Estimating an AR Model with Exogenous Driver

Yves Grenier

Estimating an AR Model with Exogenous Driver

Yves Grenier

October 2013

Institut Mines-Télécom ; Télécom ParisTech ; CNRS LTCI

Abstract

In this paper, we introduce an autoregressive model which has an evolution that is driven by an exogenous pilot signal. This model shares some properties with TAR (Threshold Auto Regressive) models and STAR (Smooth Transition Auto Regressive) models.

This text defines the model, it presents an estimator for this model, and an estimator for the variance of the innovation, which is not constant in this model. An exact computation of the likelihood of this driven autoregressive model is then presented. Two appendices present a state-space realization of this model and the expression of a Kalman filter for such a model.

Résumé

Dans cet article, nous introduisons un modèle auto régressif dont l'évolution est pilotée par un signal exogène. Ce modèle présente des analogies avec les modèles autorégressifs à seuil (TAR, Threshold AutoRegression) ainsi que les modèles STAR (Smooth Transition Autoregressive).

Le texte présente le modèle, puis un estimateur pour ce modèle ainsi que pour la variance de son innovation, qui n'est pas constante dans ce modèle. Un calcul exact de la vraisemblance du modèle autorégressif sera ensuite présenté. Deux annexes montreront la réalisation de ce type de modèle autorégressif sous forme de modèle d'état ainsi que l'expression du filtre de Kalman reposant sur ce modèle.

1 Autoregressive model with exogenous pilot

1.1 Model definition

Let us consider a signal y_t, $t \in [0, T]$, and a representation of this signal as a non stationary autoregressive process with coefficients $a_i(t)$, its innovation being denoted e_t:

$$y_t + a_1(t-1)y_{t-1} + \cdots + a_p(t-p)y_{t-p} = e_t$$

We assume that the innovation is a non-stationary white noise, which is centered ($\mathbb{E}(e_t) = 0$), with variance σ_t^2:

$$\mathbb{E}(e_t e_{t'}) = \delta_{t,t'}\sigma_t^2$$

We also assume that we observe a second signal x_t, that is taken as deterministic; this signal drives the coefficients $a_i(t)$ of the autoregressive model of the first signal y_t:

$$a_i(t) = g_i(x_t)$$

where $g_i(x)$, $i \in [1, p]$ are nonlinear functions of x; for instance, we could assume that these functions have a simple parametric representation only depending on a set of functions $f_m(.)$ such as:

$$a_i(t) = \sum_{m=0}^{M} a_{im} f_m(x_t)$$

Most often, we will take $f_m(x) = (x/\alpha)^m$ where α will be a normalization coefficient applied to the signal x_t. The autoregressive model which is written:

$$y_t + \sum_{i=1}^{p} a_i(t-i)y_{t-i} = e_t$$

may be rewritten (if we omit the normalization parameter α) as follows:

$$y_t + \sum_{i=1}^{p}\sum_{m=0}^{M} a_{im} x(t-i)^m y_{t-i} = e_t \tag{1}$$

Let us define a vector signal Y_t build from the signal y_t (which we want to model) and from the signal x_t (which drives the model):

$$Y_t = \begin{bmatrix} y_t \\ x_t y_t \\ x_t^2 y_t \\ \vdots \\ x_t^M y_t \end{bmatrix}$$

Given the expression of the autoregressive model and this definition of Y_t, the model of y_t driven by x_t is rewritten as a linear regression of signal y_t on the p past samples of the non stationary vector signal Y_t:

$$y_t + \sum_{i=1}^{p} A_i Y_{t-i} = e_t$$

1.2 Similar models

This model is inspired by the time-varying autoregressive model described in [1]. In this reference, the coefficients of the model $a_i(t)$ were represented through combinations of functions of time. In the current model, they are represented through combinations of functions of the exogenous variable x_t.

The idea of expressing the coefficients in terms of an exogenous variable appears in the definition of Threshold AutoRegression by Tong and Lim [2] which, after adaptation to the notations of the current paper, may be written as:

$$y_t + a_1^{(j)} y_{t-1} + \cdots + a_p^{(j)} y_{t-p} = \varepsilon_t,$$

conditional on $y_{t-d} \in R_j$ where the R_j form a partition of the real line, defined by $\{r_0, r_1, \ldots, r_m\}$, a linearly ordered subset of the real numbers ($r_0 = -\infty$, $r_m = \infty$, $r_j < r_{j+1}$); the a_i form a discrete set of

m vectors. In the discussion of this paper, Priestley proposes to write a similar model, where the a_i form a continuous set:

$$y_t + a_1(Y_t)y_{t-1} + \cdots + a_p(Y_t)y_{t-p} = \varepsilon_t,$$

with

$$Y_t = [y_t y_{t-1} \cdots y_{t-p}]^T, \;\; a_i(Y_t) = a_i^{(0)} + Y_{t-1}^T \beta_i.$$

