The University of Sydney Business School
The University of Sydney

BUSINESS ANALYTICS WORKING PAPER SERIES

Maximum likelihood estimation of time series models: the Kalman filter and beyond

Tommaso Proietti
Discipline of Business Analytics
The University of Sydney

Alessandra Luati
Department of Statistics
University of Bologna, Italy

Abstract

The purpose of this chapter is to provide a comprehensive treatment of likelihood inference for state space models. These are a class of time series models relating an observable time series to quantities called states, which are characterized by a simple temporal dependence structure, typically a first order Markov process.

The states have sometimes substantial interpretation. Key estimation problems in economics concern latent variables, such as the output gap, potential output, the non-accelerating-inflation rate of unemployment, or NAIRU, core inflation, and so forth. Time-varying volatility, which is quintessential to finance, is an important feature also in macroeconomics. In the multivariate framework relevant features can be common to different series, meaning that the driving forces of a particular feature and/or the transmission mechanism are the same.

The objective of this chapter is reviewing this algorithm and discussing maximum likelihood inference, starting from the linear Gaussian case and discussing the extensions to a nonlinear and non Gaussian framework.

May 2012

BA Working Paper No: 02/2012
http://sydney.edu.au/business/business_analytics/research/working_papers

Maximum likelihood estimation of time series models: the Kalman filter and beyond*

Tommaso Proietti
Discipline of Business Analytics
University of Sydney Business School
Sydney, NSW Australia

Alessandra Luati
Department of Statistics
University of Bologna
Italy

1 Introduction

The purpose of this chapter is to provide a comprehensive treatment of likelihood inference for state space models. These are a class of time series models relating an observable time series to quantities called states, which are characterized by a simple temporal dependence structure, typically a first order Markov process.

The states have sometimes substantial interpretation. Key estimation problems in economics concern latent variables, such as the output gap, potential output, the non-accelerating-inflation rate of unemployment, or NAIRU, core inflation, and so forth. Time-varying volatility, which is quintessential to finance, is an important feature also in macroeconomics. In the multivariate framework relevant features can be common to different series, meaning that the driving forces of a particular feature and/or the transmission mechanism are the same.

The main macroeconomic applications of state space models have dealt with the following topics.

- The extraction of signals such as trends and cycles in macroeconomic time series: see Watson (1986), Clark (1987), Harvey and Jaeger (1993), Hodrick and Prescott (1997), Morley, Nelson and Zivot (2003), Proietti (2006), Luati and Proietti (2011).
- Dynamic factor models, for the extraction of a single index of coincident indicators, see Stock and Watson (1989), Frale *et al.* (2011), and for large dimensional systems (Jungbacker, Koopman and van der Wel, 2011).
- Stochastic volatility models: see Shephard (2005) and Stock and Watson (2007) for applications to US inflation.
- Time varying autoregressions, with stochastic volatility: Primiceri (2005), Cogley, Primiceri and Sargent (2010).
- Structural change in macroeconomics: see Kim and Nelson (1999), Giordani, Kohn and van Dijk (2007).

*Chapter written for the Handbook of Research Methods and Applications on Empirical Macroeconomics, edited by Nigar Hashimzade and Michael Thornton, forthcoming in 2012 (Edward Elgar Publishing).

- The class of dynamic stochastic general equilibrium (DSGE) models: Sargent (1989), Fernandez-Villaverde and Rubio-Ramirez (2005), Smets and Wouters (2003), Fernandez-Villaverde (2010).

Leading macroeconomics books, such as Ljungqvist and Sargent (2004) and Canova (2007), provide a comprehensive treatment of state space models and related methods. The above list of references and topics is all but exhaustive and the literature has been growing at a fast rate.

State space methods are tools for inference in state space models, since they allow one to estimate any unknown parameters along with the states, to assess the uncertainty of the estimates, to perform diagnostic checking, to forecast future states and observations, and so forth.

The Kalman filter (Kalman, 1960; Kalman and Bucy, 1961) is a fundamental algorithm for the statistical treatment of a state space model. Under the Gaussian assumption it produces the minimum mean square estimator of the state vector along with its mean square error matrix, conditional on past information; this is used to build the one-step-ahead predictor of y_t and its mean square error matrix. Due to the independence of the one-step-ahead prediction errors, the likelihood can be evaluated via the prediction error decomposition.

The objective of this chapter is reviewing this algorithm and discussing maximum likelihood inference, starting from the linear Gaussian case and discussing the extensions to a nonlinear and non Gaussian framework. Due to space constraints we shall provide a self-contained treatment of the standard case and an overview of the possible modes of inference in the nonlinear and non Gaussian case. For more details we refer the reader to Jazwinski (1970), Anderson and Moore (1979), Hannan and Deistler (1988), Harvey (1989), West and Harrison (1997), Kitagawa and Gersch (1996) Kailath, Sayed and Hassibi (2000), Durbin and Koopman (2001), Harvey and Proietti (2005), Cappé, Moulines and Ryden (2007) and Kitagawa (2009).

The chapter is structured as follows. Section 2 introduces state space models and provides the state space representation of some commonly applied linear processes, such as univariate and multivariate autoregressive moving average processes (ARMA) and dynamic factor models. Section 3 is concerned with the basic tool for inference in state space models, that is the Kalman filter. Maximum likelihood estimation is the topic of section 4, and discusses the profile and marginal likelihood, when nonstationary and regression effects are present; section 5 deals with estimation by the Expectation Maximization (EM) algorithm. Section 6 considers inference in nonlinear and non-Gaussian models along with stochastic simulation methods and new directions of research. Section 7 concludes the chapter.

2 State space models

We begin our treatment with the linear Gaussian state space model. Let $\mathbf{y}_t$ denote an $N \times 1$ vector time series related to an $m \times 1$ vector of unobservable components, the states, $\boldsymbol{\alpha}_t$, by the so-called measurement equation,

$$\mathbf{y}_t = \mathbf{Z}_t \boldsymbol{\alpha}_t + \mathbf{G}_t \boldsymbol{\varepsilon}_t, \qquad t = 1, \ldots, n, \tag{1}$$

where $\mathbf{Z}_t$ is an $N \times m$ matrix, $\mathbf{G}_t$ is $N \times g$ and $\boldsymbol{\varepsilon}_t \sim \text{NID}(0, \sigma^2 \mathbf{I}_g)$.

The evolution of the states is governed by the transition equation:

$$\boldsymbol{\alpha}_{t+1} = \mathbf{T}_t \boldsymbol{\alpha}_t + \mathbf{H}_t \boldsymbol{\varepsilon}_t, \qquad t = 1, 2, \ldots, n, \tag{2}$$

where the transition matrix $\mathbf{T}_t$ is $m \times m$ and $\mathbf{H}_t$ is $m \times g$.

The specification of the state space model is completed by the initial conditions concerning the distribution of $\boldsymbol{\alpha}_1$. In the sequel we shall assume that this distribution is independent of $\boldsymbol{\varepsilon}_t, \forall t \geq 1$. When the system is time-invariant and $\boldsymbol{\alpha}_t$ is stationary (the eigenvalues of the transition matrix, $\mathbf{T}$, are inside the unit circle), the initial conditions are provided by the unconditional mean and covariance matrix of the state vector, $\mathrm{E}(\boldsymbol{\alpha}_1) = \mathbf{0}$ and $\mathrm{Var}(\boldsymbol{\alpha}_1) = \sigma^2 \mathbf{P}_{1|0}$, satisfying the matrix equation $\mathbf{P}_{1|0} = \mathbf{T}\mathbf{P}_{1|0}\mathbf{T}' + \mathbf{H}\mathbf{H}'$. Initialization of the system turns out to be a relevant issue when nonstationary components are present.

Often the models are specified in a way that the measurement and transition equation disturbances are uncorrelated, i.e. $\mathbf{H}_t \mathbf{G}_t' = 0, \forall t$.

The system matrices, $\mathbf{Z}_t, \mathbf{G}_t, \mathbf{T}_t$, and $\mathbf{H}_t$, are non-stochastic, i.e. they are allowed to vary over time in a deterministic fashion, and are functionally related to a set of hyperparameters, $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^k$, which are usually unknown. If the system matrices are constant, i.e. $\mathbf{Z}_t = \mathbf{Z}, \mathbf{G}_t = \mathbf{G}, \mathbf{T}_t = \mathbf{T}$ and $\mathbf{H}_t = \mathbf{H}$, the state space model is time invariant.

2.1 State Space representation of ARMA models

Let y_t be a scalar time series with ARMA(p, q) representation:

$$y_t = \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \xi_t + \theta_1 \xi_{t-1} + \cdots + \theta_q \xi_{t-q}, \xi_t \sim \mathrm{NID}(0, \sigma^2),$$

or $\phi(L) y_t = \theta(L)\xi_t$, where L is the lag operator, and $\phi(L) = 1 - \phi_1 L - \cdots - \phi_p L^p$, $\theta(L) = 1 + \theta_1 L + \cdots + \theta_q L^q$,

The state space representation proposed by Pearlman (1980), see Burridge and Wallis (1988) and de Jong and Penzer (2004), is based on $m = \max(p, q)$ state elements and it is such that $\boldsymbol{\varepsilon}_t = \xi_t$. The time invariant system matrices are

$$\mathbf{Z} = [1, \quad \mathbf{0}'_{m-1}], \mathbf{G} = 1, \mathbf{T} = \begin{bmatrix} \phi_1 & 1 & 0 & \cdots & 0 \\ \phi_2 & 0 & 1 & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & 0 \\ \vdots & \cdots & \cdots & 0 & 1 \\ \phi_m & 0 & \cdots & \cdots & 0 \end{bmatrix}, \mathbf{H} = \begin{bmatrix} \theta_1 + \phi_1 \\ \theta_2 + \phi_2 \\ \vdots \\ \vdots \\ \theta_m + \phi_m \end{bmatrix}.$$

If y_t is stationary, the eigenvalues of $\mathbf{T}$ are inside the unit circle (and viceversa). State space representations are not unique. The representation adopted by Harvey (1989) is based on $m = \max(p, q+1)$ states and has $\mathbf{Z}, \mathbf{T}$ as above, but $\mathbf{G} = 0$ and $\mathbf{H}' = [1, \theta_1, \ldots, \theta_m]$. The canonical observable representation in Brockwell and Davis (1991) has minimal state dimension, $m = \max(p, q)$, and

$$\mathbf{Z} = [1, \quad \mathbf{0}'_{m-1}], \mathbf{G} = 1, \mathbf{T} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & 0 \\ \vdots & \cdots & \cdots & 0 & 1 \\ \phi_m & \phi_{m-1} & \cdots & \cdots & \phi_1 \end{bmatrix}, \mathbf{H} = \begin{bmatrix} \psi_1 \\ \psi_2 \\ \vdots \\ \vdots \\ \psi_m \end{bmatrix},$$

where ψ_j are the coefficients of the Wold polynomial $\psi(L) = \theta(L)/\phi(L)$. The virtues of this representation is that $\boldsymbol{\alpha}_t = [\tilde{y}_{t|t-1}, \tilde{y}_{t+1|t-1}, \ldots, \tilde{y}_{t+m-1|t-1}]'$ where $\tilde{y}_{t+j|t-1} = \mathrm{E}(y_{t+j}|Y_{t-1}), Y_t = \{y_t, y_{t-1}, \ldots\}$.

In fact, the transition equation is based on the forecast updating recursions:

$$\tilde{y}_{t+j|t} = \tilde{y}_{t+j-1|t-1} + \psi_j \xi_t, j = 1, \ldots, m-1, \tilde{y}_{t+m|t} = \sum_{k=1}^{m} \phi_k \tilde{y}_{t+k|t-1} + \psi_m \xi_t.$$

2.2 AR and MA approximation of Fractional Noise

The fractional noise process $(1-L)^d y_t = \xi_t, \xi_t \sim \text{NID}(0, \sigma^2)$, is stationary if $d < 0.5$. Unfortunately such a process is not finite order Markovian and does not admit a state space representation with finite m. Chan and Palma (1998) obtained the finite m AR and MA approximations by truncating respectively the AR polynomial $\phi(L) = (1-L)^d = 1 - \sum_{j=1}^{\infty} \frac{\Gamma(j+d)}{\Gamma(d)\Gamma(j+1)} L^j$ and the MA polynomial $\theta(L) = (1-L)^{-d} = 1 + \sum_{j=1}^{\infty} \frac{\Gamma(j-d)}{\Gamma(-d)\Gamma(j+1)} L^j$. Here $\Gamma(\cdot)$ is the Gamma function. A better option is to obtain the first m AR coefficients applying the Durbin-Levison algorithm to the Toeplitz variance covariance matrix of the process.

2.3 AR(1) plus noise model

Consider an AR(1) process μ_t observed with error:

$$\begin{array}{rcll} y_t & = & \mu_t + \epsilon_t & \epsilon_t \sim \text{NID}(0, \sigma_\epsilon^2), \\ \mu_{t+1} & = & \phi \mu_t + \eta_t, & \eta_t \sim \text{NID}(0, \sigma_\eta^2) \end{array}$$

where $|\phi| < 1$ to ensure stationarity and $\text{E}(\eta_t \epsilon_{t+s}) = 0, \forall s$. The initial condition is $\mu_1 \sim \text{N}(\tilde{\mu}_{1|0}, P_{1|0})$.

Assuming that the process has started in the indefinitely remote past $\tilde{\mu}_{1|0} = 0, P_{1|0} = \frac{\sigma_\eta^2}{1-\phi^2}$. Alternatively, we may assume that the process started at time 1, so that $P_{1|0} = 0$ and μ_1 is a fixed (though possibly unknown) value.

If $\sigma_\epsilon^2 = 0$ then $y_t \sim$ AR(1); on the other hand, if $\sigma_\eta^2 = 0$ then $y_t \sim \text{NID}(0, \sigma_\epsilon^2)$; finally, if $\phi = 0$ then the model is not identifiable.

When $\phi = 1$, the local level (random walk plus noise) model is obtained.

2.4 Time-varying AR models

Consider the time varying VAR model $\mathbf{y}_t = \sum_{k=1}^{p} \mathbf{\Phi}_{kt} \mathbf{y}_{t-k} + \boldsymbol{\xi}_t, \boldsymbol{\xi}_t = \text{N}(\mathbf{0}, \mathbf{\Sigma}_t)$ with random walk evolution for the coefficients:

$$\text{vec}(\mathbf{\Phi}_{k,t+1}) = \text{vec}(\mathbf{\Phi}_{k,t}) + \boldsymbol{\eta}_{kt}, \boldsymbol{\eta}_{kt} \sim \text{NID}(\mathbf{0}, \mathbf{\Sigma}_\eta);$$

(see Primiceri, 2005). Often $\mathbf{\Sigma}_\eta$ is taken as a scalar or a diagonal matrix.

The model can be cast in state space form, setting $\boldsymbol{\alpha}_t = [\text{vec}(\mathbf{\Phi}_1)', \ldots, \text{vec}(\mathbf{\Phi}_p)']'$, $\mathbf{Z}_t = [(\mathbf{y}_{t-1}' \otimes \mathbf{I}), \ldots, (\mathbf{y}_{t-p}' \otimes \mathbf{I})], \mathbf{G} = \mathbf{\Sigma}^{1/2}, \mathbf{T}_t = \mathbf{I}, \mathbf{H} = \mathbf{\Sigma}_\eta^{1/2}$.

Time-varying volatility is incorporated by writing $\mathbf{G}_t = \mathbf{C}_t \mathbf{D}_t$ where $\mathbf{C}_t$ is lower diagonal with unit diagonal elements and $c_{ij,t+1} = c_{ij,t} + \zeta_{ij,t}, j < i, \zeta_{ij,t} \sim \text{NID}(0, \sigma_\zeta^2)$, and $\mathbf{D}_t = \text{diag}(d_{it}, i = 1, \ldots, N$, $\ln d_{i,t+1} = \ln d_{it} + \kappa_{it}$, $\kappa_{it} \sim \text{NID}(0, \sigma_\kappa^2)$. Allowing for time-varying volatility makes the model non linear.

2.5 Dynamic factor models

A simple model is $\mathbf{y}_t = \mathbf{\Lambda}\mathbf{f}_t + \mathbf{u}_t$ where $\mathbf{\Lambda}$ is the matrix of factor loadings, $\mathbf{f}_t$ are q common factors admitting a VAR representation $\mathbf{f}_{t+1} = \mathbf{\Phi}\mathbf{f}_t + \boldsymbol{\eta}_t, \boldsymbol{\eta}_t \sim \mathrm{N}(\mathbf{0}, \mathbf{\Sigma}_\eta)$, see Sargent and Sims (1977), Stock and Watson (1989). For identification we need to impose $q(q+1)/2$ restrictions (see Geweke and Singleton, 1981). One possibility is to set $\mathbf{\Sigma}_\eta = \mathbf{I}$; alternatively, we could set $\mathbf{\Lambda}$ equal to a lower triangular matrix with ones on the main diagonal.

