | z_1) = \left[(\mathbf{H}^{**})^T \mathbf{B}_1^{-1} \mathbf{H}^{**} \right]^{-1} \quad (4.8)$$

being $\mathbf{B}_1 = \text{cov}(z_1^{nd})$ a computable and finite value matrix which can always be inverted since \mathbf{H}^{**} is full rank, see De Jong (1988).

Since there is a part of x_1^D that can be estimated with the information in z_1 , it would be convenient to derive a specialized filter for (4.5)-(4.6) such that the propagation of the diffuse states distinguishes the part corresponding to α_1 , which uncertainty conditional to z_1 is finite and, accordingly, should be taken into account. Therefore, we can re-organize (4.1)-(4.3) at $t=1$, taking into account (4.5) as:

$$x_2^{d*} = \Phi x_1^{d*} \text{ with } x_1^{d*} = \mathbf{T} \mathbf{Q}_H^\perp \alpha_1^\perp$$

Taking into account that $x_2^{nd} = \Phi x_1^{nd} + \mathbf{E} w_1$, we obtain:

$$x_2^{nd*} = \Phi \mathbf{T} \mathbf{Q}_H \alpha_1 + x_2^{nd} \quad (4.9)$$

and:

$$z_1 = \mathbf{H} \mathbf{T} \mathbf{Q}_H \alpha_1 + z_1^{nd} \quad (4.10)$$

Building on (4.9)-(4.10) and Casals, Jerez and Sotoca (2000, pp. 61) the estimates for the mean and variance of x_2^{nd*} , conditional on z_1 , are:

$$\hat{x}_{2|1}^{nd*} = \hat{x}_{2|1}^{nd} + (\Phi - \mathbf{K}_1 \mathbf{H}) \mathbf{T} \mathbf{Q}_H \hat{\alpha}_1 \quad (4.11)$$

$$\mathbf{P}_{2|1}^{nd*} = \mathbf{P}_{2|1}^{nd} + (\mathbf{\Phi} - \mathbf{K}_1 \mathbf{H}) \mathbf{T} \mathbf{Q}_H \cdot c\hat{ov}(\hat{\alpha}_1 | z_1) \mathbf{Q}_H^T \cdot \mathbf{T}^T (\mathbf{\Phi} - \mathbf{K}_1 \mathbf{H})^T \quad (4.12)$$

and $\hat{\mathbf{x}}_{2|1}^{nd}$, $\mathbf{P}_{2|1}^{nd}$ and \mathbf{K}_1 (the KF gain) can be computed by applying a standard KF to the stationary subsystem. On the other hand, the estimate, $\hat{\alpha}_1$, and its covariance, $c\hat{ov}(\hat{\alpha}_1 | z_1)$, are given by (4.7) and (4.8).

4.2 Second Step: Estimation

In $t=2$, $z_2 = \mathbf{H}\mathbf{\Phi}\mathbf{x}_1^{d*} + z_2^{nd*}$, so:

$$z_2 = \mathbf{H}\mathbf{\Phi}\mathbf{T}\mathbf{Q}_H^\perp \cdot \alpha_1^\perp + z_2^{nd*} \quad (4.13)$$

where the states α_1^\perp affect z_2 but do not affect z_1 . Then, the second innovation can be written as $\tilde{z}_2 = z_2 - \hat{z}_{2|1} = \mathbf{H}\mathbf{x}_2^{d*} + \tilde{z}_2^{nd*}$ and we are in the same situation as in $t = 1$ because $cov(\tilde{z}_2^{nd*})$ is finite and has the expression:

$$cov(\tilde{z}_2^{nd*}) = \mathbf{H} \mathbf{P}_{2|1}^{nd*} \mathbf{H}^T + \mathbf{C}\mathbf{R}\mathbf{C}^T \quad (4.14)$$

but comparing (4.13) with (4.6) it is immediate to see that the number of diffuse initial states is now the dimension of α_1^\perp , that is $d_2 = d - d_1 \leq d$.

We will therefore do the same as in the first step: (a) use the results in De Jong (1988) to estimate α_1^\perp and its variance, conditional to z_2 , (b) apply the smoother due to Casals, Jerez and Sotoca (2000) to condition it to z_2 , and (c) obtain an estimate of the stationary sub-system state vector and its covariance, so that the diffuse initial conditions will not affect the filtering results, see (4.11)-(4.12).