Many variants of the above models have been created (see [3]). The Smooth Threshold AutoRegressive model (STAR) [1] presented in [4] may be written:

$$y_t + a_1 y_{t-1} + \cdots + a_p y_{t-p} + (b_1 y_{t-1} + \cdots + b_p y_{t-p})\, F\left(\gamma(y_{t-d} - c)\right) = \varepsilon_t.$$

In this model, the autoregressive coefficients appear a a smooth transition between the set of $\{a_i\}$ and the set of $\{b_i\}$, with a weight given by some function of the threshold variable y_{t-d}. The function $F(.)$ is chosen as the distribution function of the standard normal distribution in [4]. Other variants (see [5]) include the logistic STAR model and the exponential STAR model: we obtain them by rewriting the model as:

$$y_t + (a_1 y_{t-1} + \cdots + a_p y_{t-p})\,(1 - G(z_t, \gamma, c)) + (b_1 y_{t-1} + \cdots + b_p y_{t-p})\, G(z_t, \gamma, c) = \varepsilon_t,$$

where $\gamma > 0$, z_t is the exogenous variable and c is a centering value. In the logistic STAR, the function is taken as:

$$G(z_t, \gamma, c) = \frac{1}{1 + \exp(-\gamma(z_t - c))}$$

In the exponential STAR, the function is taken as:

$$G(z_t, \gamma, c) = 1 - \exp(-\gamma(z_t - c)^2).$$

Multiple regime STAR models are defined in [5] as:

$$y_t + \varphi_1^T Y_t + (\varphi_2 - \varphi_1)^T Y_t G_1(z_t) + (\varphi_3 - \varphi_2)^T Y_t G_2(z_t) + \cdots + (\varphi_m - \varphi_{m-1})^T Y_t G_m(z_t) = \varepsilon_t,$$

where φ_i is the vector containing the autoregressive coefficients of the i-th model, and $G_i(z_t)$ is a shortcut for $G_i(z_t, \gamma, c)$.

A third-order Taylor expansion of these functions in [6] leads to an equation (in the case where the exogenous variable z_t is equal to $a^T Y_t$) with the following expression:

$$y_t + \beta^T Y_t + \sum_{i=1}^{p}\sum_{j=1}^{p} \xi_{ij} y_{t-i} y_{t-j} + \sum_{i=1}^{p}\sum_{j=1}^{p} \psi_{ij} y_{t-i} y_{t-j}^2 + \sum_{i=1}^{p}\sum_{j=1}^{p} \kappa_{ij} y_{t-i} y_{t-j}^3 = \varepsilon_t. \tag{2}$$

This last model is a special case (when $z_t = a^T Y_t$ and $M = 3$) of the model that we introduced in equation (1). The expansion was introduced in [6] for the purpose of designing a statistical test of linearity and was not intended for estimation. [2]

2 Estimation of the autoregressive coefficients

2.1 Normal (Yule-Walker) equations

To estimate the autoregressive coefficients, we will assume in a first approach that the innovation variance σ_t^2 is constant ($\sigma_t^2 = \sigma^2$). We can then estimate the A_i coefficients in a least squares sense by minimizing $\mathbb{E}(e_t^2)$, or by maximizing an approximated likelihood. The likelihood of the model (where θ is the vector of A_i coefficients) is written as:

$$L(\theta) = p(y_0, y_1, \cdots y_T | \theta).$$

We approximate it by:

$$L(\theta) \simeq p(y_p, y_1, \cdots y_T | \theta, y_0, \cdots y_{p-1}).$$

Using the fact that the innovation is a white noise, and assuming that this innovation is Gaussian:

1. The STAR model is now rather called Smooth Transition AutoRegressive.

2. It is clear that a model defined as in equation (1), as well as a model described by equation (2) is no longer stable when z_t grows. We consider this as a minor problem when using this model on a given signal which remains bounded.

$$\begin{aligned} L(\theta) &= \prod_{t=p}^{T} p(\varepsilon_t|\theta, y_0, \cdots y_{p-1}) \\ &= \prod_{t=p}^{T} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{e_t^2}{\sigma^2}} \end{aligned}$$

$$\log L(\theta) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2}\sum_{t=p}^{T}\frac{e_t^2}{\sigma^2}.$$