2.6 Contemporaneous and future representations

The transition equation (2) has been specified in the so-called future form; in some treatment, e.g. Harvey (1989) and West and Harrison (1996), the contemporaneous form of the model is adopted, with (2) replaced by $\boldsymbol{\alpha}_t^* = \mathbf{T}_t\boldsymbol{\alpha}_{t-1}^* + \mathbf{H}_t\boldsymbol{\varepsilon}_t$, $t = 1, \ldots, n$, whereas the measurement equation retains the form $\mathbf{y}_t = \mathbf{Z}^*\boldsymbol{\alpha}_t^* + \mathbf{G}^*\boldsymbol{\varepsilon}_t$. The initial conditions are usually specified in terms of $\boldsymbol{\alpha}_0^* \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{P}_0)$, which is assumed to be distributed independently of $\boldsymbol{\varepsilon}_t, \forall t \geq 1$.

Simple algebra shows that we can reformulate the model in future form (1)-(2) with $\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_{t-1}^*, \mathbf{Z} = \mathbf{Z}^*\mathbf{T}^*, \mathbf{G} = \mathbf{G}^* + \mathbf{Z}^*\mathbf{H}^*$.

For instance, consider the AR(1) plus noise model in contemporaneous form, specified as $y_t = \mu_t^* + \epsilon_t^*$, $\mu_t^* = \phi\mu_{t-1}^* + \eta_t^*$, with ϵ_t^* and η_t^* mutually and serially independent. Substituting from the transition equation, $y_t = \mu_{t-1}^* + \eta_t^* + \epsilon_t^*$, and setting $\mu_t = \mu_{t-1}^*$, we can rewrite the model in future form, but the disturbances $\epsilon_t = \eta_t^* + \epsilon_t^*$ and $\eta_t = \eta_t^*$ will be (positively) correlated.

2.7 Fixed effects and explanatory variables

The linear state space model can be extended to introduce fixed and regression effects. There are essentially two ways for handling them.

If we let $\mathbf{X}_t$ and $\mathbf{W}_t$ denote fixed and known matrices of dimension $N \times k$ and $m \times k$, respectively, the state space form can be generalised as follows:

$$\mathbf{y}_t = \mathbf{Z}_t\boldsymbol{\alpha}_t + \mathbf{X}_t\boldsymbol{\beta} + \mathbf{G}_t\boldsymbol{\varepsilon}_t, \qquad \boldsymbol{\alpha}_{t+1} = \mathbf{T}_t\boldsymbol{\alpha}_t + \mathbf{W}_t\boldsymbol{\beta} + \mathbf{H}_t\boldsymbol{\varepsilon}_t. \tag{3}$$

In the sequel we shall express the initial state vector in terms of the vector $\boldsymbol{\beta}$ as follows:

$$\boldsymbol{\alpha}_1 = \tilde{\boldsymbol{\alpha}}_{1|0}^* + \mathbf{W}_0\boldsymbol{\beta} + \mathbf{H}_0\boldsymbol{\varepsilon}_0, \boldsymbol{\varepsilon}_0 \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I}), \tag{4}$$

where $\tilde{\boldsymbol{\alpha}}_{1|0}^*$, $\mathbf{W}_0$, $\mathbf{H}_0$, are known quantities.

Alternatively, $\boldsymbol{\beta}$ is included in the state vector and the state space model becomes:

$$\mathbf{y}_t = \mathbf{Z}_t^\dagger\boldsymbol{\alpha}_t^\dagger + \mathbf{G}_t\boldsymbol{\varepsilon}_t, \qquad \boldsymbol{\alpha}_{t+1}^\dagger = \mathbf{T}_t^\dagger\boldsymbol{\alpha}_t^\dagger + \mathbf{H}_t^\dagger\boldsymbol{\varepsilon}_t,$$

where

$$\boldsymbol{\alpha}_t^\dagger = \begin{bmatrix} \boldsymbol{\alpha}_t \\ \boldsymbol{\beta}_t \end{bmatrix}, \quad \mathbf{Z}_t^\dagger = [\mathbf{Z}_t \;\; \mathbf{X}_t], \quad \mathbf{T}_t^\dagger = \begin{bmatrix} \mathbf{T}_t & \mathbf{W}_t \\ \mathbf{0} & \mathbf{I}_k \end{bmatrix}, \mathbf{H}_t^\dagger = \begin{bmatrix} \mathbf{H}_t \\ \mathbf{0} \end{bmatrix}.$$

This representation opens the way to the treatment of $\boldsymbol{\beta}$ as a time varying vector.

3 The Kalman filter

Consider a stationary state space model with no fixed effect (1)-(2) with initial condition $\boldsymbol{\alpha}_1 \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{P}_{1|0})$, independent of $\boldsymbol{\varepsilon}_t, t \geq 1$, and define $\mathbf{Y}_t = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_t\}$, the information set up to and including time t, $\tilde{\boldsymbol{\alpha}}_{t|t-1} = \mathrm{E}(\boldsymbol{\alpha}_t|\mathbf{Y}_{t-1})$, and $\mathrm{Var}(\boldsymbol{\alpha}_t|\mathbf{Y}_{t-1}) = \sigma^2\mathbf{P}_{t|t-1}$.

The Kalman filter (KF) is the following recursive algorithm: for $t = 1, \ldots, n$,

$$\begin{array}{ll} \boldsymbol{\nu}_t = \mathbf{y}_t - \mathbf{Z}_t\tilde{\boldsymbol{\alpha}}_{t|t-1}, & \mathbf{F}_t = \mathbf{Z}_t\mathbf{P}_{t|t-1}\mathbf{Z}_t' + \mathbf{G}_t\mathbf{G}_t', \\ & \mathbf{K}_t = (\mathbf{T}_t\mathbf{P}_{t|t-1}\mathbf{Z}_t' + \mathbf{H}_t\mathbf{G}_t')\mathbf{F}_t^{-1}, \\ \tilde{\boldsymbol{\alpha}}_{t+1|t} = \mathbf{T}_t\tilde{\boldsymbol{\alpha}}_{t|t-1} + \mathbf{K}_t\boldsymbol{\nu}_t, & \mathbf{P}_{t+1|t} = \mathbf{T}_t\mathbf{P}_{t|t-1}\mathbf{T}_t' + \mathbf{H}_t\mathbf{H}_t' - \mathbf{K}_t\mathbf{F}_t\mathbf{K}_t'. \end{array}$$

Hence, the KF computes recursively the optimal predictor of the states and thereby of $\mathbf{y}_t$ conditional on past information as well as the variance of their prediction error. The vector $\boldsymbol{\nu}_t = \mathbf{y}_t - \mathrm{E}(\mathbf{y}_t|\mathbf{Y}_{t-1})$ is the time t innovation. i.e. the new information in $\mathbf{y}_t$ that could not be predicted from knowledge of the past, also known as the one-step-ahead prediction error; $\sigma^2\mathbf{F}_t$ is the prediction error variance at time t, that is $\mathrm{Var}(\mathbf{y}_t|\mathbf{Y}_{t-1})$. The one-step-ahead predictive distribution is $\mathbf{y}_t|\mathbf{Y}_{t-1} \sim \mathrm{N}(\mathbf{Z}_t\tilde{\boldsymbol{\alpha}}_{t|t-1}, \sigma^2\mathbf{F}_t)$. The matrix $\mathbf{K}_t$ is sometimes referred to as the Kalman gain.

3.1 Proof of the Kalman Filter

Let us assume that $\tilde{\boldsymbol{\alpha}}_{t|t-1}$, $\mathbf{P}_{t|t-1}$ are given at the t-th run of the KF. The available information set is $\mathbf{Y}_{t-1}$. Taking the conditional expectation of both sides of the measurement equations yields $\tilde{\mathbf{y}}_{t|t-1} = \mathrm{E}(\mathbf{y}_t|\mathbf{Y}_{t-1}) = \mathbf{Z}_t\tilde{\boldsymbol{\alpha}}_{t|t-1}$. The innovation at time t is $\boldsymbol{\nu}_t = \mathbf{y}_t - \mathbf{Z}_t\tilde{\boldsymbol{\alpha}}_{t|t-1} = \mathbf{Z}_t(\boldsymbol{\alpha}_t - \tilde{\boldsymbol{\alpha}}_{t|t-1}) + \mathbf{G}_t\varepsilon_t$. Moreover, $\mathrm{Var}(\mathbf{y}_t|\mathbf{Y}_{t-1}) = \sigma^2\mathbf{F}_t$, where $\mathbf{F}_t = \mathbf{Z}_t\mathbf{P}_{t|t-1}\mathbf{Z}_t' + \mathbf{G}_t\mathbf{G}_t'$. From the transition equation, $\mathrm{E}(\boldsymbol{\alpha}_{t+1}|\mathbf{Y}_{t-1}) = \mathbf{T}_t\tilde{\boldsymbol{\alpha}}_{t|t-1}$ $\mathrm{Var}(\boldsymbol{\alpha}_{t+1}|\mathbf{Y}_{t-1}) = \mathrm{Var}\left[\mathbf{T}_t(\boldsymbol{\alpha}_t - \tilde{\boldsymbol{\alpha}}_{t|t-1}) + \mathbf{H}_t\boldsymbol{\varepsilon}_t\right] = \sigma^2(\mathbf{T}_t\mathbf{P}_{t|t-1}\mathbf{T}_t' + \mathbf{H}_t\mathbf{H}_t')$, and $\mathrm{Cov}(\boldsymbol{\alpha}_{t+1}, \mathbf{y}_t|\mathbf{Y}_{t-1}) = \sigma^2(\mathbf{T}_t\mathbf{P}_{t|t-1}\mathbf{Z}_t' + \mathbf{H}_t\mathbf{G}_t')$.

The joint conditional distribution of $(\boldsymbol{\alpha}_{t+1}, \mathbf{y}_t)$ is thus:

$$\begin{pmatrix} \boldsymbol{\alpha}_{t+1} \\ \mathbf{y}_t \end{pmatrix} \Bigg| \mathbf{Y}_{t-1} \sim \mathrm{N}\left[\begin{pmatrix} \mathbf{T}_t\tilde{\boldsymbol{\alpha}}_{t|t-1} \\ \mathbf{Z}_t\tilde{\boldsymbol{\alpha}}_{t|t-1} \end{pmatrix}, \sigma^2 \begin{pmatrix} \mathbf{T}_t\mathbf{P}_{t|t-1}\mathbf{T}_t' + \mathbf{H}_t\mathbf{H}_t', & \mathbf{T}_t\mathbf{P}_{t|t-1}\mathbf{Z}_t' + \mathbf{H}_t\mathbf{G}_t' \\ \mathbf{Z}_t\mathbf{P}_{t|t-1}\mathbf{T}_t' + \mathbf{G}_t\mathbf{H}_t', & \mathbf{F}_t \end{pmatrix} \right]$$

which implies $\boldsymbol{\alpha}_{t+1}|\mathbf{Y}_{t-1}, \mathbf{y}_t \equiv \boldsymbol{\alpha}_{t+1}|\mathbf{Y}_t \sim \mathrm{N}(\tilde{\boldsymbol{\alpha}}_{t+1|t}, \sigma^2\mathbf{P}_{t+1|t})$ with $\tilde{\boldsymbol{\alpha}}_{t+1|t} = \mathbf{T}_t\tilde{\boldsymbol{\alpha}}_{t|t-1} + \mathbf{K}_t\boldsymbol{\nu}_t$, $\mathbf{K}_t = (\mathbf{T}_t\mathbf{P}_{t|t-1}\mathbf{Z}_t' + \mathbf{H}_t\mathbf{G}_t')\mathbf{F}_t^{-1}$, $\mathbf{P}_{t+1|t} = \mathbf{T}_t\mathbf{P}_{t|t-1}\mathbf{T}_t' + \mathbf{H}_t\mathbf{H}_t' - \mathbf{K}_t\mathbf{F}_t\mathbf{K}_t'$. Hence, $\mathbf{K}_t = \mathrm{Cov}(\boldsymbol{\alpha}_t, \mathbf{y}_t|\mathbf{Y}_{t-1})$ $[\mathrm{Var}(\mathbf{y}_t|\mathbf{Y}_{t-1})]^{-1}$ is the regression matrix of $\boldsymbol{\alpha}_t$ on the new information $\mathbf{y}_t$, given $\mathbf{Y}_{t-1}$.

3.2 Real time estimates and an alternative Kalman filter

The updated (real time) estimates of the state vector, $\tilde{\boldsymbol{\alpha}}_{t|t} = \mathrm{E}(\boldsymbol{\alpha}_t|\mathbf{Y}_t)$, and their covariance matrix $\mathrm{Var}(\boldsymbol{\alpha}_t|\mathbf{Y}_t) = \sigma^2\mathbf{P}_{t|t}$ are:

$$\tilde{\boldsymbol{\alpha}}_{t|t} = \tilde{\boldsymbol{\alpha}}_{t|t-1} + \mathbf{P}_{t|t-1}\mathbf{Z}_t'\mathbf{F}_t^{-1}\boldsymbol{\nu}_t, \qquad \mathbf{P}_{t|t} = \mathbf{P}_{t|t-1} - \mathbf{P}_{t|t-1}\mathbf{Z}_t'\mathbf{F}_t^{-1}\mathbf{Z}_t\mathbf{P}_{t|t-1}. \tag{5}$$

The proof of (5) is straightforward. We start writing the joint distribution of the states and the last observation, given the past:

$$\begin{pmatrix} \boldsymbol{\alpha}_t \\ \mathbf{y}_t \end{pmatrix} \Bigg| \mathbf{Y}_{t-1} \sim \mathrm{N}\left[\begin{pmatrix} \tilde{\boldsymbol{\alpha}}_{t|t-1} \\ \mathbf{Z}_t\tilde{\boldsymbol{\alpha}}_{t|t-1} \end{pmatrix}, \sigma^2 \begin{pmatrix} \mathbf{P}_{t|t-1}, & \mathbf{P}_{t|t-1}\mathbf{Z}_t' \\ \mathbf{Z}_t\mathbf{P}_{t|t-1}, & \mathbf{F}_t \end{pmatrix} \right]$$

whence it follows $\boldsymbol{\alpha}_t|\mathbf{Y}_{t-1}, \mathbf{y}_t \equiv \boldsymbol{\alpha}_t|\mathbf{Y}_t \sim \mathrm{N}(\tilde{\boldsymbol{\alpha}}_{t|t}, \sigma^2\mathbf{P}_{t|t})$ with (5) providing, respectively,

$$\mathrm{E}(\boldsymbol{\alpha}_t|\mathbf{Y}_t) = \mathrm{E}(\boldsymbol{\alpha}_t|\mathbf{Y}_{t-1}) + \mathrm{Cov}(\boldsymbol{\alpha}_t, \mathbf{y}_t|\mathbf{Y}_{t-1}) \left[\mathrm{Var}(\mathbf{y}_t|\mathbf{Y}_{t-1})\right]^{-1} \left[\mathbf{y}_t - \mathrm{E}(\mathbf{y}_t|\mathbf{Y}_{t-1})\right]$$

$$\mathrm{Var}(\boldsymbol{\alpha}_t|\mathbf{Y}_t) = \mathrm{Var}(\boldsymbol{\alpha}_t|\mathbf{Y}_{t-1}) - \mathrm{Cov}(\boldsymbol{\alpha}_t, \mathbf{y}_t|\mathbf{Y}_{t-1}) \left[\mathrm{Var}(\mathbf{y}_t|\mathbf{Y}_{t-1})\right]^{-1} \mathrm{Cov}(\mathbf{y}_t, \boldsymbol{\alpha}_t|\mathbf{Y}_{t-1}).$$

The KF recursions for the states can be broken up into an updating step, followed by a prediction step: for $t = 1, \ldots, n$,

$$\begin{array}{ll} \boldsymbol{\nu}_t = \mathbf{y}_t - \mathbf{Z}_t\tilde{\boldsymbol{\alpha}}_{t|t-1}, & \mathbf{F}_t = \mathbf{Z}_t\mathbf{P}_{t|t-1}\mathbf{Z}_t' + \mathbf{G}_t\mathbf{G}_t', \\ \tilde{\boldsymbol{\alpha}}_{t|t} = \tilde{\boldsymbol{\alpha}}_{t|t-1} + \mathbf{P}_{t|t-1}\mathbf{Z}_t'\mathbf{F}_t^{-1}\boldsymbol{\nu}_t, & \mathbf{P}_{t|t} = \mathbf{P}_{t|t-1} - \mathbf{P}_{t|t-1}\mathbf{Z}_t'\mathbf{F}_t^{-1}\mathbf{Z}_t\mathbf{P}_{t|t-1}. \\ \tilde{\boldsymbol{\alpha}}_{t+1|t} = \mathbf{T}_t\tilde{\boldsymbol{\alpha}}_{t|t} + \mathbf{H}_t\mathbf{G}_t'\mathbf{F}_t^{-1}\boldsymbol{\nu}_t, & \mathbf{P}_{t+1|t} = \mathbf{T}_t\mathbf{P}_{t|t}\mathbf{T}_t' + \mathbf{H}_t\mathbf{H}_t' - \mathbf{H}_t\mathbf{G}_t'\mathbf{F}_t^{-1}\mathbf{G}_t\mathbf{H}_t'. \end{array}$$

The last row follows from $\boldsymbol{\varepsilon}_t|\mathbf{Y}_t \sim \mathrm{N}\left(\mathbf{G}_t'\mathbf{F}_t^{-1}\boldsymbol{\nu}_t, \sigma^2(I - \mathbf{G}_t'\mathbf{F}_t^{-1}\mathbf{G}_t)\right)$. Also, when $\mathbf{H}_t\mathbf{G}_t' = 0$ (uncorrelated measurement and transition disturbances), the prediction equations in (3.2) simplify considerably.