4.3 Third Step: Filtering

By induction, it is easy to see that, at any t , the procedure given by the two previous steps reduces to an augmented filter, including the following standard KF equations:

$$\tilde{z}_t = z_t - \mathbf{H} \hat{\mathbf{x}}_{|t-1} \quad (4.15)$$

$$\mathbf{B}_t = \mathbf{H} \mathbf{P}_{t|t-1} \mathbf{H}^\top + \mathbf{C} \mathbf{R} \mathbf{C}^\top \quad (4.16)$$

$$\mathbf{K}_t = (\Phi \mathbf{P}_{t|t-1} \mathbf{H}^\top + \mathbf{E} \mathbf{S} \mathbf{C}^\top) \mathbf{B}_t^{-1} \quad (4.17)$$

$$\hat{\mathbf{x}}_{t+1|t}^* = \Phi \hat{\mathbf{x}}_{t|t-1} + \mathbf{K}_t \tilde{\mathbf{z}}_t \quad (4.18)$$

$$\begin{aligned} \mathbf{P}_{t+1|t}^* &= \bar{\Phi}_t \mathbf{P}_{t|t-1} (\bar{\Phi}_t)^\top + \mathbf{E} \mathbf{Q} \mathbf{E}^\top + \mathbf{K}_t \mathbf{C} \mathbf{R} \mathbf{C}^\top (\mathbf{K}_t)^\top - \\ &\quad - \mathbf{K}_t \mathbf{E} \mathbf{S} \mathbf{C}^\top - \mathbf{C} \mathbf{S}^\top \mathbf{E}^\top (\mathbf{K}_t)^\top \end{aligned} \quad (4.19)$$

with $\bar{\Phi}_t = \Phi - \mathbf{K}_t \mathbf{H}$. On the other hand, the augmented filter equations are:

$$\mathbf{A}_t = \left[\mathbf{H}_t^{**} \mathbf{B}_t^{-1} (\mathbf{H}_t^{**})^\top \right]^{-1} \quad (4.20)$$

$$\mathbf{a}_t = \mathbf{A}_t (\mathbf{H}_t^{**})^\top \mathbf{B}_t^{-1} \tilde{\mathbf{z}}_t \quad (4.21)$$

$$\mathbf{P}_{t+1|t} = \mathbf{P}_{t+1|t}^* + \bar{\Phi}_t \mathbf{T}_t \mathbf{Q}_{\mathbf{H}_t} \mathbf{A}_t (\mathbf{Q}_{\mathbf{H}_t})^\top (\mathbf{T}_t)^\top (\bar{\Phi}_t)^\top \quad (4.22)$$

$$\hat{\mathbf{x}}_{t+1|t} = \hat{\mathbf{x}}_{t+1|t}^* + \bar{\Phi}_t \mathbf{T}_t \mathbf{Q}_{\mathbf{H}_t} \mathbf{a}_t \quad (4.23)$$

$$\mathbf{T}_{t+1} = \Phi \mathbf{T}_t \mathbf{Q}_{\mathbf{H}_t}^\perp \quad (4.24)$$

where $\mathbf{H}_t^{**} = \mathbf{H} \mathbf{T}_t \mathbf{Q}_{\mathbf{H}_t}$

Obviously, these equations are required only if \mathbf{T}_t has not null dimension. In this case they would simplify to: $\hat{\mathbf{x}}_{t+1|t}^* = \hat{\mathbf{x}}_{t+1|t}$ and $\mathbf{P}_{t+1|t}^* = \mathbf{P}_{t+1|t}$

The matrices $\mathbf{Q}_{\mathbf{H}_t}$, $\mathbf{Q}_{\mathbf{H}_t}^\perp$ and \mathbf{H}_t^{**} can be obtained efficiently by applying a column pivoting QR decomposition to $\mathbf{H}_t^* = \mathbf{H} \mathbf{T}_t$, see Appendix 2, and the initial conditions are those corresponding to the stationary subsystem, that is, $(\mathbf{G} \bar{\mathbf{x}}_1^{\text{ND}}, \mathbf{P}_1)$.

The basic idea behind this procedure is that the number of columns of \mathbf{T}_{t+1} in (4.24) is the number of columns of \mathbf{T}_t minus the rank of the matrix \mathbf{H}_t^{**} , defined in (4.20). Therefore, the dimension of \mathbf{T}_t decreases with the number of observations processed and, in a finite number of iterations, its dimension will collapse to zero. Once

this critical size has been achieved, the augmented equations are no longer needed and the filter collapses to a standard KF for stationary systems.

4.4 Likelihood evaluation

Given the results in the previous sub-sections, we will now discuss the analytical expression of the likelihood function, conditional on the minimum number of observations required to determine the diffuse initial conditions.