Maximizing the (approximate) log likelihood is equivalent to minimizing a least squares criterion:

$$J(\theta) = \sum_{t=p}^{T} e_t^2$$

If we introduce the value of the innovation in this criterion:

$$e_t = y_t + \sum_{i=1}^{p} A_i Y_{t-i} = y_t + \theta^T \begin{bmatrix} Y_{t-1} \\ Y_{t-2} \\ \vdots \\ Y_{t-p} \end{bmatrix}$$

the minimization leads to Yule-Walker equations:

$$\sum_{t=p}^{T} \begin{bmatrix} Y_{t-1} \\ Y_{t-2} \\ \vdots \\ Y_{t-p} \end{bmatrix} \left[Y_{t-1}^T Y_{t-2}^T \cdots Y_{t-p}^T\right] \theta = \sum_{t=p}^{T} \begin{bmatrix} Y_{t-1} \\ Y_{t-2} \\ \vdots \\ Y_{t-p} \end{bmatrix} y_t.$$

2.2 Weighted estimation of the autoregressive coefficients

If the variance of the innovation is no longer constant, we can still estimate the autoregressive coefficients, conditionally to σ_t^2 (assuming it is known). The criterion being:

$$J(\theta) = \sum_{t=p}^{T} \frac{e_t^2}{\sigma_t^2},$$

the Yule-Walker equations will be written:

$$\sum_{t=p}^{T} \frac{1}{\sigma_t^2} \begin{bmatrix} Y_{t-1} \\ Y_{t-2} \\ \vdots \\ Y_{t-p} \end{bmatrix} \left[Y_{t-1}^T Y_{t-2}^T \cdots Y_{t-p}^T\right] \theta = \sum_{t=p}^{T} \frac{1}{\sigma_t^2} \begin{bmatrix} Y_{t-1} \\ Y_{t-2} \\ \vdots \\ Y_{t-p} \end{bmatrix} y_t.$$

Through these equations, we may estimate the autoregressive coefficients as soon as the values of the variances σ_t^2 are known. We will see in the next section that it is possible to estimate these values through a model which is also driven by the exogenous signal x_t. This estimation will assume that the autoregressive coefficients are already know.

This allows us to consider an iterative estimator that will alternate two phases:

1. estimate the autoregressive coefficients, given the model of the variance σ_t^2,
2. estimate the model of the variance, given the autoregressive coefficients.

3 Estimation of the innovation variance

3.1 Model for the innovation variance

Let e_t be a centered, Gaussian, non-stationary white noise; we want to estimate its variance σ_t^2. In order to build an estimate on an interval $[0, T]$, we will assume a model for the variance that will be:

$$\sigma_t = \exp\left(\sum_{m=0}^{M} b_m f_m(t)\right) \quad \leftrightarrow \quad \log(\sigma_t) = \sum_{m=0}^{M} b_m f_m(t)$$

where $\{f_m(t),\, m \in [0, M]\}$ is a set of known functions of time. The unknown parameters in the model are the b_m coefficients, and we will group these parameters in a vector θ:

$$\theta = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_M \end{bmatrix}.$$

3.2 Estimation of the variance

In a Gaussian framework, we can derive a maximum likelihood estimation. The likelihood is:

$$p(e_1, \cdots, e_T|\theta) = \prod_{t=0}^{T} \frac{1}{\sqrt{2\pi\sigma_t^2}} e^{-\frac{1}{2}\frac{e_t^2}{\sigma_t^2}}.$$

The log-likelihood is:

$$L = \log(p(e_1, \cdots, e_T|\theta)) = \sum_{t=0}^{T} \left[-\frac{1}{2}\frac{e_t^2}{\sigma_t^2} - \log(\sigma_t) \right] - \frac{T}{2}\log(2\pi).$$

By taking the expression of σ_t into account, the log-likelihood becomes:

$$L = \sum_{t=0}^{T} \left[-\frac{1}{2} e_t^2 e^{-2\sum_{m=0}^{M} b_m f_m(t)} - \sum_{m=0}^{M} b_m f_m(t) \right] - \frac{T}{2}\log(2\pi),$$

or:

$$L = -\frac{1}{2}\sum_{t=0}^{T} e_t^2 \prod_{m=0}^{M} e^{-2b_m f_m(t)} - \sum_{t=0}^{T}\sum_{m=0}^{M} b_m f_m(t) - \frac{T}{2}\log(2\pi).$$

The gradient of the log-likelihood can be computed without any difficulty:

$$\frac{\partial L}{\partial b_m} = \sum_{t=0}^{T} e_t^2 f_m(t) \prod_{q=0}^{M} e^{-2b_q f_q(t)} - \sum_{t=0}^{T} f_m(t).$$