3.3 Illustration: the AR(1) plus noise model

For the AR(1) plus noise process considered above, let $\sigma^2 = 1$ and $\mu_1 \sim N(\tilde{\mu}_{1|0}, P_{1|0}), \tilde{\mu}_{1|0} = 0, P_{1|0} = \sigma_\eta/(1-\phi^2)$. Hence, $\tilde{y}_{1|0} = \mathrm{E}(y_1|Y_0) = \tilde{\mu}_{1|0} = 0$, so that at the first update of the KF,

$$\nu_1 = y_1 - \tilde{y}_{1|0} = y_1 \quad F_1 = \mathrm{Var}(y_1|Y_0) = \mathrm{Var}(\nu_1) = P_{1|0} + \sigma_\epsilon^2 = \frac{\sigma_\eta^2}{1-\phi^2} + \sigma_\epsilon^2.$$

Note that F_1 is the unconditional variance of y_t. The updating equations will provide the mean and variance of the distribution of μ_1 given y_1:

$$\begin{aligned} \tilde{\mu}_{1|1} &= \mathrm{E}(\mu_1|Y_1) = \tilde{\mu}_{1|0} + P_{1|0}F_1^{-1}\nu_1 = \frac{\sigma_\eta^2}{1-\phi^2}\left[\frac{\sigma_\eta^2}{1-\phi^2} + \sigma_\epsilon^2\right]^{-1} y_1 \\ P_{1|1} &= \mathrm{Var}(\mu_1|Y_1) = P_{1|0} - P_{1|0}F_1^{-1}P_{1|0} = \frac{\sigma_\eta^2}{1-\phi^2}\left[1 - \frac{\sigma_\eta^2/(1-\phi^2)}{\sigma_\eta^2/(1-\phi^2) + \sigma_\epsilon^2}\right]. \end{aligned}$$

It should be noticed that if $\sigma_\epsilon^2 = 0$, $\tilde{\mu}_{1|1} = y_1$ and $P_{1|1} = 0$ as the AR(1) process is observed without error. On the contrary, when $\sigma_\epsilon^2 > 0$, y_1 will be shrunk towards zero by an amount depending on the relative contribution of the signal to the total variation.

The one-step-ahead prediction of the state and the state prediction error variance are:

$$\begin{aligned} \tilde{\mu}_{2|1} &= \mathrm{E}(\mu_2|Y_1)\tilde{\mu}_{2|1} = \phi\mathrm{E}(\mu_1|Y_1) + \mathrm{E}(\eta_1|Y_1) = \phi\tilde{\mu}_{1|1} \\ P_{2|1} &= \mathrm{Var}(\mu_2|Y_1) = \mathrm{E}(\mu_2 - \phi\tilde{\mu}_{1|0})^2 = \mathrm{E}[\phi(\mu_1 - \tilde{\mu}_{1|0}) + \eta_1]^2 = \phi^2 P_{1|1} + \sigma_\eta^2. \end{aligned}$$

At time $t = 2$, $\tilde{y}_{2|1} = \mathrm{E}(y_2|Y_1) = \tilde{\mu}_{2|1} = \phi\tilde{\mu}_{1|1}$, so that $\nu_2 = y_2 - \tilde{y}_{2|1} = y_2 - \tilde{\mu}_{2|1}$ and $F_2 = \mathrm{Var}(y_2|Y_1) = \mathrm{Var}(\nu_2) = P_{2|1} + \sigma_\epsilon^2$, and so forth.

The KF equations (9) give for $t = 1, \ldots, n$,

$$\begin{array}{ll} \nu_t = y_t - \tilde{\mu}_{t|t-1}, & F_t = P_{t|t-1} + \sigma_\epsilon^2, \\ & K_t = \phi P_{t|t-1}F_t^{-1}, \\ \tilde{\mu}_{t+1|t} = \phi\tilde{\mu}_{t|t-1} + K_t\nu_t, & P_{t+1|t} = \phi^2 P_{t|t-1} + \sigma_\eta^2 - \phi^2 P_{t|t-1}^2 F_t^{-1}. \end{array}$$

Notice that $\sigma_\epsilon^2 = 0 \Rightarrow F_t = P_{t|t-1} = \sigma_\eta^2$ and $\tilde{y}_{t+1|t} = \tilde{\mu}_{t+1|t} = \phi y_t$.

3.4 Nonstationarity and regression effects

Consider the local level model,

$$\begin{aligned} y_t &= \mu_t + \epsilon_t \quad \epsilon_t \sim \text{NID}(0, \sigma_\epsilon^2), \\ \mu_{t+1} &= \mu_t + \eta_t, \quad \eta_t \sim \text{NID}(0, \sigma_\eta^2). \end{aligned}$$

which is obtained as a limiting case of the above AR(1) plus noise model, letting $\phi = 1$. The signal is a nonstationary process. How do we handle initial conditions in this case? We may alternatively assume:

i Fixed initial conditions: the latent process has started at time $t = 0$ with μ_0 representing a fixed and unknown quantity.

ii Diffuse (random) initial conditions: the process has started in the remote past, so that at time $t = 1$, μ_1 has a degenerate distribution centered at zero, $\tilde{\mu}_{1|0} = 0$, but with variance tending to infinity: $P_{1|0} = \kappa, \kappa \to \infty$.

In the first case, the model is rewritten as $y_t = \mu_0 + \alpha_t + \epsilon_t, \alpha_{t+1} = \alpha_t + \eta_t, \alpha_1 \sim N(\tilde{\alpha}_{1|0}, P_{1|0}), \tilde{\alpha}_{1|0} = 0, P_{1|0} = \sigma_\eta^2$, which is a particular case of the augmented state space model (3). The generalized least squares estimator of μ_1 is $\hat{\mu}_0 = (\mathbf{i}'\boldsymbol{\Sigma}^{-1}\mathbf{i})^{-1}\mathbf{i}\boldsymbol{\Sigma}^{-1}\mathbf{y}$, where $\mathbf{y}$ is the stack of the observations, $\mathbf{i}$ is a vector of 1's and $\boldsymbol{\Sigma} = \sigma_\epsilon^2\mathbf{I} + \sigma_\eta^2\mathbf{C}\mathbf{C}'$, where $\mathbf{C}$ is lower triangular with unit elements. We shall provide a more systematic treatment of the filtering problem for nonstationary processes in section (4.2). In particular, the GLS estimator can be computed efficiently by the augmented KF. For the time being we show that, under diffuse initial conditions, after processing one observation, the usual KF provides proper inferences. At time $t = 1$ the first update of the KF, with initial conditions $\tilde{\mu}_{1|0} = 0$ and $P_{1|0} = \kappa$, gives:

$$\begin{aligned} \nu_1 &= y_1, & F_1 &= \kappa + \sigma_\epsilon^2, \\ & & K_1 &= \kappa/(\kappa + \sigma_\epsilon^2), \\ \tilde{\mu}_{2|1} &= y_1\kappa/(\kappa + \sigma_\epsilon^2) & P_{2|1} &= \sigma_\epsilon^2\kappa/(\kappa + \sigma_\epsilon^2) + \sigma_\eta^2. \end{aligned}$$

The distribution of ν_1 is not proper, as y_1 is nonstationary and $F_1 \to \infty$ if we let $\kappa \to \infty$. Also, by letting $\kappa \to \infty$, we obtain the limiting values $K_1 = 1$, $\tilde{\mu}_{2|1} = y_1$ $P_{2|1} = \sigma_\epsilon^2 + \sigma_\eta^2$. Notice that $P_{2|1}$ no longer depends upon κ and $\nu_2 = y_2 - y_1$ has a proper distribution, $\nu_2 \sim \text{N}(0, F_2)$, with finite $F_2 = \sigma_\eta^2 + 2\sigma_\epsilon^2$. In general, the innovations ν_t, for $t > 1$, can be expressed as a linear combination of $\Delta y_t, \Delta y_{t-1}, \ldots$, and thus they possess a proper distribution.

4 Maximum Likelihood Estimation

Let $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^k$ denote a vector containing the so-called hyperparameters, i.e. the vector of structural parameters other than the scale factor σ^2. The state space model depends on $\boldsymbol{\theta}$ via the system matrices $\mathbf{Z}_t = \mathbf{Z}_t(\boldsymbol{\theta}), \mathbf{G}_t = \mathbf{G}_t(\boldsymbol{\theta}), \mathbf{T}_t = \mathbf{T}_t(\boldsymbol{\theta}), \mathbf{H}_t = \mathbf{H}_t(\boldsymbol{\theta})$ and via the initial conditions $\tilde{\boldsymbol{\alpha}}_{1|0}, \mathbf{P}_{1|0}$.

Whenever possible, the constraints in the parameter space $\boldsymbol{\Theta}$ are handled by transformations. Also, one of the variance parameter is attributed the role of the scale parameter σ^2. For instance, for the local level model, we set: $\mathbf{Z} = \mathbf{T} = 1, \mathbf{G} = [1, 0], \sigma^2 = \sigma_\epsilon^2$, $\boldsymbol{\varepsilon}_t \sim \text{NID}(0, \sigma_\epsilon^2 I_2)$, $\mathbf{H} = [0, e^\theta], \theta = \frac{1}{2}\ln q$, where $q = \sigma_\eta^2/\sigma_\epsilon^2$ is the signal to noise ratio.

As a second example, consider the Harvey-Jaeger (1997) decomposition of US gross domestic product(GDP): $y_t = \mu_t + \psi_t$, where μ_t is a local linear trend and ψ_t is a stochastic cycle. The state space representation has $\boldsymbol{\alpha}_t = [\mu_t \;\; \beta_t \;\; \psi_t \;\; \psi_t^*]'$, $\mathbf{Z} = [1, 0, 1, 0]$, $\mathbf{G} = [0, 0, 0, 0]$, $\mathbf{T} = \text{diag}(\mathbf{T}_\mu, \mathbf{T}_\psi)$,

$$\mathbf{T}_\mu = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \quad \mathbf{T}_\psi = \rho \begin{bmatrix} \cos\lambda & \sin\lambda \\ -\sin\lambda & \cos\lambda \end{bmatrix},$$

$$\mathbf{H} = \text{diag}\left(\frac{\sigma_\eta}{\sigma_\kappa}, \frac{\sigma_\zeta}{\sigma_\kappa}, 1, 1\right); \quad \boldsymbol{\varepsilon}_t = \begin{bmatrix} \eta_t\sigma_\kappa/\sigma_\eta \\ \zeta_t\sigma_\kappa/\sigma_\zeta \\ \kappa_t \\ \kappa_t^* \end{bmatrix} \sim \text{N}(0, \sigma_\kappa^2 I_4)$$

The parameter ρ is a damping factor, taking values in (0,1), and λ is the cycle frequency, restricted in the range $[0, \pi]$. Moreover, the parameters σ_η^2 and σ_ζ^2 take nonnegative values. The parameter σ_κ^2 is the scale of the state space disturbance which will be concentrated out of the likelihood function.

We reparameterize the model in terms of the vector $\boldsymbol{\theta}$, which has four unrestricted elements, so that $\boldsymbol{\Theta} \subseteq \mathbb{R}^4$, related to the original hyperparameters by:

$$\frac{\sigma_\eta^2}{\sigma_\kappa^2} = \exp(2\theta_1), \qquad \frac{\sigma_\zeta^2}{\sigma_\kappa^2} = \exp(2\theta_2),$$

$$\rho = \frac{|\theta_3|}{\sqrt{1+\theta_3^2}}, \qquad \lambda = \frac{2\pi}{2 + \exp\theta_4}.$$

Let $\ell(\boldsymbol{\theta}, \sigma^2)$ denote the log-likelihood function, that is the logarithm of the joint density of the sample time series $\{y_1, \ldots, y_n\}$ as a function of the parameters θ, σ^2.

The log-likelihood can be evaluated by the prediction error decomposition:

$$\ell(\boldsymbol{\theta}, \sigma^2) = \ln g(\mathbf{y}_1, \ldots, \mathbf{y}_n; \boldsymbol{\theta}, \sigma^2) = \sum_{t=1}^{n} \ln g(\mathbf{y}_t | \mathbf{Y}_{t-1}; \boldsymbol{\theta}, \sigma^2).$$

Here $g(\cdot)$ denotes the Gaussian probability density function. The predictive density $g(\mathbf{y}_t | \mathbf{Y}_{t-1}; \boldsymbol{\theta}, \sigma^2)$ is evaluated with the support of the KF, as $\mathbf{y}_t | \mathbf{Y}_{t-1} \sim \text{NID}(\tilde{\mathbf{y}}_{t|t-1}, \sigma^2 \mathbf{F}_t)$, so that

$$\ell(\boldsymbol{\theta}, \sigma^2) = -\frac{1}{2}\left(Nn \ln\sigma^2 + \sum_{t=1}^{n} \ln|\mathbf{F}_t| + \frac{1}{\sigma^2}\sum_{t=1}^{n} \boldsymbol{\nu}_t' \mathbf{F}_t^{-1} \boldsymbol{\nu}_t\right). \tag{6}$$

The scale parameter σ^2 can be concentrated out of the LF: maximising $\ell(\boldsymbol{\theta}, \sigma^2)$ with respect to σ^2 yields

$$\hat{\sigma}^2 = \sum_t \boldsymbol{\nu}_t' \mathbf{F}_t^{-1} \boldsymbol{\nu}_t / (Nn).$$

The profile (or concentrated) likelihood is

$$\ell_{\sigma^2}(\boldsymbol{\theta}) = -\frac{1}{2}\left[Nn(\ln\hat{\sigma}^2 + 1) + \sum_{t=1}^{n} \ln|\mathbf{F}_t|\right]. \tag{7}$$

This function can be maximised numerically by a quasi-Newton optimisation routine, by iterating the following updating scheme:

$$\tilde{\boldsymbol{\theta}}_{k+1} = \tilde{\boldsymbol{\theta}}_k - \lambda_k \left[\nabla^2 \ell_{\sigma^2}(\tilde{\boldsymbol{\theta}}_k)\right]^{-1} \nabla \ell_{\sigma^2}(\tilde{\boldsymbol{\theta}}_k),$$

where λ_k is a variable step-length, and $\nabla \ell_{\sigma^2}(\tilde{\boldsymbol{\theta}}_k)$ and $\nabla^2 \ell_{\sigma^2}(\tilde{\boldsymbol{\theta}}_k)$ are respectively the gradient and hessian, evaluated at $\tilde{\boldsymbol{\theta}}_k$. The analytical gradient and hessian can be obtained in parallel to the Kalman filter recursions; see Harvey (1989) and Proietti (1999), for an application.

The innovations are a martingale difference sequence, $\mathrm{E}(\boldsymbol{\nu}_t|\mathbf{Y}_{t-1}) = 0$, which implies that they are uncorrelated with any function of their past: using the law of iterated expectations $\mathrm{E}(\boldsymbol{\nu}_t\boldsymbol{\nu}_{t-j}|\mathbf{Y}_{t-1}) = 0$. Under Gaussianity they will also be independent.

The KF performs a linear transformation of the observations with unit Jacobian: if $\boldsymbol{\nu}$ denotes the stack of the innovations and $\mathbf{y}$ that of the observations: then $\boldsymbol{\nu} = \mathbf{C}^{-1}\mathbf{y}$, where $\mathbf{C}^{-1}$ is a lower triangular matrix such that $\boldsymbol{\Sigma}_y = \mathrm{Cov}(\mathbf{y}) = \sigma^2\mathbf{C}\mathbf{F}\mathbf{C}'$,

$$\mathbf{C} = \begin{bmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ -\mathbf{Z}_2\mathbf{K}_1 & \mathbf{I} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ -\mathbf{Z}_3\mathbf{L}_{3,2}\mathbf{K}_1 & -\mathbf{Z}_3\mathbf{K}_2 & \mathbf{I} & \ddots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ -\mathbf{Z}_{n-1}\mathbf{L}_{n-1,2}\mathbf{K}_1, & -\mathbf{Z}_{n-1}\mathbf{L}_{n-1,3}\mathbf{K}_2, & \dots & \ddots & \mathbf{I} & \mathbf{0} \\ -\mathbf{Z}_n\mathbf{L}_{n,2}\mathbf{K}_1, & -\mathbf{Z}_n\mathbf{L}_{n,3}\mathbf{K}_2, & -\mathbf{Z}_n\mathbf{L}_{n,4}\mathbf{K}_3, & \dots & -\mathbf{Z}_n\mathbf{K}_{n-1}, & \mathbf{I} \end{bmatrix}, \quad (8)$$

where $\mathbf{L}_t = \mathbf{T}_t - \mathbf{K}_t\mathbf{Z}_t'$, and $\mathbf{L}_{t,s} = \mathbf{L}_{t-1}\mathbf{L}_{t-2}\cdots\mathbf{L}_s$ for $t > s$, $\mathbf{L}_{t,t} = \mathbf{I}$ and $\mathbf{F} = \mathrm{diag}(\mathbf{F}_1, \dots, \mathbf{F}_t, \dots, \mathbf{F}_n)$. Hence, $\boldsymbol{\nu}_t$ is a linear combination of the current and past observations and is orthogonal to the information set $\mathbf{Y}_{t-1}$. As a result $|\boldsymbol{\Sigma}_y| = \sigma^{2n}|\mathbf{F}| = \sigma^{2n}\prod_t |\mathbf{F}_t|$ and $\mathbf{y}'\boldsymbol{\Sigma}_y^{-1}\mathbf{y} = \frac{1}{\sigma^2}\boldsymbol{\nu}'\mathbf{F}^{-1}\boldsymbol{\nu} = \frac{1}{\sigma^2}\sum_t \boldsymbol{\nu}_t\mathbf{F}_t^{-1}\boldsymbol{\nu}_t$.