Consider a system such as (2.1)-(2.2), with diffuse initial states and, therefore, with a finite conditional covariance and an infinite unconditional uncertainty. Its innovations, $\tilde{\mathbf{z}}_t$, can be decomposed as:

$$\tilde{\mathbf{z}}_t = \mathbf{H}_t^{**} \boldsymbol{\alpha}_t + \tilde{\mathbf{z}}_t^{nd} \quad (4.25)$$

where $\tilde{\mathbf{z}}_t^{nd}$ is the stationary component and $\boldsymbol{\alpha}_t$ is a vector of $d_t < n$ diffuse states. Under these conditions, if we define \mathbf{U} , the subset of $\tilde{\mathbf{Z}}$ corresponding to the first d_t linearly independent rows of \mathbf{H}_t^{**} , the Gaussian log-likelihood of $\tilde{\mathbf{Z}}$ conditional to \mathbf{U} according to (2.26) would be:

$$\begin{aligned} \ell(\mathbf{Z}/\mathbf{U}) = \frac{1}{2} \sum_{t=1}^T \{ & \tilde{\mathbf{z}}_t^T \mathbf{B}_t^{-1} \tilde{\mathbf{z}}_t - \mathbf{a}_t^T \mathbf{A}_t^{-1} \mathbf{a}_t + \log |\mathbf{B}_t| - \log |\mathbf{A}_t| - \\ & - \log |\mathbf{H}_{t0}^{*T} \mathbf{H}_{t0}^*| + [\log(2\pi)](N - d) \} \end{aligned} \quad (4.26)$$

where \mathbf{H}_{t0}^* includes the first d_t linearly independent rows of \mathbf{H}_t^{**} and the values of T , N and d are defined as in (3.12). Note that:

- 1) All the terms in (4.26) can be evaluated by propagating the filter (4.15)-(4.24). In particular, the computation of the term $\log |\mathbf{H}_{t0}^{*T} \mathbf{H}_{t0}^*|$ and the number of initial diffuse states, d , are obtained as by-products of the QR decomposition, see Appendix 3.
- 2) This procedure is efficient, as its only computational overhead in comparison with evaluating the likelihood of a stationary system, results from the

calculation of \mathbf{a}_t and \mathbf{A}_t , given in (4.20) and (4.21), until these terms collapse to zero. Furthermore, it does not require generalized inverses, see Koopman (1997), eqs. (11)-(12).

- 3) The term $\log|\mathbf{H}_{t_0}^{*T}\mathbf{H}_{t_0}^*|$ in (4.26) is the difference between the diffuse and the minimally-conditioned likelihood and, therefore, is equivalent to the addend in (2.26).
- 4) Last, when the model does not include cointegration restrictions and the sample does not have missing values, the log-likelihood values given by (3.12) and (4.26) coincide with those obtaining by differencing the data, see Appendix 1. In the cointegration or missing value cases our method is more efficient because differencing leads to a loss of sample information.

5. Examples

5.1. Airline model

This example illustrates the consistency of the conditional likelihood approach and its ability to work with missing values. To this end, we will use the famous series G of international airline passengers, from January 1949 to December 1960, see Box, Jenkins and Reinsel (1994).

Table 1 compares the estimation results obtained with the full sample for both, for the stationary $\text{MA}(1) \times \text{MA}(1)_{12}$ airline model:

$$z_t = (1 - \theta B)(1 - \Theta B^{12})\varepsilon_t \quad (5.1)$$

and the nonstationary version of (5.1):

$$(1 - B)(1 - B^{12})\log P_t = (1 - \theta B)(1 - \Theta B^{12})\varepsilon_t \quad (5.2)$$

where P_t is the number of airline passengers at time t , B is the backshift operator, $z_t = (1 - B)(1 - B^{12})\log P_t$ and the error variance is $\text{var}(\varepsilon_t) = \sigma_\varepsilon^2$.

[Insert Table 1]

Note that:

- 1) the estimates displayed in the table are identical up to the third decimal place; in fact, the differences between actual values are in the order of 10^{-14} ,
- 2) in the case of the stationary model (5.1), the diffuse and conditional likelihood values on convergence are identical, but
- 3) when one estimates the nonstationary model (5.2) the conditional likelihood value on convergence is identical to those obtained with model (5.1), while the diffuse likelihood value is substantially smaller.

Consistency of the likelihood values over various difference orders may be important when one wants to apply common econometric tools such as LR tests or Information criteria.

A residual analysis of the previous model shows that observations # 62 and 135 may be impulse-type outliers. An efficient way to deal with these values consists of tagging them as missing values, see Gomez, Maravall and Peña (1999). Table 2 compares the estimates of models (5.1) and (5.2) obtained for the sample with these missing values. Note that the estimates corresponding to the stationary and nonstationary models are remarkably different. This happens because differencing propagates the missing values, thus destroying potentially valuable sample information. In this case working with model (5.2) is certainly more adequate.