Writing that this gradient is zero does not permit to derive a simple estimate. But since the computation of the Hessian is not complicated, a Newton-Raphson technique can be used to maximize the log-likelihood:

$$\frac{\partial^2 L}{\partial b_m \partial b_{m'}} = -2\sum_{t=0}^{T} e_t^2 f_m(t) f_{m'}(t) \prod_{q=0}^{M} e^{-2b_q f_q(t)}.$$

3.3 Initial estimation

Using a Newton-Raphson procedure for the maximization of the likelihood can be very efficient, provided that the procedure starts from a good guess of the solution. A simple manner to obtain this initial guess is to perform a regression of the standard deviation of a set of slices from the innovation signal, over the basis of functions $f_m(t)$ taken at instants corresponding to the slices. The idea would be to split the set of samples $\{\varepsilon_t;\, t \in [0,T]\}$ into K subsets S_1, ..., S_K, associate one instant τ_k, $k \in [1,K]$ to each subset, evaluate the variance of the samples in each subset:

$$\sigma_k^2 = \sum_{t \in S_k} e_t^2,$$

and perform a regression of the values of the logarithm of the standard deviations over the basis:

$$\begin{bmatrix} f_0(\tau_1) & f_1(\tau_1) & \cdots & f_M(\tau_1) \\ f_0(\tau_2) & f_1(\tau_2) & \cdots & f_M(\tau_2) \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ f_0(\tau_K) & f_1(\tau_K) & \cdots & f_M(\tau_K) \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_M \end{bmatrix} = \begin{bmatrix} \frac{1}{2}\log\sigma_1^2 \\ \frac{1}{2}\log\sigma_2^2 \\ \vdots \\ \vdots \\ \frac{1}{2}\log\sigma_K^2 \end{bmatrix}.$$

Such an initial guess is satisfactory as soon as K is more than 3 or 4 times $M+1$, the number of unknowns.

How can we choose the subsets S_1, ..., S_K? It would be wrong to split the time interval $[0,T]$ into K successive intervals with equal length given by T/K. Considering the fact that the exogenous signal x_t drives

the model, a better choice would be to select the slices from the amplitudes of $\{x_t,\, t \in [0,T]\}$. Let us sort the values of x_t in increasing order. This defines a new set of indexes $t' = o(t)$, such that:

$$\forall t \in [0, T-1] : x_{o(t)} \leq x_{o(t+1)}.$$

With this new set of indexes, we define the slice S_k by:

$$S_k = \left\{ t : \frac{kT}{K} \leq o(t) < \frac{(k+1)T}{K} \right\},$$

and the representative instant τ_k by:

$$\tau_k : o(\tau_k) = \frac{(k+\frac{1}{2})T}{K}.$$

This way to split the values of the driving signal into K slices is clearly inspired by the definition of the Threshold AutoRegressive (TAR) model.

4 Likelihood of the model

In this section, we describe an exact computation of the likelihood of the model. For completeness, we not only assume an autoregressive model but also an autoregressive moving-average model.

4.1 State equations for the ARMA model

An ARMA model written as:

$$y_t + \sum_{i=1}^{n} a_i(t-i) y_{t-i} = \sum_{j=0}^{n} b_j(t-j) e_{t-j}$$

has a state-space representation (see appendix A):

$$X_{t+1} = \begin{bmatrix} -a_1(t) & 1 & 0 & \cdots & 0 \\ -a_2(t) & 0 & 1 & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ -a_{n-1}(t) & 0 & \cdots & 0 & 1 \\ -a_n(t) & 0 & \cdots & \cdots & 0 \end{bmatrix} X_t + \begin{bmatrix} \beta_1(t) \\ \beta_2(t) \\ \vdots \\ \beta_{n-1}(t) \\ \beta_n(t) \end{bmatrix} e_t \tag{3}$$

$$y_t = \begin{bmatrix} 1 & 0 & \cdots & \cdots & 0 \end{bmatrix} x_t + b_0(t) e_t \tag{4}$$

where:

$$\beta_i(t) = b_i(t) - a_i(t) b_0(t)$$

and

$$\mathbb{E}(e_t^2) = 1.$$

This state-space representation is a modified version of the canonical observable representation given in [7, page 374]. In order to compute the likelihood of the model, we will apply a Kalman filter based upon this representation, to compute the innovation and deduce from it the likelihood. The representation in [7, page 374] does not easily allow the use of the Kalman filter.