4.1 Properties of maximum likelihood estimators

Under regularity conditions, the maximum likelihood estimators of $\boldsymbol{\theta}$ are consistent and asymptotically normal, with covariance matrix equal to the inverse of the asymptotic Fisher information matrix (see Caines, 1988). Besides the technical conditions regarding the existence of derivatives and their continuity about the true parameter, regularity requires that the model is identifiable and the true parameter values do not lie on the boundary of the parameter space. For the AR(1) plus noise model introduced in section 2.3 these conditions are violated, for instance, when $\phi = 0$ and when $\phi = 1$ or $\sigma_\epsilon^2 = 0$, respectively. While testing for the null hypothesis $\phi = 0$ against the alternative $\phi \neq 0$ is standard, based on the t-statistics of the coefficient y_{t-1} in the regression of y_t on y_{t-1} or on the first order autocorrelation, testing for unit roots or deterministic effects is much more involved, since likelihood ratio tests do not have the usual chi square distribution. Testing for deterministic and non stationary effects in unobserved component models is considered in Nyblom (1996) and Harvey (2001).

Pagan (1980) has derived sufficient conditions for asymptotic identifiability in stationary models and sufficient conditions for consistency and asymptotic normality of the maximum likelihood estimators in non stationary but asymptotically identifiable models. Strong consistency of the maximum likelihood estimator in the general case of a non compact parameter space is proved in Hannan and Deistler (1988). Recently, full asymptotic theory for maximum likelihood estimation of nonstationary state space models has been provided by Chang, Miller and Park (2009).

4.2 Profile and Marginal likelihood for Nonstationary Models with Fixed and Regression Effects

Let us consider the case when nonstationary state elements and exogenous variables are present. The relevant state space form is (3), and the initial conditions are stated in (4).

Let us start from the simple case when the vector $\boldsymbol{\beta}$ is fixed and known, so that $\boldsymbol{\alpha}_1 \sim \mathrm{N}(\tilde{\boldsymbol{\alpha}}^*_{1|0} + \mathbf{W}_0\boldsymbol{\beta}, \sigma^2\mathbf{P}^*_{1|0})$, where $\mathbf{P}^*_{1|0} = \mathbf{H}_0\mathbf{H}_0'$.

The KF for this model becomes, for $t = 1, \ldots, n$:

$$\begin{aligned} \boldsymbol{\nu}^*_t &= \mathbf{y}_t - \mathbf{Z}_t\tilde{\boldsymbol{\alpha}}^*_{t|t-1} - \mathbf{X}_t\boldsymbol{\beta}, & \mathbf{F}^*_t &= \mathbf{Z}_t\mathbf{P}^*_{t|t-1}\mathbf{Z}_t' + \mathbf{G}_t\mathbf{G}_t', \\ & & \mathbf{K}^*_t &= (\mathbf{T}_t\mathbf{P}_{t|t-1}\mathbf{Z}_t' + \mathbf{H}_t\mathbf{G}_t')\mathbf{F}^{*-1}_t, \\ \tilde{\boldsymbol{\alpha}}^*_{t+1|t} &= \mathbf{T}_t\tilde{\boldsymbol{\alpha}}^*_{t|t-1} + \mathbf{W}_t\boldsymbol{\beta} + \mathbf{K}^*_t\boldsymbol{\nu}^*_t, & \mathbf{P}^*_{t+1|t} &= \mathbf{T}_t\mathbf{P}^*_{t|t-1}\mathbf{T}_t' + \mathbf{H}_t\mathbf{H}_t' - \mathbf{K}^*_t\mathbf{F}^*_t\mathbf{K}^{*'}_t \end{aligned} \tag{9}$$

We refer to this filter as KF($\boldsymbol{\beta}$). Apart from a constant term, the log likelihood is as given in (6), whereas, (7) is the profile likelihood.

The KF and the definition of the likelihood need to be amended when nonstationary and regression effects are present. An instance is provided by the local level model, for which $\mathbf{Z}_t = 1$, $\mathbf{X}_t = 0$, $\boldsymbol{\alpha}_t = \mu_t$, $\mathbf{G}_t = [1, 0]$, $\sigma^2 = \sigma^2_\epsilon$, $\boldsymbol{\varepsilon}_t = [\epsilon_t, \sigma_\epsilon\eta_t/\sigma_\eta]'$, $\mathbf{H}_t = [0, \sigma_\eta/\sigma_\epsilon]$, $\mathbf{T}_t = 1$, $\mathbf{W}_t = 0$,

$$\tilde{\boldsymbol{\alpha}}^*_{1|0} = 0, \mathbf{W}_0 = 1, \boldsymbol{\beta} = \mu_0, \mathbf{H}_0 = [0, \sigma_\eta/\sigma_\epsilon].$$

If a scalar explanatory variable is present, x_t, with coefficient γ: $\mathbf{X}_t = [0, x_t], \boldsymbol{\beta} = [\mu_0, \gamma]', \mathbf{W}_0 = [1, 0], \mathbf{W}_t = [0, 0], t > 0$.

When $\boldsymbol{\beta}$ is fixed but unknown, Rosenberg (1973) showed that it can be concentrated out of the likelihood function and that its generalised least square estimate is obtained from the output of an augmented KF. In fact, $\boldsymbol{\alpha}_1$ has mean $\tilde{\boldsymbol{\alpha}}_{1|0} = \tilde{\boldsymbol{\alpha}}^*_{1|0} + \mathbf{W}_0\boldsymbol{\beta}$ and a covariance matrix $\mathbf{P}^*_{1|0} = \sigma^2\mathbf{H}_0\mathbf{H}_0'$. Defining $\mathbf{A}_{1|0} = -\mathbf{W}_0$, rewriting $\tilde{\boldsymbol{\alpha}}_{1|0} = \tilde{\boldsymbol{\alpha}}^*_{1|0} - \mathbf{A}_{1|0}\boldsymbol{\beta}$, and running the KF recursions for a fixed $\boldsymbol{\beta}$, we obtain the set of innovations $\boldsymbol{\nu}_t = \boldsymbol{\nu}^*_t - \mathbf{V}_t\boldsymbol{\beta}$ and one-step-ahead state predictions $\tilde{\boldsymbol{\alpha}}_{t+1|t} = \tilde{\boldsymbol{\alpha}}^*_{t+1|t} - \mathbf{A}_{t+1|t}\boldsymbol{\beta}$, as a linear function of $\boldsymbol{\beta}$.

In the above expressions the starred quantities, $\boldsymbol{\nu}^*_t$ and $\tilde{\boldsymbol{\alpha}}^*_{t+1|t}$, are produced by the KF run with $\boldsymbol{\beta} = \mathbf{0}$, i.e. with initial conditions $\tilde{\boldsymbol{\alpha}}^*_{1|0}$ and $\mathbf{P}^*_{1|0}$, hereby denoted KF($\mathbf{0}$). The latter also computes the matrices $\mathbf{F}^*_t$, $\mathbf{K}^*_t$ and $\mathbf{P}^*_{t+1|t}, t = 1, \ldots, n$, that do not depend on $\boldsymbol{\beta}$.

The matrices $\mathbf{V}_t$ and $\mathbf{A}_{t+1|t}$ are generated by the following recursions, that are run in parallel to KF($\mathbf{0}$):

$$\mathbf{V}_t = \mathbf{X}_t - \mathbf{Z}_t\mathbf{A}_{t|t-1}, \quad \mathbf{A}_{t+1|t} = \mathbf{T}_t\mathbf{A}_{t|t-1} + \mathbf{W}_t + \mathbf{K}^*_t\mathbf{V}_t, \quad t = 1, \ldots, T, \tag{10}$$

with initial value $\mathbf{A}_{1|0} = -\mathbf{W}_0$. Notice that this amounts to running the same filter, KF($\mathbf{0}$), on each of the columns of the matrix $\mathbf{U}_t$.

Then, replacing $\boldsymbol{\nu}_t = \boldsymbol{\nu}^*_t - \mathbf{V}_t\boldsymbol{\beta}$ into the expression for the log-likelihood (6), and defining $\mathbf{s}_n = \sum_1^n \mathbf{V}_t'\mathbf{F}^{*-1}_t\boldsymbol{\nu}^*_t$ and $\mathbf{S}_n = \sum_1^n \mathbf{V}_t'\mathbf{F}^{*-1}_t\mathbf{V}_t$, yields, apart from a constant term:

$$\ell(\boldsymbol{\theta}, \sigma^2, \boldsymbol{\beta}) = -\frac{1}{2}\left(Nn\ln\sigma^2 \sum_{t=1}^n \ln|\mathbf{F}^*_t| + \sigma^{-2}\left[\sum_{t=1}^n \boldsymbol{\nu}^{*\prime}_t\mathbf{F}^{*-1}_t\boldsymbol{\nu}^*_t - 2\boldsymbol{\beta}'\mathbf{s}_n + \boldsymbol{\beta}'\mathbf{S}_n\boldsymbol{\beta}\right]\right). \tag{11}$$

Hence, the maximum likelihood estimate of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = \mathbf{S}_n^{-1}\mathbf{s}_n$. This is coincident with the generalized least square estimator. The profile likelihood (with respect to $\boldsymbol{\beta}$) is

$$\ell_\beta(\boldsymbol{\theta}, \sigma^2) = -\frac{1}{2}\left(Nn \ln \sigma^2 + \sum_{t=1}^{n} \ln |\mathbf{F}_t^*| + \sigma^{-2}\left[\sum_{t=1}^{n} \boldsymbol{\nu}_t^{*\prime} \mathbf{F}_t^{*-1} \boldsymbol{\nu}_t^* - \mathbf{s}_n' \mathbf{S}_n^{-1} \mathbf{s}_n \right] \right) \tag{12}$$

The MLE of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{Nn}\left[\sum_{t=1}^{n} \boldsymbol{\nu}_t^{*\prime} \mathbf{F}_t^{*-1} \boldsymbol{\nu}_t^* - \mathbf{s}_n' \mathbf{S}_n^{-1} \mathbf{s}_n \right]$$

and the profile likelihood (also with respect to σ^2) is

$$\ell_{\beta,\sigma^2}(\boldsymbol{\theta}) = -\frac{1}{2}\left[Nn(\ln \hat{\sigma}^2 + 1) + \sum_{t=1}^{n} \ln |\mathbf{F}_t^*| \right]. \tag{13}$$

The vector $\boldsymbol{\beta}$ is said to be diffuse if $\boldsymbol{\beta} \sim \mathrm{N}(0, \boldsymbol{\Sigma}_\beta)$, where $\boldsymbol{\Sigma}_\beta^{-1} \to \mathbf{0}$. The diffuse likelihood is defined as the limit of $\ell(\boldsymbol{\theta}, \sigma^2, \boldsymbol{\beta})$ as $\boldsymbol{\Sigma}_\beta^{-1} \to \mathbf{0}$. This yields

$$\ell_\infty(\boldsymbol{\theta}, \sigma^2) = -\frac{1}{2}\left\{ N(n-k) \ln \sigma^2 + \sum \ln |\mathbf{F}_t^*| + \ln |\mathbf{S}_n| + \sigma^{-2}\left[\sum \boldsymbol{\nu}_t^{*\prime} \mathbf{F}_t^{*-1} \boldsymbol{\nu}_t^* - \mathbf{s}_n' \mathbf{S}_n^{-1} \mathbf{s}_n \right], \right\}$$

where k is the number of elements of $\boldsymbol{\beta}$. The MLE of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{N(n-k)}\left[\sum_{t=1}^{n} \boldsymbol{\nu}_t^{*\prime} \mathbf{F}_t^{*-1} \boldsymbol{\nu}_t^* - \mathbf{s}_n' \mathbf{S}_n^{-1} \mathbf{s}_n \right]$$

and the profile likelihood is

$$\ell_{\infty,\sigma^2}(\boldsymbol{\theta}) = -\frac{1}{2}\left[N(n-k)(\ln \hat{\sigma}^2 + 1) + \sum_{t=1}^{n} \ln |\mathbf{F}_t^*| + \ln |\mathbf{S}_n| \right]. \tag{14}$$

The notion of a diffuse likelihood is close to that of a marginal likelihood, being based on reduced rank linear transformation of the series that eliminates dependence on $\boldsymbol{\beta}$; see the next subsection and Francke, Koopman and de Vos (2010).

de Jong (1991) has further shown that the limiting expressions for the innovations, the one-step-ahead prediction of the state vector and the corresponding covariance matrices are

$$\begin{array}{lcllcl} \boldsymbol{\nu}_t & = & \boldsymbol{\nu}_t^* - \mathbf{V}_t \mathbf{S}_{t-1}^{-1} \mathbf{s}_{t-1}, & \mathbf{F}_t & = & \mathbf{F}_t^* + \mathbf{V}_t \mathbf{S}_{t-1}^{-1} \mathbf{V}_t', \\ \tilde{\boldsymbol{\alpha}}_{t|t-1} & = & \tilde{\boldsymbol{\alpha}}_{t|t-1}^* - \mathbf{A}_{t|t-1} \mathbf{S}_{t-1}^{-1} \mathbf{s}_{t-1}, & \mathbf{P}_{t|t-1} & = & \mathbf{P}_{t|t-1}^* + \mathbf{A}_{t|t-1} \mathbf{S}_{t-1}^{-1} \mathbf{A}_{t|t-1}'. \end{array} \tag{15}$$

de Jong and Chu-Chun-Lin (1994) show that the additional recursions (10) referring to initial conditions can be collapsed after a suitable number of updates (given by the rank of $\mathbf{W}_0$).

4.3 Discussion

The augmented state space model (3) can be represented as a linear regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ for a suitable choice of the matrice $\mathbf{X}$, Under the Gaussian assumption $\mathbf{y} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}_u)$, the MLE of $\boldsymbol{\beta}$ is the GLS estimator

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Sigma}_u^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}_u^{-1}\mathbf{y}.$$

Consider the LDL decomposition (see, for instance, Golub and Van Loan, 1996) of the matrix $\boldsymbol{\Sigma}_u$, $\boldsymbol{\Sigma}_u = \mathbf{C}^*\mathbf{F}^*\mathbf{C}^{*'}$, where $\mathbf{C}^*$ has the same structure as (8). The KF($\mathbf{0}$) applied to $\mathbf{y}$ yields $\mathbf{v}^* = \mathbf{C}^{*-1}\mathbf{y}$. When applied to each of the deterministic regressors making up the columns of the $\mathbf{X}$ matrix, it gives $\mathbf{V} = \mathbf{C}^{*-1}\mathbf{X}$. The GLS estimate of $\boldsymbol{\beta}$ is thus obtained from the augmented KF as follows:

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{C}^{*-1'}\mathbf{F}^{*-1}\mathbf{C}^{*-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{C}^{*-1'}\mathbf{F}^{*-1}\mathbf{C}^{*-1}\mathbf{y} \\
&= (\mathbf{V}'\mathbf{F}^{*-1}\mathbf{V})^{-1}\mathbf{V}'\mathbf{F}^{*-1}\mathbf{v}^* \\
&= \left(\textstyle\sum_t \mathbf{V}_t\mathbf{F}_t^{*-1}\mathbf{V}_t'\right)^{-1}\sum_t \mathbf{V}_t\mathbf{F}_t^{*-1}\mathbf{v}_t^*
\end{aligned}$$

The restricted or marginal log-likelihood estimator of $\boldsymbol{\theta}$ is the maximiser of the marginal likelihood defined by Patterson and Thompson (1971) and Harville (1977):

$$\begin{aligned}
\ell_R(\boldsymbol{\theta}, \sigma^2) &= \ell_\beta(\boldsymbol{\theta}, \sigma^2) - \tfrac{1}{2}\left[\ln\left|\mathbf{X}'\boldsymbol{\Sigma}_u^{-1}\mathbf{X}\right| - \ln|\mathbf{X}'\mathbf{X}|\right] \\
&= -\tfrac{1}{2}\left\{\ln|\boldsymbol{\Sigma}_u| + \ln\left|\mathbf{X}'\boldsymbol{\Sigma}_u^{-1}\mathbf{X}\right| - \ln|\mathbf{X}'\mathbf{X}| + \mathbf{y}'\boldsymbol{\Sigma}_u^{-1}\mathbf{y} - \mathbf{y}'\boldsymbol{\Sigma}_u^{-1}\mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}_u^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}_u^{-1}\mathbf{y}\right\}.
\end{aligned}$$

Simple algebra shows that $\ell_R(\boldsymbol{\theta}, \sigma^2) = \ell_\infty(\boldsymbol{\theta}, \sigma^2) + 0.5\ln|\mathbf{X}'\mathbf{X}|$. Thus the marginal MLE is obtained from the assumption that the vector $\boldsymbol{\beta}$ is a diffuse random vector, i.e. it has an improper distribution with a mean of zero and an arbitrarily large variance matrix.