[Insert Table 2]

5.2. Dynamic factor model

Consider two observable time series generated by a common dynamic factor:

$$\begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} = \begin{bmatrix} \alpha \\ \beta \end{bmatrix} f_t + \begin{bmatrix} a_{1t} \\ a_{2t} \end{bmatrix} \quad (5.3)$$

where α and β are unknown parameters and the common factor, f_t , is given by the process:

$$(1 - \phi_f B)(1 - B)f_t = \varepsilon_t \quad (5.4)$$

with a stationary autoregressive parameter ϕ_f . Finally, the error terms ε_t , a_{1t} and a_{2t} are mutually independent gaussian white noise sequences with constant variances σ_ε^2 , σ_1^2 and σ_2^2 .

As it is well known, factor models such as (5.3)-(5.4) are not identified. To estimate them one must therefore impose a normalizing constraint.

Tables 3 and 4 summarize the results of an exercise consisting of: simulating 100 values of y_{1t} and y_{2t} , with the true parameter values indicated in the first column, and then estimating the model parameters, considering different normalizing constraints and likelihood functions.

[Insert Table 3]

[Insert Table 4]

The results in Table 3 display a remarkable stability. Parameter estimates and likelihood values on convergence are practically identical, clearly showing that different normalizing constraints yield observationally equivalent models. On the other hand, the diffuse likelihood results in Table 4 are practically identical and good enough for the normalizing constraints $\alpha = 1$ and $\beta = 1$, but change substantially when the constraint is $\sigma_\varepsilon = .1$. This sensitivity to the normalizing constraint is due to the last addend in (2.26) which, for this model, depends on the parameters to be estimated. As this

example shows, its omission in the diffuse likelihood can be the source of substantial changes in the parameter estimates.

6. Concluding remarks

In this paper we discussed the minimally conditioned likelihood for a state-space model, allowing for unit roots. This approach is relatively simple, as it is based on a standard KF, and has specific advantages in comparison with data transformation, diffuse likelihood and differencing.

About the former, our method avoids using nonstandard filters and is general, meaning that it allows for missing data and can be applied to any dynamic model including, for example, cointegrated structures.

Our approach also has advantages in comparison with diffuse likelihood because, as we showed in sub-section 2.1, the diffuse likelihood value depends in some cases on the scale of the state vector. Our log-likelihood includes a normalizing addend, see Exp. (2.26), which avoids this problem by making it insensitive to scale factors. In some nontrivial cases such as, e.g., dynamic factor models or models with cointegration constraints, this addend depends on the parameters to be estimated. As the example in sub-section 5.2 shows, ignoring this term makes the estimates sensitive to identifying constraints that should be neutral.

Last, we have proved that the minimally conditioned likelihood is consistent with the results provided by differencing (see Appendix 1) so, when both methods are comparable, their results are identical. On the other hand our method is more complex than differencing, but has many advantages as it is: (a) more flexible, as it can be applied to estimate non-multiplicative models, such as time-varying parameter regressions or structural time series models; (b) more efficient when there are

cointegration constraints or missing in-sample, because it avoids unnecessary data losses; and (c) more convenient when one wants to compute forecasts or apply a signal extraction procedure.

The terms of the minimally conditioned likelihood can be computed using either the SD or CD procedures described in Sections 3 and 4, respectively. Both methods are mutually consistent in the sense that, when applied to a given sample and model, they return the same likelihood value, allowing for insignificant numerical differences. Despite this equivalence, each one has specific advantages in different situations.

Specifically, the filters required by both algorithms have some computational overhead in comparison with a standard KF. In particular, the SD procedure requires propagating equations (3.9)-(3.11) for all t , while the CD method only requires additional calculations until the augmented filter collapses to a standard KF. Furthermore, the CD algorithm includes an efficient and stable method, based on the QR decomposition, to include the diffuse initial conditions in the filter. As a consequence, the CD procedure is more efficient in the general case. However, the SD method is computationally cheaper when there are no missing values in the sample and the model parameters are time-invariant. This happens because, under these conditions, one can take advantage of the simplifications derived from the null solution to the Riccati equation, which more than compensate its intrinsic overhead.

Computational efficiency in both cases is further enhanced by the QR algorithm employed to compute the determinants $\log|\mathbf{O}_1^T \mathbf{O}_1|$ or $\log|\mathbf{H}_{t_0}^{*T} \mathbf{H}_{t_0}^*|$ which, when computed using standard approaches, can add a substantial computational overhead.

The procedures described in this article are implemented in the E^4 functions “lfsd” (SD method) and “lfcd” (CD method). E^4 is a MATLAB toolbox for time series modeling, which can be downloaded at: www.ucm.es/info/icae/e4. The source code for all the functions in the toolbox is freely provided under the terms of the GNU General

Public License. This site also includes a complete user manual and other reference materials.