4.2 State estimation through Kalman filter

We will use the formulation of the Kalman filter given by [8, pp 105 et seq.] and rewrite equations (3) and (4), introducing obvious notations F_t, G_t and H:

$$X_{t+1} = F_t X_t + G_t e_t \tag{5}$$

$$y_t = H^T X_t + b_0(t) e_t \tag{6}$$

With these notations, the filter is summarized by (see appendix B where $w_t = e_t$, $v_t = b_0(t) e_t$ and the time index on H_t may be dropped here):

$$
\begin{aligned}
\varepsilon_t &= y_t - H^T X_{t/t-1} \\
\Gamma_t &= H^T \Sigma_{t/t-1} H + b_0(t)^2 \\
K_t &= \left(F_t \Sigma_{t/t-1} H + G_t b_0(t)\right) \Gamma_t^{-1} \\
X_{t+1/t} &= F_t X_{t/t-1} + K_t \varepsilon_t \\
\Sigma_{t+1/t} &= \left(F_t - K_t H^T\right) \Sigma_{t/t-1} \left(F_t - K_t H^T\right)^T + [G_t, -K_t] \begin{bmatrix} 1 & b_0(t) \\ b_0(t) & b_0(t)^2 \end{bmatrix} \begin{bmatrix} G_t^T \\ -K_t^T \end{bmatrix} \\
&= \left(F_t - K_t H^T\right) \Sigma_{t/t-1} \left(F_t - K_t H^T\right)^T + \left(G_t - b_0(t) K_t\right) \left(G_t - b_0(t) K_t\right)^T .
\end{aligned}
$$

where ε_t is the innovation, $\Gamma_t = \mathbb{E}(\varepsilon_t^2)$ is the variance of the innovation, K_t is the Kalman gain, $X_{t+1/t}$ is the prediction of the state X_t at time $t+1$, given the observations up to t, $\Sigma_{t+1/t}$ is the variance of the state prediction error: $\Sigma_{t+1/t} = \mathbb{E}\left((X_{t+1/t} - X_t)(X_{t+1/t} - X_t)^T\right)$.

The Kalman filter has initial conditions $X_{0/-1}$ and $\Sigma_{0/-1}$. We will use $X_{0/-1} = 0$. Since we cannot make any assumption concerning the state, a realistic way to initialize $\Sigma_{0/-1}$ is to make this matrix proportional to the identity matrix $\Sigma_{0/-1} = \lambda \mathbb{I}$ where λ is a large value, for instance $\lambda = 1000 \times \mathbb{E}(y_t^2)$.

4.3 Estimation of the log likelihood

The likelihood is

$$L(\theta) = p(y_0, y_1, \cdots y_T|\theta) = p(\varepsilon_0, \varepsilon_1, \cdots \varepsilon_T|\theta).$$

Given the orthogonality of the sequence $\{\varepsilon_t, t \in [0, T]\}$, under the assumption that the signals are Gaussian, this is:

$$L(\theta) = \prod_{t=0}^{T} p(\varepsilon_t|\theta) = \prod_{t=0}^{T} \frac{1}{\sqrt{2\pi\Gamma_t}} e^{-\frac{1}{2}\frac{\varepsilon_t^2}{\Gamma_t}}.$$

The log likelihood is therefore:

$$\log L(\theta) = -\frac{T}{2} \log 2\pi - \frac{1}{2} \sum_{t=0}^{T} \left(\frac{\varepsilon_t^2}{\Gamma_t} + \log \Gamma_t \right)$$

References

[1] Y.Grenier. Time-dependent ARMA Modeling of Nonstationary Signals. *IEEE Transactions on ASSP*, 31(4):899–911, 1983.

[2] H. Tong and K.S. Lim. Threshold Autoregression, Limit Cycles and Cyclical Data. *J. Roy. Statist. Soc., Series B*, 42(3):245–292, 1980.

[3] H. Tong. Threshold Models in Time Series Analysis - 30 Years on. Technical Report 471, University of Hong-Kong, Department of Statistics and Actuarial Science, April 2010.

[4] K. S. Chan and H. Tong. On Estimating Thresholds in Autoregressive Models. *Journal of Time Series Analysis*, 7(3):179–190, 1986.

[5] D. Van Dijk, T. Teräsvirta, and P.H. Franses. Smooth Transition Autoregressive Models - a Survey of Recent Developments. *Econometrics Reviews*, (21):1–47, 2002.

[6] R. Luukkonen, P. Saikkonen, and T. Teräsvirta. Testing Linearity against Smooth Transition Autoregressive Models. *Biometrika*, 75(3):pp. 491–499, 1988.