The restricted likelihood is the likelihood of a non-invertible linear transformation of the data, $(\mathbf{I} - \mathbf{Q}_X)\mathbf{y}$, $\mathbf{Q}_X = \mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}_y^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}_y^{-1}$, which eliminates the dependence on $\boldsymbol{\beta}$. The maximiser of $\ell_R(\boldsymbol{\theta}, \sigma^2)$ is preferable to the profile likelihood estimator when n is small and the variance of the random signal is small compared to that of the noise.

4.4 Missing values and sequential processing

In univariate models missing values are handled by skipping the KF updating operations: if y_i is missing at time i, ν_i and F_i cannot be computed and $\tilde{\boldsymbol{\alpha}}_{i+1|i-1} = \mathbf{T}_i\tilde{\boldsymbol{\alpha}}_{i|i-1}$, $\mathbf{P}_{i+1|i-1} = \mathbf{T}_i\mathbf{P}_{i|i-1}\mathbf{T}' + \mathbf{H}_i\mathbf{H}_i'$ are the moments of the two-step-ahead predictive distribution.

For multivariate models, when $\mathbf{y}_i$ is only partially missing, sequential processing must be used. This technique, illustrated by Anderson and Moore (1979) and further developed by Koopman and Durbin (2000) for nonstationary models, provides a very flexible and convenient device for filtering and smoothing and for handling missing values. Our treatment is prevalently based on Koopman and Durbin (2000). However, for the treatment of regression effects and initial conditions we adopt the augmentation approach by de Jong (1991).

Assume, for notation simplicity, a time invariant model with $\mathbf{H}\mathbf{G}' = \mathbf{0}$ (uncorrelated measurement and transition disturbances) and $\mathbf{G}\mathbf{G}' = \mathrm{diag}\{g_i^2, i = 1, \ldots, N\}$, so that the measurements $y_{t,i}$ are conditionally independent, given $\boldsymbol{\alpha}_t$. The latter assumption can be relaxed: a possibility is to include $\mathbf{G}\boldsymbol{\varepsilon}_t$ in the state vector, and set $g_i^2 = 0, \forall i$; alternatively, we can transform the measurement equation so as to achieve that the measurement disturbances are fully idiosyncratic.

The multivariate vectors $\mathbf{y}_t$, $t = 1, \ldots, n$, where some elements can be missing, are stacked one on top of the other to yield a univariate time series $\{y_{t,i}, i = 1, \ldots, N, t = 1, \ldots, n\}$, whose elements are processed sequentially. The state space model for the univariate time series $\{y_{t,i}\}$ is constructed as follows.

The new measurement equation for the i-th element of the vector $\mathbf{y}_t$ is:

$$y_{t,i} = \mathbf{z}_i^{'}\boldsymbol{\alpha}_{t,i} + \mathbf{x}_{t,i}^{\prime}\boldsymbol{\beta} + g_i\varepsilon_{t,i}^{*}, \quad t = 1, \ldots, n, \;\; i = 1, \ldots, N, \;\; \varepsilon_{t,i}^{*} \sim \text{NID}(0, \sigma^2) \tag{16}$$

where $\mathbf{z}_i^{'}$ and $\mathbf{x}_{t,i}^{\prime}$ denote the i-th rows of $\mathbf{Z}$ and $\mathbf{X}_t$, respectively. Notice that (16) has two indices: the time index runs first and it is kept fixed as series index runs.

The transition equation varies with the two indices. For a fixed time index, the transition equation is the identity $\boldsymbol{\alpha}_{t,i} = \boldsymbol{\alpha}_{t,i-1}$, for $i = 2, \ldots, N$, whereas, for $i = 1$,

$$\boldsymbol{\alpha}_{t,1} = \mathbf{T}\boldsymbol{\alpha}_{t-1,N} + \mathbf{W}\boldsymbol{\beta} + \mathbf{H}\boldsymbol{\epsilon}_{t,1}$$

The state space form is completed by the initial state vector which is $\boldsymbol{\alpha}_{1,1} = \mathbf{a}_{1,1} + \mathbf{W}_0\boldsymbol{\beta} + \mathbf{H}_0\boldsymbol{\epsilon}_{1,1}$, where $\text{Var}(\boldsymbol{\epsilon}_{1,1}) = \text{Var}(\boldsymbol{\epsilon}_{t,1}) = \sigma^2\mathbf{I}$.

The augmented Kalman filter, taking into account the presence of missing values, is given by the following definitions and recursive formulae.

- Set the initial values $\mathbf{a}_{1,1} = \mathbf{0}, \mathbf{A}_{1,1} = -\mathbf{W}_0, \mathbf{P}_{1,1} = \mathbf{H}_0\mathbf{H}_0^{\prime}, q_{1,1} = 0, \mathbf{s}_{1,1} = \mathbf{0}, \mathbf{S}_{1,1} = \mathbf{0}$, $d_{1,1} = 0$,
- for $t = 1, \ldots, n$, $i = 1, \ldots, N-1$,
  - if $y_{t,i}^{\dagger}$ is available:

$$\begin{array}{rclrcl}
v_{t,i} & = & y_{t,i} - \mathbf{z}_i^{'}\mathbf{a}_{t,i}, & \mathbf{V}_{t,i}^{\prime} & = & \mathbf{x}_{t,i}^{\prime} - \mathbf{z}_i^{'}\mathbf{A}_{t,i}, \\
f_{t,i} & = & \mathbf{z}_i^{'}\mathbf{P}_{t,i}\mathbf{z}_i^{'} + g_i^2, & \mathbf{K}_{t,i} & = & \mathbf{P}_t\mathbf{z}_i^{'}/f_{t,i} \\
\mathbf{a}_{t,i+1} & = & \mathbf{a}_{t,i} + \mathbf{K}_{t,i}v_{t,i}, & \mathbf{A}_{t,i+1} & = & \mathbf{A}_{t,i} + \mathbf{K}_{t,i}\mathbf{V}_{t,i}^{\prime}, \\
\mathbf{P}_{t,i+1} & = & \mathbf{P}_{t,i} - \mathbf{K}_{t,i}\mathbf{K}_{t,i}^{'}f_t, & & & \\
q_{t,i+1} & = & q_{t,i} + v_{t,i}^2/f_{t,i}, & \mathbf{s}_{t,i+1} & = & \mathbf{s}_{t,i} + \mathbf{V}_{t,i}v_{t,i}/f_{t,i} \\
\mathbf{S}_{t,i+1} & = & \mathbf{S}_{t,i} + \mathbf{V}_{t,i}\mathbf{V}_{t,i}^{\prime}/f_{t,i} & d_{t,i+1} & = & d_{t,i} + \ln f_{t,i} \\
cn & = & cn + 1 & & &
\end{array} \tag{17}$$

    Here, cn counts the number of observations.

  - Else, if $y_{t,i}$ is missing:

$$\begin{array}{rcl}
\mathbf{a}_{t,i+1} & = & \mathbf{a}_{t,i}, \quad \mathbf{A}_{t,i+1} = \mathbf{A}_{t,i}, \\
\mathbf{P}_{t,i+1} & = & \mathbf{P}_{t,i}, \\
q_{t,i+1} & = & q_{t,i}, \quad \mathbf{s}_{t,i+1} = \mathbf{s}_{t,i}, \quad \mathbf{S}_{t,i+1} = \mathbf{S}_{t,i}, \quad d_{t,i+1} = d_{t,i}.
\end{array} \tag{18}$$

- For $i = N$, compute:

$$\begin{array}{rclrcl}
\mathbf{a}_{t+1,1} & = & \mathbf{T}\mathbf{a}_{t,N}, & \mathbf{A}_{t+1,1} & = & \mathbf{W} + \mathbf{T}\mathbf{A}_{t,N}, \\
\mathbf{P}_{t+1,1} & = & \mathbf{T}\mathbf{P}_{t,N}\mathbf{T}^{'} + \mathbf{H}\mathbf{H}^{'}, & & & \\
q_{t+1,1} & = & q_{t,N}, \quad \mathbf{s}_{t+1,1} = \mathbf{s}_{t,N}, & \mathbf{S}_{t+1,1} & = & \mathbf{S}_{t,N}, \quad d_{t+1,1} = d_{t,N}.
\end{array} \tag{19}$$

Under the fixed effects model maximising the likelihood with respect to $\boldsymbol{\beta}$ and σ^2 yields:

$$\hat{\boldsymbol{\beta}} = \mathbf{S}_{n+1,1}^{-1}\mathbf{s}_{n+1,1}, \text{Var}(\hat{\boldsymbol{\beta}}) = \mathbf{S}_{n+1,1}^{-1}, \quad \hat{\sigma}^2 = \frac{q_{n+1,1} - \mathbf{s}'_{n+1,1}\mathbf{S}_{n+1,1}^{-1}\mathbf{s}_{n+1,1}}{cn}, \tag{20}$$

The profile likelihood is $\ell_{\beta,\sigma^2} = -0.5\left[d_{n+1,1} + cn\left(\ln\hat{\sigma}^2 + \ln(2\pi) + 1\right)\right]$.

When $\boldsymbol{\beta}$ is diffuse, the maximum likelihood estimate of the scale parameter is

$$\hat{\sigma}^2 = \frac{q_{n+1,1} - \mathbf{s}'_{n+1,1}\mathbf{S}_{n+1,1}^{-1}\mathbf{s}_{n+1,1}}{cn - k},$$

and the diffuse profile likelihood is:

$$\ell_{\infty} = -0.5\left[d_{n+1,1} + (cn-k)\left(\ln\hat{\sigma}^2 + \ln(2\pi) + 1\right) + \ln|\mathbf{S}_{n+1,1}|\right]. \tag{21}$$

This treatment is useful for handling estimation with mixed frequency data. Also, temporal aggregation can be converted into a systematic sampling problem an handled by sequential processing; see Harvey and Chung (2000) and Frale *et al.* (2011), among others.

4.5 Linear constraints

Suppose that the vector $\boldsymbol{\alpha}_t$ is subject to c linear binding constraints $\mathbf{C}_t\boldsymbol{\alpha}_t = \mathbf{c}_t$, with $\mathbf{C}_t$ and $\mathbf{c}_t$ fixed and known. An example is a Cobb-Douglas production function with time varying elasticities, but constant returns to scale in every time period. See Doran (1992) for further details.

These constraints are handled by augmenting the measurement equation with further c observations:

$$\begin{bmatrix} \mathbf{y}_t \\ \mathbf{c}_t \end{bmatrix} = \begin{bmatrix} \mathbf{Z}_t \\ \mathbf{C}_t \end{bmatrix}\boldsymbol{\alpha}_t + \begin{bmatrix} \mathbf{G}_t \\ \mathbf{0} \end{bmatrix}\boldsymbol{\varepsilon}_t.$$

Non-binding constraints are easily accommodated.

4.6 A simulated example

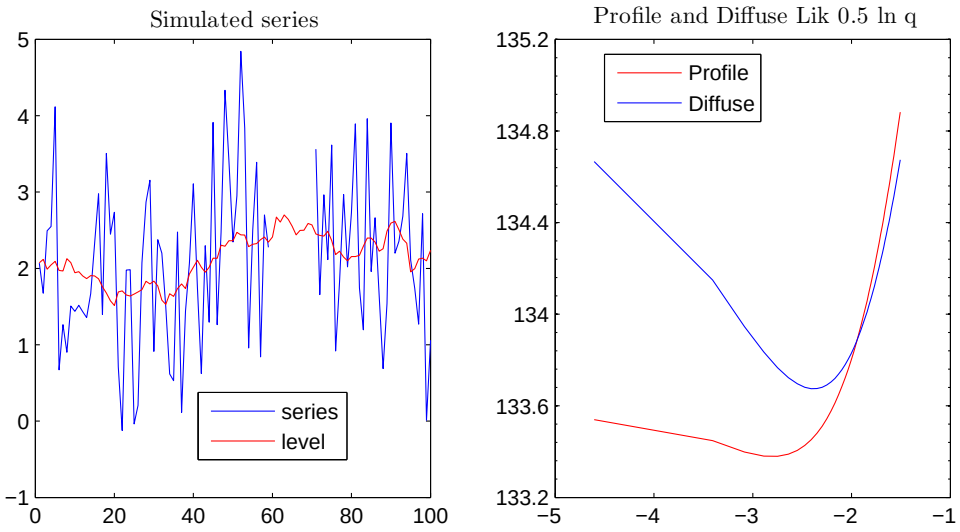
We simulated $n = 100$ observations from a local level model with signal tp noise ratio $q = 0.01$. Subsequently, 10 observations (for t = 60-69) were deleted, and the parameter $0.5\ln q$ estimated by profile and diffuse MLE. Figure 1 displays the simulated series and true level (left), and the profile and diffuse likelihood (right).

The maximiser of the diffuse likelihood is higher and closer to the true value, which amounts to -2.3. This illustrates that the diffuse likelihood in small samples provides a more accurate estimate of the signal to noise ratio when the latter is close to the boundary of the parameter space.

5 The EM Algorithm

Maximum likelihood estimation of the standard time invariant state space model can be carried out by the EM algorithm (see See Shumway and Stoffer, 1982, and Cappè, Moulines and Rydén, 2007). In the sequel we will assume without loss of generality $\sigma^2 = 1$.

Figure 1: Simulated series from a local level model with $q = 0.1$ ($0.5 \ln q = -2.3$) and underlying level (left). Plot of the profile and diffuse likelihood of the parameter $0.5 \ln q$.



Let $\mathbf{y} = [\mathbf{y}_1', \ldots, \mathbf{y}_n]'$, $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1', \ldots, \boldsymbol{\alpha}_n']'$. The log-posterior of the states is $\ln g(\boldsymbol{\alpha}|\mathbf{y}; \boldsymbol{\theta}) = \ln g(\mathbf{y}, \boldsymbol{\alpha}; \boldsymbol{\theta}) - \ln g(\mathbf{y}; \boldsymbol{\theta})$, where the first term on the right hand side is the joint probability density function of the observations and the states, also known as the complete data likelihood, and the subtrahend is the likelihood, $\ell(\boldsymbol{\theta}) = \ln g(\mathbf{y}; \boldsymbol{\theta})$, of the observed data.

The complete data log-likelihood can be evaluated as follows: $\ln g(\mathbf{y}, \boldsymbol{\alpha}; \boldsymbol{\theta}) = \ln g(\mathbf{y}|\boldsymbol{\alpha}; \boldsymbol{\theta}) + \ln g(\boldsymbol{\alpha}; \boldsymbol{\theta})$, where $\ln g(\mathbf{y}|\boldsymbol{\alpha}; \boldsymbol{\theta}) = \sum_{t=1}^{n} \ln g(\mathbf{y}_t|\boldsymbol{\alpha}_t)$, and $\ln g(\boldsymbol{\alpha}; \boldsymbol{\theta}) = \sum_{t=1}^{n} \ln g(\boldsymbol{\alpha}_{t+1}|\boldsymbol{\alpha}_t; \boldsymbol{\theta}) + \ln g(\boldsymbol{\alpha}_1; \boldsymbol{\theta})$. Thus, from (1)-(2),

$$\begin{aligned} \ln g(\mathbf{y}, \boldsymbol{\alpha}; \boldsymbol{\theta}) \quad = \quad & -\tfrac{1}{2}\left[n \ln |\mathbf{G}\mathbf{G}'| + \operatorname{tr}\left\{(\mathbf{G}\mathbf{G}')^{-1} \textstyle\sum_{t=1}^{n} (\mathbf{y}_t - \mathbf{Z}\boldsymbol{\alpha}_t)(\mathbf{y}_t - \mathbf{Z}\boldsymbol{\alpha}_t)'\right\}\right] \\ & -\tfrac{1}{2}\left[n \ln |\mathbf{H}\mathbf{H}'| + \operatorname{tr}\left\{(\mathbf{H}\mathbf{H}')^{-1} \textstyle\sum_{t=2}^{n} (\boldsymbol{\alpha}_{t+1} - \mathbf{T}\boldsymbol{\alpha}_t)(\boldsymbol{\alpha}_{t+1} - \mathbf{T}\boldsymbol{\alpha}_t)'\right\}\right] \\ & -\tfrac{1}{2}\left[\ln |\mathbf{P}_{1|0}| + \operatorname{tr}\left\{\mathbf{P}_{1|0}^{-1} \boldsymbol{\alpha}_1 \boldsymbol{\alpha}_1'\right\}\right] \end{aligned}$$

where $\mathbf{P}_0$ satisfies the matrix equation $\mathbf{P}_{1|0} = \mathbf{T}\mathbf{P}_{1|0}\mathbf{T}' + \mathbf{H}\mathbf{H}'$ and we take, with little loss in generality, $\tilde{\boldsymbol{\alpha}}_{1|0} = \mathbf{0}$.