References

- Ansley, C.F. and R. Kohn (1985). "Estimation, Filtering and Smoothing in State Space Models with Incompletely Specified Initial Conditions," *Annals of Statistics* 13, 1286-1316.
- Ansley, C.F. and Kohn, R. (1990). "Filtering and Smoothing in State Space Models with Partially Diffuse Initial Conditions," *Journal Time Series Analysis* 11, 275-93.
- Bell, W., and Hillmer, S. C. (1991). "Initializing the Kalman Filter for Nonstationary Time Series Models," *Journal of Time Series Analysis*, 12, 4, 283-300.
- Box, G.E.P.; G. M. Jenkins y G. C. Reinsel (2008). *Time Series Analysis: Forecasting and Control*, Wiley, New York.
- Casals, J., S. Sotoca and M. Jerez (1999) "A Fast and Stable Method to Compute the Likelihood of Time Invariant State-Space Models," *Economics Letters*, 65, 329-337.
- Casals, J., M. Jerez and S. Sotoca (2000). "Exact Smoothing for Stationary and Nonstationary Time Series," *International Journal of Forecasting* 16, 1, 59-69.
- De Jong, P. (1988). "The Likelihood for a State Space Model," *Biometrika* 75, 1, 165-169.
- De Jong, P. (1991). "The Diffuse Kalman Filter," *Annals of Statistics* 19, 1073-1083.
- De Jong, P and S. Chu-Chun-Lin (1994). "Fast Likelihood Evaluation and Prediction for Nonstationary State Space Models," *Biometrika*, 81, 133-142.
- Francke, M.K., S. J. Koopman and A. de Vos (2010). "Likelihood Functions for State Space Models with Diffuse Initial Conditions," *Journal of Time Series Analysis*, 31, 6, 407-414.
- Gomez, V. and Maravall, A. (1994). "Estimation, Prediction and Interpolation for Nonstationary Series with the Kalman Filter," *Journal of the American Statistical Association*, 89, 611-624.

- Gomez, V., Maravall, A., and D. Peña, (1999). "Missing Observations in ARIMA Models: Skipping Approach versus Additive Outlier Approach," *Journal of Econometrics*, 88, 2, 341-363.
- Harvey, A.C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge University Press, Cambridge (UK).
- Kohn, R. and Ansley, C.F. (1986). "Estimation, Prediction, and Interpolation for ARIMA Models with Missing Data," *Journal of the American Statistical Association*, 81, 751-761.
- Koopman, S.J. (1997). "Exact Initial Kalman Filtering and Smoothing for Non-Stationary Time Series Models," *Journal of the American Statistical Association*, 92, 440, 1630-1638.
- Mauricio, A. (2006). "Exact Maximum Likelihood Estimation of Partially Nonstationary Vector ARMA Models," *Computational Statistics and Data Analysis* 50, 12, 3644-3662.

Acknowledgements: Financial support from Universidad Complutense de Madrid, ref. GR35/10-B-940223 and Plan Nacional de I+D+i. through grants ECO2008-02588/ECON and ECO2011-23972, is gratefully acknowledged.

APPENDIX 1. Equivalence between the likelihood of differenced data and the minimally-conditioned likelihood.

Consider the multivariate stochastic process $\{\mathbf{z}_t\}_{t=1}^N$ and assume, without loss of generality, that it has m first-order integrated components, such that its first-order difference, $\mathbf{w}_t = (1-B)\mathbf{z}_t$, is a stationary and invertible process. This stationary transformation can be written in compact form as: $\mathbf{w} = \mathbf{F} \mathbf{z}$, where \mathbf{z} and \mathbf{w} are $N \times 1$ vectors and \mathbf{F} is a $(m \times N) \times (m \times N)$ block-matrix, composed of $m \times m$ null and identity matrices, with the following structure:

$$\mathbf{F} = \begin{bmatrix} \mathbf{I} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ -\mathbf{I} & \mathbf{I} & \dots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & -\mathbf{I} & \mathbf{I} \end{bmatrix}$$

Under these conditions, the Gaussian density of \mathbf{z} is:

$$f(\mathbf{z}) = f(\mathbf{F} \mathbf{z}) = f(\mathbf{z}_1, \mathbf{w}_2, \dots, \mathbf{w}_N)$$

where \mathbf{z}_1 contains the first observation in the sample, $\mathbf{w}_2 = \mathbf{z}_2 - \mathbf{z}_1$ and so on. Under diffuse initial conditions it holds that $\text{cov}(\mathbf{z}_1) \rightarrow \infty$, because the dimension of \mathbf{z}_1 coincides with the number of unit roots. Accordingly $[\text{cov}(\mathbf{z}_1)]^{-1} \rightarrow 0$ and then:

$$f(\mathbf{z}|\mathbf{z}_1) = f(\mathbf{z}_1, \mathbf{w}_2, \dots, \mathbf{w}_N | \mathbf{z}_1) = f(\mathbf{w}_2, \dots, \mathbf{w}_N)$$

So the density of the stationary transformation $f(\mathbf{w}_2, \dots, \mathbf{w}_N)$ coincides with the conditional density $f(\mathbf{z}|\mathbf{z}_1)$, being \mathbf{z}_1 is the minimal sub-sample required to determine the diffuse initial conditions in this case.