[7] Y. Grenier. Parametric Time-Frequency Representation. In R. Stora J.L. Lacoume, T.S. Durrani, editor, *Signal Processing*, Les Houches 1985, Session XLV, ISBN 0 444 87058 X, pages 339–397. North-Holland, 1987.

[8] B.D.O.Anderson and J.B.Moore. *Optimal Filtering*. Prentice Hall, 1979.

A State-space realization of a non-stationary ARMA model

We consider a signal y_t that is described by the following ARMA model:

$$y_t + \sum_{i=1}^{n} a_i(t-i) y_{t-i} = \sum_{j=0}^{n} b_j(t-j) e_{t-j} \tag{7}$$

The purpose of this section is to show that the model in equation (7) has a state-space representation that can be written as:

$$X_{t+1} = \begin{bmatrix} -a_1(t) & 1 & 0 & \cdots & 0 \\ -a_2(t) & 0 & 1 & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ -a_{n-1}(t) & 0 & \cdots & 0 & 1 \\ -a_n(t) & 0 & \cdots & \cdots & 0 \end{bmatrix} X_t + \begin{bmatrix} \beta_1(t) \\ \beta_2(t) \\ \vdots \\ \beta_{n-1}(t) \\ \beta_n(t) \end{bmatrix} e_t, \tag{8}$$

$$y_t = \begin{bmatrix} 1 & 0 & \cdots & \cdots & 0 \end{bmatrix} X_t + b_0(t) e_t. \tag{9}$$

A.1 Expression of the state vector

The k-th component of the state vector X_t will be denoted $x_t(k)$. Let us first compute the values of the components $x_k(t)$. For $k = 1$, we may use the observation equation (9) which gives:

$$x_t(1) = y_t - b_0(t) e_t. \tag{10}$$

The first row of matrix equation (3) links $x_{t+1}(1)$ to $x_t(2)$:

$$x_{t+1}(1) = -a_1(t) x_t(1) + x_t(2) + \beta_1(t) e_t. \tag{11}$$

We can then obtain the expression of $x_t(2)$ by mixing equations (10) and (11):

$$\begin{aligned} x_t(2) &= x_{t+1}(1) + a_1(t) x_t(1) - \beta_1(t) e_t, \\ &= y_{t+1} - b_0(t+1) e_{t+1} + a_1(t)(y_t - b_0(t) e_t) - \beta_1(t) e_t \\ &= y_{t+1} + a_1(t) y_t - b_0(t+1) e_{t+1} - (a_1(t) b_0(t) e_t + \beta_1(t))\, e_t. \end{aligned}$$

The same procedure leads to the successive components, and the expression of $x_t(k)$ is:

$$\begin{aligned} x_t(k) = &\sum_{i=0}^{k-1} a_i(t+k-1-i) y_{t+k-1-i} \\ &- b_0(t+k-1) e_{t+k-1} \\ &- \sum_{i=1}^{k-1} \left(a_i(t+k-1-i) b_0(t+k-1-i) + \beta_i(t+k-1-i)\right) e_{t+k-1-i}, \end{aligned} \tag{12}$$

where we assume that $a_0(t) = 1$, $\forall t$.

Proof: Equation is true for $k = 1$ (equation (12) coincides with equation (10)). It suffices to prove that if equation (12) is true for k, it is also true for $k+1$.

For $k < n$, the k-th row of matrix equation (3) is written as:

$$x_{t+1}(k) = -a_k(t) x_t(1) + x_t(k+1) + \beta_k(t) e_t.$$

Therefore:

$$
\begin{aligned}
x_t(k+1) &= x_{t+1}(k) + a_k(t)x_t(1) - \beta_k(t)e_t \\
&= \sum_{i=0}^{k-1} a_i(t+k-i)y_{t+k-i} + a_k(t)y_t \\
&\quad -b_0(t+k)e_{t+k} \\
&\quad -\sum_{i=1}^{k-1} \left(a_i(t+k-i)b_0(t+k-i) + b_i(t+k-i)\right) e_{t+k-i} \\
&\quad -\left(a_k(t)b_0(t) + \beta_k(t)\right) e_t
\end{aligned}
$$

which is identical to equation (12) for $k+1$.