Given an initial parameter value, $\boldsymbol{\theta}^*$, the EM algorithm iteratively maximizes, with respect to $\boldsymbol{\theta}$, the intermediate quantity (Dempster *et al.*, 1977):

$$Q(\boldsymbol{\theta}; \boldsymbol{\theta}^*) = \mathrm{E}_{\theta^*}\left[\ln g(\mathbf{y}, \boldsymbol{\alpha}; \boldsymbol{\theta})\right] = \int \ln g(\mathbf{y}, \boldsymbol{\alpha}; \boldsymbol{\theta}) g(\boldsymbol{\alpha}|\mathbf{y}; \boldsymbol{\theta}^*) d\boldsymbol{\alpha},$$

which is interpreted as the expectation of the complete data log-likelihood with respect to $g(\boldsymbol{\alpha}|\mathbf{y}; \boldsymbol{\theta}^*)$, which is the conditional probability density function of the unobservable states, given the observations,

evaluated using $\boldsymbol{\theta}^*$. Now,

$$\begin{aligned} Q(\boldsymbol{\theta};\boldsymbol{\theta}^*) &= -\tfrac{1}{2}\left[n\ln|\mathbf{G}\mathbf{G}'| + \mathrm{tr}\left\{(\mathbf{G}\mathbf{G}')^{-1}\textstyle\sum_{t=1}^{n}\left[(\mathbf{y}_t - \mathbf{Z}\tilde{\boldsymbol{\alpha}}_{t|n})(\mathbf{y}_t - \mathbf{Z}\tilde{\boldsymbol{\alpha}}_{t|n})' + \mathbf{Z}\mathbf{P}_{t|n}\mathbf{Z}'\right]\right\}\right] \\ &\quad -\tfrac{1}{2}\left[n\ln|\mathbf{H}\mathbf{H}'| + \mathrm{tr}\left\{(\mathbf{H}\mathbf{H}')^{-1}(\mathcal{S}_\alpha - \mathcal{S}_{\alpha,\alpha-1}\mathbf{T}' - \mathbf{T}\mathcal{S}'_{\alpha,\alpha-1} + \mathbf{T}\mathcal{S}_{\alpha-1}\mathbf{T}')\right\}\right] \\ &\quad -\tfrac{1}{2}\left[\ln|\mathbf{P}_0| + \mathrm{tr}\left\{\mathbf{P}_0^{-1}(\tilde{\boldsymbol{\alpha}}_{0|n}\tilde{\boldsymbol{\alpha}}'_{0|n} + \mathbf{P}_{0|n})\right\}\right] \end{aligned}$$

where $\tilde{\boldsymbol{\alpha}}_{t|n} = \mathrm{E}(\boldsymbol{\alpha}_t|\mathbf{y};\boldsymbol{\theta}^{(j)})$, $\mathbf{P}_{t|n} = \mathrm{Var}(\boldsymbol{\alpha}_t|\mathbf{y};\boldsymbol{\theta}^{(j)})$, and

$$\mathcal{S}_\alpha = \left[\sum_{t=2}^{n}\left(\mathbf{P}_{t+1|n} + \tilde{\boldsymbol{\alpha}}_{t+1|n}\tilde{\boldsymbol{\alpha}}'_{t+1|n}\right)\right],$$

$$\mathcal{S}_{\alpha-1} = \left[\sum_{t=2}^{n}\left(\mathbf{P}_{t|n} + \tilde{\boldsymbol{\alpha}}_{t|n}\tilde{\boldsymbol{\alpha}}'_{t|n}\right)\right], \mathcal{S}_{\alpha,\alpha-1} = \left[\sum_{t=2}^{n}\left(\mathbf{P}_{t+1,t|n} + \tilde{\boldsymbol{\alpha}}_{t+1|n}\tilde{\boldsymbol{\alpha}}'_{t|n}\right)\right].$$

These quantities are evaluated with the support of the Kalman filter and smoother (KFS, see below), adapted to the state space model (1)-(2) with parameter values $\boldsymbol{\theta}^*$. Also, $\mathbf{P}_{t+1,t|n} = \mathrm{Cov}(\boldsymbol{\alpha}_{t+1}, \boldsymbol{\alpha}_t|\mathbf{y};\boldsymbol{\theta}^*)$ is computed using the output of the KFS recursions, as it will be detailed below.

Dempster *et al*. (1977) show that the parameter estimates maximising the log-likelihood $\ell(\boldsymbol{\theta})$, can be obtained by a sequence of iterations, each consisting of an expectation step (E-step) and a maximization step (M-step), that aim at locating a stationary point of $Q(\boldsymbol{\theta};\boldsymbol{\theta}^*)$. At iteration j, given the estimate $\boldsymbol{\theta}^{(j)}$, the E-step deals with the evaluation of $Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(j)})$; this is carried out with the support of the KFS applied to the state space representation (1)-(2) with hyperparameters $\boldsymbol{\theta}^{(j)}$.

The M-step amounts to choosing a new value $\boldsymbol{\theta}^{(j+1)}$, so as to maximize with respect to $\boldsymbol{\theta}$ the criterion $Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(j)})$, i.e., $Q(\boldsymbol{\theta}^{(j+1)};\boldsymbol{\theta}^{(j)}) \geq Q(\boldsymbol{\theta}^{(j)};\boldsymbol{\theta}^{(j)})$. The maximization is in closed form, if we assume that $\mathbf{P}_0$ is an independent unrestricted parameter. Actually, the latter depends on the matrices $\mathbf{T}$ and $\mathbf{H}\mathbf{H}'$, but we will ignore this fact, as it is usually done. For the measurement matrix the M-step consists of maximizing $Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(j)})$ with respect to $\mathbf{Z}$, which gives

$$\hat{\mathbf{Z}}^{(j+1)} = \left(\sum_{t=1}^{n}\mathbf{y}_t\tilde{\boldsymbol{\alpha}}'_{t|n}\right)\mathcal{S}_\alpha^{-1}.$$

The $(j+1)$ update of the matrix $\mathbf{G}\mathbf{G}'$ is given by

$$\widehat{\mathbf{G}\mathbf{G}'}^{(j+1)} = \mathrm{diag}\left\{\frac{1}{n}\sum_{t=1}^{n}\left[\mathbf{y}_t\mathbf{y}'_t - \hat{\mathbf{Z}}^{(j+1)}\tilde{\boldsymbol{\alpha}}_{t|n}\mathbf{y}'_t\right]\right\}.$$

Further, we have:

$$\hat{\mathbf{T}}^{(j+1)} = \mathcal{S}_{\alpha,\alpha-1}\mathcal{S}_{\alpha-1}^{-1}, \quad \widehat{\mathbf{H}\mathbf{H}'}^{(j+1)} = \frac{1}{n}\left(\mathcal{S}_f - \hat{\mathbf{T}}^{(j+1)}\mathcal{S}'_{\alpha,\alpha-1}\right).$$

5.1 Smoothing algorithm

The smoothed estimates $\tilde{\boldsymbol{\alpha}}_{t|n} = \mathrm{E}(\boldsymbol{\alpha}_t|\mathbf{y};\boldsymbol{\theta})$, and their covariance matrix $\mathbf{P}_{t|n} = \mathrm{E}[(\boldsymbol{\alpha}_t - \tilde{\boldsymbol{\alpha}}_{t|n})(\boldsymbol{\alpha}_t - \tilde{\boldsymbol{\alpha}}_{t|n})'|\mathbf{y};\boldsymbol{\theta}]$, are computed by the following backwards recursive formulae, given by Bryson and Ho

(1969) and de Jong (1989), starting at $t = n$, with initial values $\mathbf{r}_n = 0, \mathbf{R}_n = \mathbf{0}$ and $\mathbf{N}_n = 0$: for $t = n-1, \ldots, 1$,

$$\begin{array}{rclrcl} \mathbf{r}_{t-1} & = & \mathbf{L}_t'\mathbf{r}_t + \mathbf{Z}_t'\mathbf{F}_t^{-1}\mathbf{v}_t, & \mathbf{M}_{t-1} & = & \mathbf{L}_t'\mathbf{M}_t\mathbf{L}_t + \mathbf{Z}_t'\mathbf{F}_t^{-1}\mathbf{Z}_t, \\ \tilde{\boldsymbol{\alpha}}_{t|n} & = & \tilde{\boldsymbol{\alpha}}_{t|t-1} + \mathbf{P}_{t|t-1}\mathbf{r}_{t-1}, & \mathbf{P}_{t|n} & = & \mathbf{P}_{t|t-1} - \mathbf{P}_{t|t-1}\mathbf{M}_{t-1}\mathbf{P}_{t|t-1}. \end{array} \tag{22}$$

where $\mathbf{L}_t = \mathbf{T}_t - \mathbf{K}_t\mathbf{Z}'$.

Finally, it can be shown that $\mathbf{P}_{t,t-1|n} = \mathrm{Cov}(\boldsymbol{\alpha}_t, \boldsymbol{\alpha}_{t-1}|\mathbf{y}) = \mathbf{T}_t\mathbf{P}_{t-1|n} - \mathbf{H}_t\mathbf{H}_t'\mathbf{M}_{t-1}\mathbf{L}_{t-1}\mathbf{P}_{t-1|t-2}$.

6 Nonlinear and Non-Gaussian Models

A general state space model is such that the density of the observations is conditionally independent, given the states, i.e.

$$p(\mathbf{y}_1, \ldots, \mathbf{y}_n|\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n; \boldsymbol{\theta}) = \prod_{t=1}^{n} p(\mathbf{y}_t|\boldsymbol{\alpha}_t; \boldsymbol{\theta}), \tag{23}$$

and the transition density has the Markovian structure,

$$p(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n|\boldsymbol{\theta}) = p(\boldsymbol{\alpha}_0|\boldsymbol{\theta}) \prod_{t=0}^{n-1} p(\boldsymbol{\alpha}_{t+1}|\boldsymbol{\alpha}_t; \boldsymbol{\theta}). \tag{24}$$

The measurement and the transition density belong to a given family. The linear Gaussian state space model (1)-(2) arises when $p(y_t|\boldsymbol{\alpha}_t; \boldsymbol{\theta}) \sim \mathrm{N}(\mathbf{Z}_t\boldsymbol{\alpha}_t, \sigma^2\mathbf{G}_t\mathbf{G}_t')$ and $p(\boldsymbol{\alpha}_{t+1}|\boldsymbol{\alpha}_t; \boldsymbol{\theta}) \sim \mathrm{N}(\mathbf{T}_t\boldsymbol{\alpha}_t, \sigma^2\mathbf{H}_t\mathbf{H}_t')$.

An important special case is the class of generalized linear state space models, which are such that the states are Gaussian and the transition model retains its linearity, whereas the observation density belongs to the exponential family. Models for time series observations originating from the exponential family, such as count data with Poisson, binomial, negative binomial and multinomial distributions, and continuous data with skewed distributions such as the exponential and gamma have been considered by West and Harrison (1997), Fahrmeir and Tutz (2000) and Durbin and Koopman (2001), among others. In particular, the latter perform MLE by importance sampling; see section 6.2.

Models for which some or all of the state have discrete support (multinomial) are often referred to as Markov switching models; usually, conditionally on those states, the model retains a Gaussian and linear structure. See Cappé, Moulines and Rydén (2007) and Kim and Nelson (1999) for macroeconomic applications.

In a more general framework, the predictive densities required to form the likelihood via the prediction error decomposition, need not be available in closed form and their evaluation calls for Monte Carlo or deterministic integration methods. Likelihood inference is straightforward only for a class of models with a single source of disturbance, known as observation driven models; see Ord, Koehler and Snyder (1997) and section 6.5.

6.1 Extended Kalman Filter

A nonlinear time series model is such that the observations are functionally related in a nonlinear way to the states, and/or the states are subject to a nonlinear transition function. Nonlinear state space representations typically arise in the context of DSGE models. Assume that the state space model is formulated

as

$$\begin{array}{lcl} \mathbf{y}_t & = & \mathcal{Z}_t(\boldsymbol{\alpha}_t) + \mathcal{G}_t(\boldsymbol{\alpha}_t)\boldsymbol{\varepsilon}_t \\ \boldsymbol{\alpha}_{t+1} & = & \mathcal{T}_t(\boldsymbol{\alpha}_t) + \mathcal{H}_t(\boldsymbol{\alpha}_t)\boldsymbol{\varepsilon}_t, \quad \boldsymbol{\alpha}_1 \sim \mathrm{N}(\tilde{\boldsymbol{\alpha}}_{1|0}, \mathbf{P}_{1|0}), \end{array} \tag{25}$$

where $\mathcal{Z}_t(\cdot)$ and $\mathcal{T}_t(\cdot)$ are known smooth and differentiable functions.

Let $\mathbf{a}_t$ denote a representative value of $\boldsymbol{\alpha}_t$. Then, by Taylor series expansion, the model can be linearized around the trajectory $\mathbf{a}_t, t = 1, \ldots, n$, giving,

$$\begin{array}{lcl} \mathbf{y}_t & = & \tilde{\mathbf{Z}}_t\boldsymbol{\alpha}_t + \mathbf{c}_t + \mathbf{G}_t\boldsymbol{\varepsilon}_t, \\ \boldsymbol{\alpha}_{t+1} & = & \tilde{\mathbf{T}}_t\boldsymbol{\alpha}_t + \mathbf{d}_t + \mathbf{H}_t\boldsymbol{\varepsilon}_t, \quad \boldsymbol{\alpha}_1 \sim \mathrm{N}(\tilde{\boldsymbol{\alpha}}_{1|0}, \mathbf{P}_{1|0}), \end{array} \tag{26}$$

where

$$\tilde{\mathbf{Z}}_t = \left.\frac{\partial \mathcal{Z}_t(\boldsymbol{\alpha}_t)}{\partial \boldsymbol{\alpha}_t}\right|_{\boldsymbol{\alpha}_t = \mathbf{a}_t}, \quad \mathbf{c}_t = \mathcal{Z}_t(\mathbf{a}_t) - \tilde{\mathbf{Z}}_t\mathbf{a}_t, \mathbf{G}_t = \mathcal{G}_t(\mathbf{a}_t),$$

and

$$\tilde{\mathbf{T}}_t = \left.\frac{\partial \mathcal{T}_t(\boldsymbol{\alpha}_t)}{\partial \boldsymbol{\alpha}_t}\right|_{\boldsymbol{\alpha}_t = \mathbf{a}_t}, \quad \mathbf{d}_t = \mathcal{T}_t(\mathbf{a}_t) - \tilde{\mathbf{T}}_t\mathbf{a}_t, \mathbf{H}_t = \mathcal{H}_t(\mathbf{a}_t).$$

The extended Kalman filter results from applying the KF to linearized model. The latter depends on $\mathbf{a}_t$ and we stress this dependence by writing KF($\mathbf{a}_t$). The likelihood of the linearized model is then evaluated by KF($\mathbf{a}_t$), and can be maximized with respect to the unknown parameters. See Jazwinski (1970) and Anderson and Moore (1979, ch. 8).

The issue is the choice of the value $\mathbf{a}_t$ around which the linearization is taken. One possibility is to choose $\mathbf{a}_t = \boldsymbol{\alpha}_{t|t-1}$, where the latter is delivered recursively on line as the observations are processed in (9). A more accurate solution is to use $\mathbf{a}_t = \boldsymbol{\alpha}_{t|t-1}$ for the linearization of the measurement equation and $\mathbf{a}_t = \boldsymbol{\alpha}_{t|t}$ for that of the transition equation, using the prediction-updating variant of the filter of section (3.2).

Assuming, for simplicity $\mathcal{G}_t(\boldsymbol{\alpha}_t) = \mathbf{G}_t$, $\mathcal{H}_t(\boldsymbol{\alpha}) = \mathbf{H}_t$, and $\boldsymbol{\varepsilon}_t \sim \mathrm{NID}(\mathbf{0}, \sigma^2\mathbf{I})$, the linearization can be performed using the *iterated extended KF* (Jazwinski, 1970, ch. 8), which determines the trajectory $\{\mathbf{a}_t\}$ as the maximizer of the posterior kernel:

$$\sum_t (\mathbf{y}_t - \mathcal{Z}_t(\mathbf{a}_t))' (\mathbf{G}_t\mathbf{G}_t')^{-1} (\mathbf{y}_t - \mathcal{Z}_t(\mathbf{a}_t)) + \sum_t (\mathbf{a}_{t+1} - \mathcal{T}_t(\mathbf{a}_t))' (\mathbf{H}_t\mathbf{H}_t')^{-1} (\mathbf{a}_{t+1} - \mathcal{T}_t(\mathbf{a}_t))$$

with respect to $\{\mathbf{a}_t, t = 1, \ldots, n\}$. This is referred to as *posterior mode estimation*, as it locates the posterior mode of $\boldsymbol{\alpha}$ given $\mathbf{y}$, and is carried out iteratively by the following algorithm:

1. Start with at trial trajectory $\{\mathbf{a}_t\}$
2. Linearize the model around it
3. Run the Kalman filter and smoothing algorithm (22) to obtain a new trajectory $\mathbf{a}_t = \tilde{\boldsymbol{\alpha}}_{t|n}$
4. Iterate steps 2-3 until convergence.

Rather than approximating a nonlinear function, the unscented KF (Julier and Uhlmann, 1996, 1997), is based on an approximation of the distribution of $\boldsymbol{\alpha}_t|\mathbf{Y}_t$ based on a deterministic sample of representative *sigma points*, characterised by the same mean and covariance as the true distribution of $\boldsymbol{\alpha}_t|\mathbf{Y}_t$. When these points are propagated using the true nonlinear measurement and transition equations, the mean and covariance of the predictive distributions $\boldsymbol{\alpha}_{t+1}|\mathbf{Y}_t$ and $\mathbf{y}_{t+1}|\mathbf{Y}_t$ can be approximated accurately (up to the second order) by the weighted average of the transformation of the chosen sigma points.