APPENDIX 2. Proof of the Theorem.

Part (1): Denoting $V \equiv \text{cov}(\mathbf{Z}^{ND}/U)$ and $V^* \equiv \text{cov}(\mathbf{Z}^*/U)$, see expression (3.2), we first want to prove that:

$$|V| |\mathbf{O}_D^T V^{-1} \mathbf{O}_D| = |V^*| |\Pi + \bar{\mathbf{O}}^T V^{*-1} \bar{\mathbf{O}}| |P_1| \quad (\text{A.1})$$

$$\text{with } \bar{\mathbf{O}} = [\mathbf{O}_D \ \mathbf{O}_{ND}] \text{ and } \Pi = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & P_1^{-1} \end{bmatrix}$$

First, we know that:

$$|\Pi + \bar{\mathbf{O}}^T V^{*-1} \bar{\mathbf{O}}| = |\Pi + \bar{\mathbf{O}}^T \mathbf{W}^T \mathbf{W} \bar{\mathbf{O}}| \quad (\text{A.2})$$

where \mathbf{W} is a matrix such that $\mathbf{W}V^*\mathbf{W}^T = \mathbf{I}$, so $V^{*-1} = \mathbf{W}^T \mathbf{W}$. Defining $\hat{\mathbf{O}} \equiv \mathbf{W} \bar{\mathbf{O}}$ expression (A.2) can be reformulated as:

$$\begin{aligned} |\Pi + \bar{\mathbf{O}}^T V^{*-1} \bar{\mathbf{O}}| &= |\Pi + \bar{\mathbf{O}}^T \mathbf{W}^T \mathbf{W} \bar{\mathbf{O}}| = |\Pi + \hat{\mathbf{O}}^T \hat{\mathbf{O}}| = \\ &= \begin{vmatrix} \hat{\mathbf{O}}_D^T \hat{\mathbf{O}}_D & \hat{\mathbf{O}}_D^T \hat{\mathbf{O}}_{ND} \\ \hat{\mathbf{O}}_{ND}^T \hat{\mathbf{O}}_D & \hat{\mathbf{O}}_{ND}^T \hat{\mathbf{O}}_{ND} + P_1^{-1} \end{vmatrix} = |\hat{\mathbf{O}}_D^T \bar{V}^{-1} \hat{\mathbf{O}}_D| |\bar{V}| |P_1^{-1}| \end{aligned} \quad (\text{A.3})$$

where $V = \mathbf{O}_{ND} P_1 \mathbf{O}_{ND}^T + V^*$ [see equation (3.1)] and $\mathbf{W}V\mathbf{W}^T = \mathbf{W} \mathbf{O}_{ND} P_1 \mathbf{O}_{ND}^T \mathbf{W}^T + \mathbf{I}$

Therefore, it is easy to see that:

$$\bar{V} = \hat{\mathbf{O}}_{ND} P_1 \hat{\mathbf{O}}_{ND}^T + \mathbf{I} \quad (\text{A.4})$$

with $\mathbf{W}V\mathbf{W}^T = \bar{V}$. Applying the matrix inversion lemma to expression (A.4):

$$\bar{V}^{-1} = \mathbf{I} - \hat{\mathbf{O}}_{ND} \left[\hat{\mathbf{O}}_{ND}^T \hat{\mathbf{O}}_{ND} + P_1^{-1} \right]^{-1} \hat{\mathbf{O}}_{ND}^T \quad (\text{A.5})$$

Last, (A.3) is:

$$|\hat{\mathbf{O}}_D^T \bar{V}^{-1} \hat{\mathbf{O}}_D| |\bar{V}| |P_1^{-1}| = |\mathbf{O}_D^T V^{-1} \mathbf{O}_D| |\mathbf{W}| |V| |\mathbf{W}^T| |P_1^{-1}| \quad (\text{A.6})$$

and substituting (A.6) in (A.1) we obtain:

$$|V| |\mathbf{O}_D^T V^{-1} \mathbf{O}_D| = |V^*| |\mathbf{\Pi} + \mathbf{O}_D^T V^{-1} \mathbf{O}_D| |V| |V^*|^{-1} |\mathbf{P}_1^{-1}| |\mathbf{P}_1|$$

which is the result that we wanted to prove.