A.2 Equivalence with the ARMA model

It remains to show that the state-space equations (3) and (9) are a realization of the ARMA model in (7). We will use the n-th row of equation (3):

$$x_{t+1}(n) = -a_n(t)x_t(1) + \beta_n(t)e_t.$$

Using equation (10) leads to:

$$
\begin{aligned}
x_{t+1}(n) &= -a_n(t)\left(y_t - b_0(t)e_t\right) + \beta_n(t)e_t \\
&= -a_n(t)y_t + \left(a_n(t)b_0(t) + \beta_n(t)\right) e_t. \qquad (13)
\end{aligned}
$$

Applying equation (12) for $k=n$ at time $t+1$ leads to:

$$
\begin{aligned}
x_{t+1}(n) &= \sum_{i=0}^{n-1} a_i(t+n-i)y_{t+n-i} \\
&\quad -b_0(t+n)e_{t+n} \\
&\quad -\sum_{i=1}^{n-1} \left(a_i(t+n-i)b_0(t+n-i) + \beta_i(t+n-i)\right) e_{t+n-i}, \qquad (14)
\end{aligned}
$$

Finally, equating the two expressions of $x_{t+1}(n)$ in equations (13) and (14) gives:

$$
\begin{aligned}
\sum_{i=0}^{n} a_i(t+n-i)y_{t+n-i} &= \sum_{i=1}^{n} \left(a_i(t+n-i)b_0(t+n-i) + \beta_i(t+n-i)\right) e_{t+n-i} \\
&\quad +b_0(t+n)e_{t+n},
\end{aligned}
$$

and replacing $t+n$ by t gives:

$$\sum_{i=0}^{n} a_i(t-i)y_{t-i} = b_0(t)e_t + \sum_{i=1}^{n} \left(a_i(t-i)b_0(t-i) + \beta_i(t-i)\right) e_{t-i}$$

which coincides with equation (7) provided that:

$$b_i(t) = a_i(t)b_0(t) + \beta_i(t) \;\; \forall i \in [1,n],$$

or

$$\beta_i(t) = b_i(t) - a_i(t)b_0(t) + \;\; \forall i \in [1,n].$$

B Kalman filter with correlated state and observation noises

In this section, we recall the expression of the Kalman filter as it is given in [8, section 8.4]. In this reference, contrary to most textbooks, the state noise and the observation noise are not assumed to be uncorrelated. Apart from that, their derivation also takes into account a known input sequence, but we shall ignore such an input in our presentation. The state equations are:

$$\begin{aligned} X_{t+1} &= F_t X_t + G_t w_t & (15)\\ y_t &= H_t^T X_t + v_t & (16) \end{aligned}$$

with correlated noises w_t and v_t:

$$\mathbb{E}\left(\begin{bmatrix} w_t \\ v_t \end{bmatrix} \begin{bmatrix} w_{t'}^T & v_{t'}^T \end{bmatrix} \right) = \begin{bmatrix} Q_t & S_t \\ S_t^T & R_t \end{bmatrix} \delta_{t,t'}.$$

B.1 Prediction and Kalman gain

A recursive equation is sought for the state prediction $\hat{X}_{t/t-1}$:

$$\hat{X}_{t/t-1} = \mathbb{E}(X_t | y_0,\, y_1,\, \cdots y_{t-1}) = \mathbb{E}(X_t | \tilde{y}_0,\, \tilde{y}_1,\, \cdots \tilde{y}_{t-1})$$

where $\{\tilde{y}_t\}$ is the innovation sequence:

$$\tilde{y}_t = y_t - \mathbb{E}(y_t | y_0,\, y_1,\, \cdots y_{t-1}).$$

Define the prediction error $\tilde{X}_{t/t-1} = X_t - \hat{X}_{t/t-1}$ and its covariance $\Sigma_{t/t-1} = \mathbb{E}(\tilde{X}_{t/t-1}\tilde{X}_{t/t-1}^T)$. We use the independence of the innovations (and the fact that the state is centered, $\mathbb{E}(X_{t+1}) = 0$, since the observations and noises are centered):

$$\mathbb{E}(X_{t+1} | \tilde{y}_0,\, \tilde{y}_1,\, \cdots \tilde{y}_t) = \mathbb{E}(X_{t+1} | \tilde{y}_t) + \mathbb{E}(X_{t+1} | \tilde{y}_0,\, \tilde{y}_1,\, \cdots \tilde{y}_{t-1}). \quad (17)$$

The first term can be computed through the projection theorem:

$$\mathbb{E}(X_{t+1} | \tilde{y}_t) = \mathbb{E}(X_{t+1}\tilde{y}_t^T)\mathbb{E}(\tilde{y}_t\tilde{y}_t^T)^{-1}\tilde{y}_t. \quad (18)$$