6.2 Likelihood Evaluation via Importance Sampling

Let $p(\mathbf{y})$ denote the joint density of the n observations (as a function of $\boldsymbol{\theta}$, omitted from the notation), as implied by the original non Gaussian and nonlinear model. Let $g(\mathbf{y})$ be the likelihood of the associated linearized model. See Durbin and Koopman (2001) for the linearization of exponential family models, non Gaussian observation densities such as Student's t, as well as non Gaussian state disturbances; for functionally nonlinear models see above.

The estimation of the likelihood via importance sampling is based on the following identity:

$$p(\mathbf{y}) \quad = \quad \int p(\mathbf{y}, \boldsymbol{\alpha}) d\boldsymbol{\alpha} = g(\mathbf{y}) \int \frac{p(\mathbf{y}, \boldsymbol{\alpha})}{g(\mathbf{y}, \boldsymbol{\alpha})} g(\boldsymbol{\alpha}|\mathbf{y}) d\boldsymbol{\alpha} = g(\mathbf{y}) \mathrm{E}_g \left[\frac{p(\mathbf{y}, \boldsymbol{\alpha})}{g(\boldsymbol{\alpha}|\mathbf{y})} \right] \tag{27}$$

The expectation, taken with respect to the conditional Gaussian density $g(\boldsymbol{\alpha}|\mathbf{y})$, can be estimated by Monte Carlo simulation using importance sampling: in particular, after having linearized the model by posterior mode estimation, M samples $\boldsymbol{\alpha}^{(m)}, m = 1, \ldots, M$, are drawn from $g(\boldsymbol{\alpha}|\mathbf{y})$, the importance sampling weights

$$w_m = \frac{p(\mathbf{y}, \boldsymbol{\alpha}^{(m)})}{g(\mathbf{y}, \boldsymbol{\alpha}^{(m)})} = \frac{p(\mathbf{y}|\boldsymbol{\alpha}^{(m)}) p(\boldsymbol{\alpha}^{(m)})}{g(\mathbf{y}|\boldsymbol{\alpha}^{(m)}) g(\boldsymbol{\alpha}^{(m)})},$$

are computed and the the above expectation is estimated by the average $\frac{1}{M} \sum_m w_m$. Sampling from $g(\boldsymbol{\alpha}|\mathbf{y})$ is carried out by the simulation smoother illustrated in the next subsection. The proposal distribution is multivariate normal with mean equal to the posterior mode $\tilde{\boldsymbol{\alpha}}_{t|n}$. The curvature around the mode can also be matched in special cases, in the derivation of the Gaussian linear auxiliary model. See Shepard and Pitt (1997), Durbin and Koopman (2001), and Richard and Zhang (2007) for further details.

6.3 The simulation smoother

The simulation smoother is an algorithm which draws samples from the conditional distribution of the states, or the disturbances, given the observations and the hyperparameters. We focus on the simulation smoother proposed by Durbin and Koopman (2002).

Let $\boldsymbol{\eta}_t$ denote a random vector (e.g. a selection of states or disturbances) and let $\tilde{\boldsymbol{\eta}} = \mathrm{E}(\boldsymbol{\eta}|\mathbf{y})$, where $\boldsymbol{\eta}$ is the stack of the vectors $\boldsymbol{\eta}_t$; $\tilde{\boldsymbol{\eta}}$ is computed by the Kalman filter and smoother. We can write $\boldsymbol{\eta} = \tilde{\boldsymbol{\eta}} + \mathbf{e}$, where $\mathbf{e} = \boldsymbol{\eta} - \tilde{\boldsymbol{\eta}}$ is the smoothing error, with conditional distribution $\mathbf{e}|\mathbf{y} \sim \mathrm{N}(\mathbf{0}, \mathbf{V})$, such that the covariance matrix $\mathbf{V}$ does not depend on the observations, and thus does not vary across the simulations (the diagonal blocks are computed by the smoothing algorithm).

A sample $\boldsymbol{\eta}^*$ from $\boldsymbol{\eta}|\mathbf{y}$ is constructed as follows:

- Draw $(\boldsymbol{\eta}^+, \mathbf{y}^+) \sim g(\boldsymbol{\eta}, \mathbf{y})$.

  As $p(\boldsymbol{\eta}, \mathbf{y}) = g(\boldsymbol{\eta}) g(\mathbf{y}|\boldsymbol{\eta})$, this is achieved by first drawing $\boldsymbol{\eta}^+ \sim g(\boldsymbol{\eta})$ from an unconditional Gaussian distribution, and constructing the pseudo observations $\mathbf{y}^+$ recursively from $\boldsymbol{\alpha}^+_{t+1} = \mathbf{T}_t \boldsymbol{\alpha}^+_t + \mathbf{H}_t \boldsymbol{\epsilon}^+_t, \mathbf{y}^+_t = \mathbf{Z}_t \boldsymbol{\alpha}^+_t + \mathbf{G}_t \boldsymbol{\epsilon}^+_t, t = 1, 2, \ldots, n$, where the initial draw is $\boldsymbol{\alpha}^+_1 \sim \mathrm{N}(\tilde{\boldsymbol{\alpha}}_{1|0}, \mathbf{P}_{1|0})$, so that $\mathbf{y}^+ \sim g(\mathbf{y}|\boldsymbol{\eta})$.

- The Kalman filter and smoother computed on the simulated observations $\mathbf{y}^+_t$ will produce $\tilde{\boldsymbol{\eta}}^+$, and $\boldsymbol{\eta}^+ - \tilde{\boldsymbol{\eta}}^+$ will be the required draw from $\mathbf{e}|\mathbf{y}$.

Hence , $\tilde{\boldsymbol{\eta}} + \boldsymbol{\eta}^+ - \tilde{\boldsymbol{\eta}}^+$ is the required sample from $\boldsymbol{\eta}|\mathbf{y} \sim \mathrm{N}(\tilde{\boldsymbol{\eta}}, \mathbf{V})$.

6.4 Sequential Monte Carlo Methods

For a general state space model, the one-step-ahead predictive densities of the states and the observations, and the filtering density are respectively:

$$\begin{aligned} p(\boldsymbol{\alpha}_{t+1}|\mathbf{Y}_t) &= \int p(\boldsymbol{\alpha}_{t+1}|\boldsymbol{\alpha}_t)p(\boldsymbol{\alpha}_t|\mathbf{Y}_t)d\boldsymbol{\alpha}_t = \mathrm{E}_{\alpha_t|Y_t}\left[p(\boldsymbol{\alpha}_{t+1}|\boldsymbol{\alpha}_t)\right] \\ p(\mathbf{y}_{t+1}|\mathbf{Y}_t) &= \int p(\mathbf{y}_{t+1}|\boldsymbol{\alpha}_{t+1})p(\boldsymbol{\alpha}_{t+1}|\mathbf{Y}_t)d\boldsymbol{\alpha}_{t+1} = \mathrm{E}_{\alpha_{t+1}|Y_t}\left[p(\mathbf{y}_{t+1}|\boldsymbol{\alpha}_{t+1})\right] \\ p(\boldsymbol{\alpha}_{t+1}|\mathbf{Y}_{t+1}) &= p(\boldsymbol{\alpha}_{t+1}|\mathbf{Y}_t)p(\mathbf{y}_{t+1}|\boldsymbol{\alpha}_{t+1})/p(\mathbf{y}_{t+1}|\mathbf{Y}_t) \end{aligned} \tag{28}$$

Sequential Monte Carlo methods provide algorithms, known as *particle filters*, for recursive, or *on-line*, estimation of the predictive and filtering densities in (28). They deal with the estimation of the above expectations as averages over Monte Carlo samples from the reference density, exploiting the fact that $p(\boldsymbol{\alpha}_{t+1}|\boldsymbol{\alpha}_t)$ and $p(\mathbf{y}_{t+1}|\mathbf{Y}_t)$ are easy to evaluate, as they depend solely on the model prior specification.

Assume that at any time t an IID sample of size M from the filtering density $p(\boldsymbol{\alpha}_t|\mathbf{Y}_t)$ is available, with each draw representing a "particle", $\boldsymbol{\alpha}_t^{(i)}, i = 1, \ldots, M$, so that the true density is approximated by the empirical density function:

$$\hat{p}(\boldsymbol{\alpha}_t \in A|\mathbf{Y}_t) = \frac{1}{M}\sum_{i=1}^{M} I(\boldsymbol{\alpha}_t^{(i)} \in A), \tag{29}$$

where $I(\cdot)$ is the indicator function.

The Monte Carlo approximation to the state and measurement predictive densities is obtained by generating $\boldsymbol{\alpha}_{t+1|t}^{(i)} \sim p(\boldsymbol{\alpha}_{t+1}|\boldsymbol{\alpha}_t^{(i)}), i = 1, \ldots, M$ and $\mathbf{y}_{t+1|t}^{(i)} \sim p(\mathbf{y}_{t+1}|\boldsymbol{\alpha}_{t+1}^{(i)}), i = 1, \ldots, M$.

The crucial issue is to obtain a new particle characterisation of the filtering density $p(\boldsymbol{\alpha}_{t+1}|\mathbf{Y}_{t+1})$, avoiding particle degeneracy, i.e. a non representative sample of particles. To iterate the process it is necessary to generate new particles from $p(\boldsymbol{\alpha}_{t+1}|\mathbf{Y}_{t+1})$ with probability mass equal to $1/M$, so that the approximation to the filtering density will have the same form as (29), and the sequential simulation process can progress. A direct application of the last row in 28 suggest a weighted resampling (Rubin, 1987) of the particles $\boldsymbol{\alpha}_{t+1|t}^{(i)} \sim p(\boldsymbol{\alpha}_{t+1}|\boldsymbol{\alpha}_t^{(i)})$, with importance weights $w_i = p(\mathbf{y}_{t+1}|\boldsymbol{\alpha}_{t+1|t}^{(i)})/\sum_{j=1}^{M} p(\mathbf{y}_{t+1}|\boldsymbol{\alpha}_{t+1|t}^{(j)})$. the *resampling step* eliminates particles with low importance weights and propagates those with high w_i's. This basic particle filter is known as the bootstrap (or Sampling/Importance Resampling, SIR) filter; see Gordon, Salmond and Smith (1993) and Kitagawa (1996).

A serious limitation is that the particles, $\boldsymbol{\alpha}_{t+1|t}^{(i)}$, originate from the prior density and are "blind" to the information carried by $\mathbf{y}_{t+1}$; this may deplete the representativeness of the particles when the prior is at conflict with the likelihood, $p(\mathbf{y}_{t+1}|\boldsymbol{\alpha}_{t+1|t}^{(i)})$, resulting in a highly uneven distribution of the weights w_i. A variety of sampling schemes have been proposed to overcome this conflict, such as the *auxiliary particle filter*; see Pitt and Shephard (1999) and Doucet, de Freitas and Gordon (2001).

More generally, in a sequential setting, we aim at simulating $\boldsymbol{\alpha}_{t+1}^{(i)}$ from the target distribution:

$$p(\boldsymbol{\alpha}_{t+1}|\boldsymbol{\alpha}_t, \mathbf{Y}_{t+1}) = \frac{p(\boldsymbol{\alpha}_{t+1}|\boldsymbol{\alpha}_t)p(\mathbf{y}_{t+1}|\boldsymbol{\alpha}_{t+1})}{p(\mathbf{y}_{t+1}|\boldsymbol{\alpha}_t)},$$

where typically, only the numerator is available. Let $g(\boldsymbol{\alpha}_{t+1}|\boldsymbol{\alpha}_t, \mathbf{Y}_{t+1})$ be an importance density, available for sampling $\boldsymbol{\alpha}_{t+1}^{(i)} \sim g(\boldsymbol{\alpha}_{t+1}|\boldsymbol{\alpha}_t^{(i)}, \mathbf{Y}_{t+1})$ and let

$$w_i \propto \frac{p(\mathbf{y}_{t+1}|\boldsymbol{\alpha}_{t+1}^{(i)})p(\boldsymbol{\alpha}_{t+1}^{(i)}|\boldsymbol{\alpha}_t^{(i)})}{g(\boldsymbol{\alpha}_{t+1}|\boldsymbol{\alpha}_t^{(i)}, \mathbf{Y}_{t+1})};$$

M particles are resampled with probabilities proportional to w_i. Notice that SIR arises as a special case with proposals $g(\boldsymbol{\alpha}_{t+1}|\boldsymbol{\alpha}_t, \mathbf{Y}_{t+1}) = p(\boldsymbol{\alpha}_{t+1}|\boldsymbol{\alpha}_t)$, that ignore $\mathbf{y}_{t+1}$. Merwe *et al.* (2000) used the unscented transformation of Julier and Uhlmann (1997) to generate a proposal density. Amisano and Tristani (2010) obtain the proposal density by a local linearization of the observation and transition density. Recently, Winschel and Krätzig (2010) proposed a particle filter that obtains the first two moments of the predictive distributions in (28) by Smolyak Gaussian quadrature use a normal proposal $g(\boldsymbol{\alpha}_{t+1}|\boldsymbol{\alpha}_t, \mathbf{y}_{t+1})$, with mean and variance resulting from a standard updating Kalman filter step (see section 3.2).

Essential and comprehensive references for the literature on sequential MC are Doucet, de Freitas and Gordon (2001) and Cappè, Moulines and Rydén (2007). For macroeconomic applications see Fernández-Villaverde and Rubio Ramírez (2007) and the recent survey by Creal (2012). Poyiadjis, Doucet and Singh (2011) propose sequential MC methods for approximating the score and the information matrix and use it for recursive and batch parameter estimation of nonlinear state space models.

At each update of the particle filter, the contribution to the likelihood of each observation can be thus estimated. However, maximum likelihood estimation by quasi-Newton method is unfeasible as the likelihood is not a continuous function of the parameters. Grid search approaches are only feasible when the size of the parameter space is small. A pragmatic solution consists of adding the parameters in the state vector and assigning a random walk evolution with fixed disturbance variance, as in Kitagawa (1998). In the iterated filtering approach proposed by Ionides, Breto, and King (2006), generalized in Ionides *et al.* (2011), the evolution variance is allowed to tend deterministically to zero.

6.5 Observation driven score models

Observation driven models based on the score of the conditional likelihood are a class of models independently developed by Harvey and Chakravarty (2008), Harvey (2010) and Creal, Koopman and Lucas (2011a, 2011b).

The model specification starts with the conditional probability distribution of $\mathbf{y}_t$, for $t = 1, \ldots, n$,

$$p(\mathbf{y}_t|\boldsymbol{\lambda}_{t|t-1}, \mathbf{Y}_{t-1}; \boldsymbol{\theta}),$$

where $\boldsymbol{\lambda}_{t|t-1}$ is a set of time varying parameters that are fixed at time $t-1$, $\mathbf{Y}_{t-1}$ is the information set up to time $t-1$, and $\boldsymbol{\theta}$ is a vector of static parameters that enter in the specification of the probability distribution of $\mathbf{y}_t$ and in the updating mechanism for $\boldsymbol{\lambda}_t$. The defining feature of these models is that the dynamics that govern the evolution of the time varying parameters are driven by the score of the conditional distribution:

$$\boldsymbol{\lambda}_{t+1|t} = f(\boldsymbol{\lambda}_{t|t-1}, \boldsymbol{\lambda}_{t-1|t-2}, \ldots, \mathbf{s}_t, \mathbf{s}_{t-1}, \ldots, \boldsymbol{\theta})$$

where

$$\mathbf{s}_t \propto \frac{\partial \ell(\boldsymbol{\lambda}_{t|t-1})}{\partial \boldsymbol{\lambda}_{t|t-1}}$$

and $\ell(\boldsymbol{\lambda}_{t|t-1})$ is the log-likelihood function of $\boldsymbol{\lambda}_{t|t-1}$. Given that $\boldsymbol{\lambda}_t$ is updated through the function f, maximum likelihood estimation eventually concerns the parameter vector $\boldsymbol{\theta}$. The proportionality constant linking the score function to $\mathbf{s}_t$ is a matter of choice and may depend on $\boldsymbol{\theta}$ and other features of the distribution, as the following examples show.

The basic GAS(p, q) models (Creal, Koopman and Lucas, 2011) consists in the specification of the conditional observation density

$$p(\mathbf{y}_t|\boldsymbol{\lambda}_{t|t-1}, \mathbf{Y}_{t-1}, \boldsymbol{\theta})$$

along with the generalized autoregressive updating mechanism

$$\boldsymbol{\lambda}_{t+1|t} = \boldsymbol{\delta} + \sum_{i=1}^{p} \mathbf{A}_i(\boldsymbol{\theta})\mathbf{s}_{t-i+1} + \sum_{j=1}^{q} \mathbf{B}_i(\boldsymbol{\theta})\boldsymbol{\lambda}_{t-i+1}$$

where $\boldsymbol{\delta}$ is a vector of constants and $\mathbf{A}_i(\boldsymbol{\theta})$ and $\mathbf{B}_i(\boldsymbol{\theta})$ are coefficient matrices and where $\mathbf{s}_t$ is defined as the standardized score vector, i.e. the score pre-multiplied by the inverse Fisher information matrix $\mathcal{I}_{t|t-1}^{-1}$,

$$\mathbf{s}_t = \mathcal{I}_{t|t-1}^{-1} \frac{\partial \ell(\boldsymbol{\lambda}_{t|t-1})}{\partial \boldsymbol{\lambda}_{t|t-1}}.$$

The recursive equation for $\boldsymbol{\lambda}_{t+1|t}$ can be interpreted as a Gauss-Newton algorithm for estimating $\boldsymbol{\lambda}_{t+1|t}$ through time.