Part (2). Now, we want to prove that:

$$\begin{aligned} V^{-1} - V^{-1} \mathbf{O}_D \left[(\mathbf{O}_D)^T V^{-1} \mathbf{O}_D \right] (\mathbf{O}_D)^T V^{-1} &= \\ &= V^{*-1} - V^{*-1} \bar{\mathbf{O}} \left[\mathbf{\Pi} + \bar{\mathbf{O}}^T V^{*-1} \bar{\mathbf{O}} \right]^{-1} \bar{\mathbf{O}}^T V^{*-1} \end{aligned} \quad (\text{A.7})$$

see expression (3.3). To prove (3.3) it is enough to pre and post-multiply (A.7) by \mathbf{Z}^T and \mathbf{Z} , respectively. Hereafter, we will denote the terms in (A.7) as $AA - BB = CC - DD$ where:

$$AA = V^{-1}$$

$$BB = V^{-1} \mathbf{O}_D \left[(\mathbf{O}_D)^T V^{-1} \mathbf{O}_D \right] (\mathbf{O}_D)^T V^{-1}$$

$$CC = V^{*-1}$$

$$DD = V^{*-1} \bar{\mathbf{O}} \left[\mathbf{\Pi} + \bar{\mathbf{O}}^T V^{*-1} \bar{\mathbf{O}} \right]^{-1} \bar{\mathbf{O}}^T V^{*-1}$$

Under these conditions, we will prove that $DD = BB + CC - AA$

$$\begin{aligned} DD &= \mathbf{W}^T \mathbf{W} \bar{\mathbf{O}} \left[\mathbf{\Pi} + \bar{\mathbf{O}}^T \mathbf{W}^T \mathbf{W} \bar{\mathbf{O}} \right]^{-1} \bar{\mathbf{O}}^T \mathbf{W}^T \mathbf{W} = \\ &= \mathbf{W}^T \hat{\mathbf{O}} \left[\mathbf{\Pi} + \hat{\mathbf{O}}^T \hat{\mathbf{O}} \right]^{-1} \hat{\mathbf{O}}^T \mathbf{W} = \mathbf{W}^T \hat{\mathbf{\Pi}} \mathbf{W} \end{aligned} \quad (\text{A.8})$$

with $\hat{\mathbf{\Pi}} = \hat{\mathbf{O}} \left[\mathbf{\Pi} + \hat{\mathbf{O}}^T \hat{\mathbf{O}} \right]^{-1} \hat{\mathbf{O}}^T$ or,

$$\begin{aligned} \hat{\mathbf{\Pi}} &= \begin{bmatrix} \hat{\mathbf{O}}_D & \hat{\mathbf{O}}_{ND} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{O}}_D^T \hat{\mathbf{O}}_D & \hat{\mathbf{O}}_D^T \hat{\mathbf{O}}_{ND} \\ \hat{\mathbf{O}}_{ND}^T \hat{\mathbf{O}}_D & \hat{\mathbf{O}}_{ND}^T \hat{\mathbf{O}}_{ND} + \mathbf{P}_1^{-1} \end{bmatrix}^{-1} \begin{bmatrix} \hat{\mathbf{O}}_D^T \\ \hat{\mathbf{O}}_{ND}^T \end{bmatrix} = \\ &= \bar{\mathbf{V}}^{-1} \hat{\mathbf{O}}_D \left[\hat{\mathbf{O}}_D^T \bar{\mathbf{V}}^{-1} \hat{\mathbf{O}}_D \right]^{-1} \hat{\mathbf{O}}_D^T \bar{\mathbf{V}}^{-1} + \mathbf{I} - \bar{\mathbf{V}}^{-1} \end{aligned} \quad (\text{A.9})$$

where we applied the partitioned matrix inversion lemma and expression (A.5).

Substituting (A.9) in (A.8) we obtain:

$$\begin{aligned} \mathbf{D}\mathbf{D} &= \mathbf{W}^T \hat{\mathbf{\Pi}} \mathbf{W} = \mathbf{W}^T \left\{ \mathbf{V}^{-1} \hat{\mathbf{O}}_d \left[\hat{\mathbf{O}}_d^T \bar{\mathbf{V}}^{-1} \hat{\mathbf{O}}_d \right]^{-1} \hat{\mathbf{O}}_d^T \bar{\mathbf{V}}^{-1} + \mathbf{I} - \bar{\mathbf{V}}^{-1} \right\} \mathbf{W} = \\ &= \mathbf{W}^T \left\{ \mathbf{V}^{-1} \hat{\mathbf{O}}_d \left[\hat{\mathbf{O}}_d^T \bar{\mathbf{V}}^{-1} \hat{\mathbf{O}}_d \right]^{-1} \hat{\mathbf{O}}_d^T \bar{\mathbf{V}}^{-1} \right\} \mathbf{W} + \mathbf{W}^T \mathbf{W} - \mathbf{W}^T \bar{\mathbf{V}}^{-1} \mathbf{W} \end{aligned}$$