The covariances are estimated as:

$$\begin{aligned} \mathbb{E}(X_{t+1}\tilde{y}_t^T) &= \mathbb{E}\left((F_t X_t + G_t w_t)(H_t^T \tilde{X}_t + v_t)^T \right) \\ &= \mathbb{E}\left(F_t X_t \tilde{X}_t^T H_t \right) + G_t S_t \\ &= F_t \left(\mathbb{E}(\hat{X}_{t/t-1}\tilde{X}_t^T) + \mathbb{E}(\tilde{X}_t \tilde{X}_t^T) \right) H_t + G_t S_t \\ &= F_t \Sigma_{t/t-1} H_t + G_t S_t \end{aligned}$$

and (given that $\tilde{X}_t$ and v_t are independent):

$$\begin{aligned} \mathbb{E}\left(\tilde{y}_t\tilde{y}_t^T\right) &= \mathbb{E}\left((H_t^T \tilde{X}_t + v_t)(H_t^T \tilde{X}_t + v_t)^T \right) \\ &= H_t^T \Sigma_{t/t-1} H_t + R_t \end{aligned}$$

Therefore equation (18) becomes:

$$\mathbb{E}(X_{t+1} | \tilde{y}_t) = (F_t \Sigma_{t/t-1} H_t + G_t S_t)(H_t^T \Sigma_{t/t-1} H_t + R_t)^{-1} \tilde{y}_t.$$

Given that w_t is independent of $\tilde{y}_0,\, \tilde{y}_1,\, \cdots \tilde{y}_{t-1}$, the second term in equation (17) can be computed as:

$$\begin{aligned} \mathbb{E}(X_{t+1} | \tilde{y}_0,\, \tilde{y}_1,\, \cdots \tilde{y}_{t-1}) &= \mathbb{E}(F_t X_t + G_t w_t | \tilde{y}_0,\, \tilde{y}_1,\, \cdots \tilde{y}_{t-1}) \\ &= F_t \mathbb{E}(X_t | \tilde{y}_0,\, \tilde{y}_1,\, \cdots \tilde{y}_{t-1}) \\ &= F_t \hat{X}_{t/t-1} \end{aligned}$$

The recursive equation (17) can now be written as:

$$\hat{X}_{t+1/t} = F_t \hat{X}_{t/t-1} + K_t (y_t - H_t^T \hat{X}_{t/t-1})$$

where K_t is the Kalman gain defined as:

$$K_t = (F_t \Sigma_{t/t-1} H_t + G_t S_t)(H_t^T \Sigma_{t/t-1} H_t + R_t)^{-1}$$

B.2 Evolution of the covariance matrix

The evolution of the covariance matrix $\Sigma_{t/t-1}$ is deduced from the evolution of the state error $\tilde{X}_{t/t-1}$ which is now seen as:

$$\begin{aligned}
\tilde{X}_{t+1/t} &= X_{t+1} - \hat{X}_{t+1/t} \\
&= F_t X_t + G_t w_t - F_t \hat{X}_{t/t-1} - K_t(y_t - H_t^T \hat{X}_{t/t-1}) \\
&= F_t(X_t - \hat{X}_{t/t-1}) + G_t w_t - K_t(H_t^T X_t + v_t - H_t^T \hat{X}_{t/t-1}) \\
&= F_t(X_t - \hat{X}_{t/t-1}) + G_t w_t - K_t H_t^T (X_t - \hat{X}_{t/t-1}) - K_t v_t \\
&= (F_t - K_t H_t^T)\tilde{X}_{t/t-1} + G_t w_t - K_t v_t \\
&= (F_t - K_t H_t^T)\tilde{X}_{t/t-1} + [G_t \ -K_t] \left[w_t^T v_t^T\right]^T
\end{aligned}$$

Given the independence of $\tilde{X}_{t/t-1}$ and $\left[w_t^T v_t^T\right]^T$, the covariance evolves as:

$$\begin{aligned}
\Sigma_{t+1/t} &= \mathbb{E}(\tilde{X}_{t+1/t}\tilde{X}_{t+1/t}^T) \\
&= (F_t - K_t H_t^T)\mathbb{E}(\tilde{X}_{t/t-1}\tilde{X}_{t/t-1}^T)(F_t - K_t H_t^T)^T + [G_t \ -K_t] \begin{bmatrix} Q_t & S_t \\ S_t^T & R_t \end{bmatrix} \begin{bmatrix} G_t \\ -K_t \end{bmatrix} \\
&= (F_t - K_t H_t^T)\Sigma_{t/t-1}(F_t - K_t H_t^T)^T + [G_t \ -K_t] \begin{bmatrix} Q_t & S_t \\ S_t^T & R_t \end{bmatrix} \begin{bmatrix} G_t \\ -K_t \end{bmatrix}
\end{aligned}$$