The first order Beta-t-EGARCH model (Harvey and Chakravarty, 2008) is specified as follows,

$$p(y_t|\lambda_{t|t-1}, Y_{t-1}, \boldsymbol{\theta}) \sim t_\nu(0, e^{\lambda_{t|t-1}})$$

$$\lambda_{t+1|t} = \delta + \phi\lambda_{t|t-1} + \kappa s_t$$

where

$$s_t = \frac{(\nu+1)y_t^2}{\nu e^{\lambda_{t|t-1}} + y_t^2} - 1$$

is the score of the conditional density and $\boldsymbol{\theta} = (\delta, \phi, \kappa, \nu)$. It follows from the properties of the Student-t distribution that the random variable

$$b_t = \frac{s_t + 1}{\nu + 1} = \frac{(s_t + 1)/(\nu e^{\lambda_{t|t-1}})}{(\nu + 1)/(\nu e^{\lambda_{t|t-1}})}$$

is distributed like a Beta $\left(\frac{1}{2}, \frac{\nu}{2}\right)$. Based on this property of the score, it is possible to develop full asymptotic theory for the maximum likelihood estimator of $\boldsymbol{\theta}$ (Harvey, 2010). In practice, having fixed an initial condition such as, for $|\phi| < 1$, $\lambda_{1|0} = \frac{\delta}{1-\phi}$, likelihood optimization may be carried out with a Fisher scoring or Newton-Raphson algorithm.

Notice that observation driven models based on the score have the further interpretation of approximating models for non Gaussian state space models, e.g. the AR(1) plus noise model considered in section 2.3. The use of the score as a driving mechanism for time varying parameters was originally introduced by Masreliez (1975) as an approximation of the Kalman filter for treating non Gaussian state space models. The intuition behind using the score is mainly related to its dependence of the on the whole distribution of the observations rather than on the first and second moment.

7 Conclusions

The focus of this chapter was on likelihood inference for time series models that can be represented in state space. Although we have not touched upon the vast area of Bayesian inference, the state space methods presented in this chapter are a key ingredient in designing and implementing Markov chain Monte Carlo sampling schemes.

References

Amisano, G. and Tristani, O. (2010). Euro area inflation persistence in an estimated nonlinear DSGE model. *Journal of Economic Dynamics and Control*, 34, 1837–1858.

Anderson, B.D.O., and J.B. Moore (1979). *Optimal Filtering*. Englewood Cliffs: Prentice-Hall.

Brockwell, P.J. and Davis, R.A. (1991), *Time Series: Theory and Methods*, Springer.

Bryson, A.E., and Ho, Y.C. (1969). *Applied optimal control: optimization, estimation, and control*. Blaisdell Publishing, Waltham, Mass.

Burridge, P. and Wallis, K.F. (1988). Prediction Theory for Autoregressive-Moving Average Processes. *Econometric Reviews*, 7, 65-9.

Caines P.E. (1988). *Linear Stochastic Systems*. Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, New York.

Canova, F. (2007), *Methods for Applied Macroeconomic Research*. Princeton University Press,

Cappé, O., Moulines, E., and Rydén, T. (2005). *Inference in hidden markov models*. Springer Series in Statistics. Springer, New York.

Chang, Y., Miller, J.I., and Park, J.Y. (2009), Extracting a Common Stochastic Trend: Theory with some Applications, *Journal of Econometrics*, 15, 231–247.

Clark, P.K. (1987). The Cyclical Component of U. S. Economic Activity, *The Quarterly Journal of Economics*, 102, 4, 797–814.

Cogley, T., Primiceri, G.E., Sargent, T.J. (2010), Inflation-Gap Persistence in the U.S., *American Economic Journal: Macroeconomics*, 2(1), January 2010, 43–69.

Creal,D. , (2012) A survey of sequential Monte Carlo methods for economics and finance, *Econometric Reviews*, 31, 3, 245–296.

Creal, D., Koopman, S.J. and Lucas A. (2011a), Generalized Autoregressive Score Models with Applications, *Journal of Applied Econometrics*, forthcoming.

Creal, D., Koopman, S.J. and Lucas A. (2011b), A Dynamic Multivariate Heavy-Tailed Model for Time-Varying Volatilities and Correlations, *Journal of Business and Economics Statistics*, 29, 4, 552–563.

de Jong, P. (1988a). The likelihood for a state space model. *Biometrika* 75: 165-9.

de Jong, P. (1989). Smoothing and interpolation with the state space model. *Journal of the American Statistical Association*, 84, 1085-1088.

de Jong, P (1991). The diffuse Kalman filter. *Annals of Statistics* 19, 1073-83.

de Jong, P., and Chu-Chun-Lin, S. (1994). Fast Likelihood Evaluation and Prediction for Nonstationary State Space Models. *Biometrika*, 81, 133-142.

de Jong, P. and Penzer, J. (2004), The ARMA model in state space form, *Statistics and Probability Letters*, 70, 119–125

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data. *Journal of the Royal Statistical Society*, 14, 1:38.

Doran, E. (1992). Constraining Kalman Filter and Smoothing Estimates to Satisfy Time-Varying Restrictions. *Review of Economics and Statistics*, 74, 568-572.

Doucet, A., de Freitas, J. F. G. and Gordon, N. J. (2001). *Sequential Monte Carlo Methods in Practice*. New York: Springer-Verlag.

Durbin, J., and S.J. Koopman (1997). Monte Carlo maximum likelihood estimation for non-Gaussian state space models. *Biometrika* 84, 669-84.

Durbin, J., and Koopman, S.J. (2000). Time series analysis of non-Gaussian observations based on state-space models from both classical and Bayesian perspectives (with discussion). *Journal of Royal Statistical Society, Series* B, 62, 3-56.

Durbin, J., and S.J. Koopman (2001). *Time Series Analysis by State Space Methods*. Oxford University Press, Oxford.

Durbin, J., and S.J. Koopman (2002). A simple and efficient simulation smoother for state space time series analysis. *Biometrika*, 89, 603-615.

Farhmeir, L. and Tutz G. (1994). *Multivariate Statistical Modelling Based Generalized Linear Models*, Springer-Verlag, New-York.

Fernndez-Villaverde, J. and Rubio-Ramrez, J.F. (2005), Estimating Dynamic Equilibrium Economies: Linear versus Non-Linear Likelihood, *Journal of Applied Econometrics*, 20, 891910.

Fernndez-Villaverde, J. and Rubio-Ramrez, J.F. (2007). Estimating Macroeconomic Models: A Likelihood Approach. *Review of Economic Studies*, 74, 1059–1087.

Fernndez-Villaverde, J. (2010), The Econometrics of DSGE Models, *Journal of the Spanish Economic Association* 1, 3–49.

Frale, C., Marcellino, M., Mazzi, G. and Proietti, T. (2011), EUROMIND: A Monthly Indicator of the Euro Area Economic Conditions, *Journal of the Royal Statistical Society - Series A*, 174, 2, 439–470.

Francke, M.K., Koopman, S.J., de Vos, A. (2010), Likelihood functions for state space models with diffuse initial conditions, *Journal of Time Series Analysis* 31, 407–414.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer Series in Statistics. Springer, New York.

Gamerman, D. and Lopes H. F. (2006). *Markov Chain Monte Carlo. Stochastic Simulation for Bayesian Inference*, Second edition, Chapman & Hall, London.

Geweke, J.F., and Singleton, K.J. (1981). Maximum likelihood confirmatory factor analysis of economic time series. *International Economic Review*, 22, 1980.

Golub, G.H., and van Loan, C.F. (1996), *Matrix Computations*, third edition, The John Hopkins University Press.

Gordon, N. J., Salmond, D. J. and Smith, A. F. M. (1993). A novel approach to non-linear and non-Gaussian Bayesian state estimation. *IEE-Proceedings F 140*, 107-113.

Hannan, E.J., and Deistler, M. (1988). *The Statistical Theory of Linear Systems*. Wiley Series in Probability and Statistics, John Wiley & Sons.

Harvey, A.C. (1989). *Forecasting, Structural Time Series and the Kalman Filter*. Cambridge University Press, Cambridge, UK.

Harvey, A.C. (2001). Testing in Unobserved Components Models. *Journal of Forecasting*, 20, 1-19.

Harvey, A.C., (2010), *Exponential Conditional Volatility Models* , working paper CWPE 1040.

Harvey, A.C., and Chung, C.H. (2000). Estimating the underlying change in unemployment in the UK. *Journal of the Royal Statistics Society*, Series A, Statistics in Society, 163, Part 3, 303-339.

Harvey, A.C., and Jäger, A. (1993). Detrending, stylized facts and the business cycle. *Journal of Applied Econometrics*, 8, 231-247.

Harvey, A.C., and Proietti, T. (2005). *Readings in Unobserved Components Models*. Advanced Texts in Econometrics. Oxford University Press, Oxford, UK.

Harvey, A.C., and Chakravarty, T. (2008). *Beta-t(E)GARCH*, working paper, CWPE 0840.

Harville, D. A. (1977) Maximum likelihood approaches to variance component estimation and to related problems, *Journal of the American Statistical Association*, 72, 320–340.

Hodrick, R., and Prescott, E.C. (1997). Postwar U.S. Business Cycle: an Empirical Investigation, *Journal of Money, Credit and Banking*, 29, 1, 1-16.

Ionides, E. L., Breto, C. and King, A. A. (2006), Inference for nonlinear dynamical systems, *Proceedings of the National Academy of Sciences* 103, 18438–18443.

Ionides, E. L, Bhadra, A., Atchade, Y. and King, A. A. (2011), Iterated filtering, *Annals of Statistics*, 39, 1776–1802.

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

Julier S.J., and Uhlmann, J.K. (1996), A General Method for Approximating Nonlinear Transformations of Probability Distributions, Robotics Research Group, Oxford University, 4, 7, 1–27.

Julier S.J., and Uhlmann, J.K. (1997), A New Extension of the Kalman Filter to Nonlinear Systems, *Proceedings of AeroSense: The 11th International Symposium on Aerospace/Defense Sensing, Simulation and Controls*.

Jungbacker, B., Koopman, S.J., and van der Wel, M., (2011), Maximum likelihood estimation for dynamic factor models with missing data, *Journal of Economic Dynamics and Control*, 35, 8, 1358–1368.

Kailath, T., Sayed, A.H., and Hassibi, B. (2000), *Linear Estimation*, Prentice Hall, Upper Saddle River, New Jersey.

Kalman, R.E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering, Transactions ASME. Series D* 82: 35-45.

Kalman, R.E., and R.S. Bucy (1961). New results in linear filtering and prediction theory, *Journal of Basic Engineering, Transactions ASME, Series D* 83: 95-108.

Kim, C.J. and C. Nelson (1999). *State-Space Models with Regime-Switching*. Cambridge MA: MIT Press.

Kitagawa, G. (1987). Non-Gaussian State-Space Modeling of Nonstationary Time Series (with discussion), *Journal of the American Statistical Association*, 82, 10321063.

Kitagawa, G. (1998). A self-organising state-space model, *Journal of the American Statistical Association*, 93, 1203-1215.

Kitagawa, G. (1996). Monte Carlo Filter and Smoother for Non-Gaussian Nonlinear State-Space Models, *Journal of Computational and Graphical Statistics*, 5, 125.

Kitagawa, G., and W Gersch (1996). *Smoothness priors analysis of time series*. Berlin: Springer-Verlag.

Koopman, S.J., and Durbin, J. (2000). Fast filtering and smoothing for multivariate state space models, *Journal of Time Series Analysis*, 21, 281–296.

Luati, A. and Proietti, T. (2010). Hyper-spherical and Elliptical Stochastic Cycles, *Journal of Time Series Analysis*, 31, 169–181.

Morley, J.C., Nelson, C.R., and Zivot, E. (2002). Why are Beveridge-Nelson and Unobserved-Component Decompositions of GDP So Different?, *Review of Economics and Statistics*, 85, 235-243.

Nelson, C.R., and Plosser, C.I. (1982). Trends and random walks in macroeconomic time series: some evidence and implications. *Journal of Monetary Economics*, 10, 139-62.

Nerlove, M., Grether, D. M., and Carvalho, J. L. (1979), *Analysis of Economic Time Series: A Synthesis*, New York: Academic Press.

Nyblom, J. (1986). Testing for deterministic linear trend in time series. *Journal of the American Statistical Association*, 81: 545-9.

Nyblom, J.(1989). Testing for the constancy of parameters over time. *Journal of the American Statistical Association*, 84, 223-30.

Nyblom, J., and Harvey, A.C. (2000). Tests of common stochastic trends, *Econometric Theory*, 16, 176-99.

Nyblom J., Mäkeläinen T. (1983). Comparison of tests for the presence of random walk coefficients in a simple linear model. *Journal of the American Statistical Association*, 78, 856864.

Ord J.K., Koehler A.B., and Snyder, R.D. (1997). Estimation and prediction for a class of Dynamic nonlinear statistical models. *Journal of the American Statistical Association*, 92, 1621-1629.

Pagan, A. (1980). Some Identification and Estimation Results for Regression Models with Stochastically Varying Coefficients *Journal of Econometrics*, 13, 341–363.

Patterson, H.D. and Thompson, R. (1971) Recovery of inter-block information when block sizes are unequal, *Biometrika*, 58, 545–554.

Pearlman, J. G. (1980). An Algorithm for the Exact Likelihood of a High-Order Autoregressive-Moving Average Process. *Biometrika*, 67: 232-233.

Pitt, M.K. and Shephard, N. (1999). Filtering via simulation: auxiliary particle filters. *Journal of the American Statistical Association*, 94, 590-599.

Poyiadjis, G and Doucet, A and Singh, SS (2011) Particle approximations of the score and observed information matrix in state space models with application to parameter estimation. *Biometrika*, 98, 65–80.

Primiceri, G.E. (2005), Time Varying Structural Vector Autoregressions and Monetary Policy, *The Review of Economic Studies*, 72, 821–852

Proietti T. (1999). Characterising Business Cycle Asymmetries by Smooth Transition Structural Time Series Models. *Studies in Nonlinear Dynamics and Econometrics*, 3, 141–156.

Proietti T. (2006), Trend–Cycle Decompositions with Correlated Components. *Econometric Reviews*, 25, 61-84

Richard, J.F. and Zhang, W. (2007), Efficient high-dimensional importance sampling, *Journal of Econometrics* 127, , 1385–1411.

Rosenberg, B. (1973). Random coefficient models: the analysis of a cross-section of time series by stochastically convergent parameter regression. *Annals of Economic and Social Measurement*, 2, 399-428.

Rubin, D. B. (1987). A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when the fraction of missing information is modest: the SIR algorithm. Discussion of Tanner and Wong (1987). *Journal of the American Statistical Association*, 82, 543-546.

Sargent, T.J. (1989), Two Models of Measurements and the Investment Accelerator, *Journal of Political Economy*, 97, 2, 251–287,

Sargent, T.J., and C.A. Sims (1977), Business Cycle Modeling Without Pretending to Have Too Much A-Priori Economic Theory, in *New Methods in Business Cycle Research*, ed. by C. Sims *et al.*, Minneapolis: Federal Reserve Bank of Minneapolis.

Smets, F. and Wouters, R. (2003), An Estimated Dynamic Stochastic General Equilibrium Model of the Euro Area, *Journal of the European Economic Association*,1, 5, 1123–1175.

Shephard, N. (2005). *Stochastic Volatility: Selected Readings*. Advanced Texts in Econometrics. Oxford University Press, Oxford, UK.

Shephard, N. and Pitt, M. K. (1997). Likelihood analysis of non-Gaussian measurement time series. *Biometrika*, 84, 653-667.

Shumway, R.H., and Stoffer, D.S. (1982). An approach to time series smoothing and forecasting using the EM algorithm. *Journal of Time Series Analysis*, 3, 253-264.

Stock, J.H., and M.W. Watson (1989), *New Indexes of Coincident and Leading Economic Indicators*, NBER Macroeconomics Annual 1989, 351-393.

Stock, J.H., and Watson M.W. (1991). A probability model of the coincident economic indicators. In *Leading Economic Indicators*, Lahiri K, Moore GH (eds), Cambridge University Press, New York.

Stock, J.H. and Watson, M.W. (2007), Why Has U.S. Inflation Become Harder to Forecast?, *Journal of Money, Credit and Banking*, 39(1), 3-33.

Tunnicliffe-Wilson, G. (1989). On the use of marginal likelihood in time series model estimation. *Journal of the Royal Statistical Society*, Series B, 51, 15-27.

van der Merwe, R., Doucet, A., De Freitas, N., Wan, E. (2000), The Unscented Particle Filter, *Advances in Neural Information Processing Systems*, 13, 584-590.

Watson, M.W. (1986). Univariate detrending methods with stochastic trends. *Journal of Monetary Economics,* 18, 49-75.

West, M. and P.J.Harrison (1989). *Bayesian Forecasting and Dynamic Models*. New York: Springer-Verlag.

Winschel, W. and Krätzig, M. (2010), Solving, Estimating, and Selecting Nonlinear Dynamic Models without the Curse of Dimensionality, *Econometrica*, 39, 1, 3–33.