and, taking into account that, $\mathbf{V}^{*-1} = \mathbf{W}^T \mathbf{W}$ and $\mathbf{W}\mathbf{V}\mathbf{W}^T = \bar{\mathbf{V}}$

$$\mathbf{D}\mathbf{D} = \mathbf{V}^{-1} \mathbf{O}_d \left[\mathbf{O}_d^T \mathbf{V}^{-1} \mathbf{O}_d \right]^{-1} \mathbf{O}_d^T \mathbf{V}^{-1} + \mathbf{V}^{*-1} - \mathbf{V}^{-1} = \mathbf{B}\mathbf{B} + \mathbf{C}\mathbf{C} - \mathbf{A}\mathbf{A}$$

which is the result that we wanted to prove.

APPENDIX 3. QR algorithm.

It is well known that the QR decomposition of any $m \times n$ real-valued matrix A is:

$$A = RQ$$

Where Q is an orthonormal $n \times n$ matrix and R is a $m \times n$ lower triangular matrix. The column-pivoting QR decomposition is a reordering such that the elements in the main diagonal of R are sorted in decreasing order according to their absolute value. In this case, the decomposition results in:

$$A = ERQ$$

where E is a $m \times m$ permutation matrix. After the previously defined re-ordering, R can be written as:

$$R = [R_1 \quad 0]$$

Where the dimensions of R_1 are $m \times d$, being d the rank of A , which can be determined as the number of nonzero elements in the main diagonal of R . If we partition Q as:

$$Q = \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix}$$

being Q_1 a $d \times n$ matrix, then A can be written as $A = ER_1Q_1$, where Q_1 spans the row-space of A and Q_2 is orthogonal to this row-space. In the Column Deletion algorithm we apply this decomposition directly to $H_t^* = HT_t$, obtaining $Q_{H_t^*} = (Q_1)^T$, $Q_{H_t^*}^\perp = (Q_2)^T$, $H_t^{**} = ER_1$ and the term of the likelihood given in (4.26) can be computed as $\log |H_{t0}^{*T} H_{t0}^*| = \sum_{i=1}^d \log [R_1(i, i)]$

The State Decomposition methods requires including the decomposition of HT in the filter, where T is given by (2.4). The matrix $M^{-1} = [T \ G]$ can be easily computed by the initialization procedure proposed by De Jong and Chu-Chun-Lin (1994).

Table 1. Estimation results obtained for the full airline dataset.

	Model (5.1), diffuse and conditional likelihood	Model (5.2), diffuse likelihood	Model (5.2), conditional likelihood
Minus log- likelihood	-244.697	-323.750	-244.697
$\hat{\theta}$.402		
$\hat{\Theta}$.557		
$\hat{\sigma}_\epsilon$.037		

Table 2. Estimation results obtained for the airline dataset, treating observations # 62 and 135 as missing values.

	Model (5.1), diffuse and conditional likelihood	Model (5.2), diffuse likelihood	Model (5.2), conditional likelihood
Minus log-likelihood	-237.780	-236.903	-250.687
$\hat{\theta}$.325	.359	
$\hat{\Theta}$.625	.568	
$\hat{\sigma}_\varepsilon$.034	.034	

Table 3. Results obtained with the State Decomposition algorithm (identical to those of the Column Deletion method). The true parameter values are given in the first column and have been employed in all the cases as initial conditions for the iterative optimization. Constrained parameters are denoted by an asterisk.

	Normalizing constraint		
	$\alpha = 1$	$\beta = 1$	$\sigma_a = .1$
Likelihood value on convergence	130.887		
$\alpha = 1$	1.000*	1.011	.913
$\beta = 1$.989	1.000*	.903
$\sigma_1 = .4$.404	.404	.404
$\sigma_2 = .4$.389	.389	.389
$\phi_f = .7$.776	.776	.776
$\sigma_\varepsilon = .1$.091	.090	.100*

Table 4. Results obtained with the diffuse likelihood. The true parameter values are given in the first column and have been employed in all the cases as initial conditions for the iterative optimization. Constrained parameters are denoted by an asterisk.

	Normalizing constraint		
	$\alpha = 1$	$\beta = 1$	$\sigma_a = .1$
Likelihood value on convergence	130.887	130.897	130.746
$\alpha = 1$	1.000*	1.010	.827
$\beta = 1$.989	1.000*	.818
$\sigma_1 = .4$.404	.404	.405
$\sigma_2 = .4$.389	.389	.390
$\phi_f = .7$.776	.776	.806
$\sigma_\varepsilon = .1$.091	.090	.100*