11 Chapter 12: State Space Models and the Kalman Filter

11.1 Introduction

Many time-series models used in econometrics are special cases of the class of linear state space models developed by engineers to describe physical systems. The Kalman filter, an efficient recursive method for computing optimal linear forecasts in such models, can be exploited to compute the exact Gaussian likelihood function.

The linear state-space model postulates that an observed time series is a linear function of a (generally unobserved) state vector and the law of motion for the state vector is first-order vector autoregression. More precisely, let y_t be the observed variable at time t and let x_t denote the values taken at time t by a vector of p state variables. Let A and b be $p \times p$ and $p \times 1$ matrices of constants. We assume that $\{y_t\}$ is generated by

$$y_t = b' x_t + v_t \tag{11.1}$$

$$x_t = A x_{t-1} + w_t \tag{11.2}$$

where the scalar v_t and the vector w_t are mean zero, white-noise processes, uncorrelated of each other and of the initial value x_0. We denote $\sigma^2 = Q_v = E(v_t^2)$ and $R_w = E(w_t w_t')$. Eq. (11.1) is sometimes called the "measurement" equation while (11.2) is called the "transition" equation. The assumption that the autoregression is first-order is not restrictive, since higher-order systems can be handled by adding additional state variables.

In most engineering (and some economic) applications, the x's represent meaningful but imperfectly measured physical variables. Models based on the "permanent" income hypothesis are classic examples. All ARMA models for y_t can be put in state space form even though the state variables x_t have no particular economic meaning. An even richer class of (possibly nonstationary) state space models can be produced by introducing an observed exogenous forcing variable X_t into the measurement equation, by letting b, A, σ^2, and Σ depend on t, and by letting y_t be a vector.

The general model which we shall consider is of the form

$$Y_t = \Phi_t X_t + v_t \tag{11.3}$$

$$X_{t+1} = F_t X_t + w_t; \tag{11.4}$$

where $w_t \sim WN(0, R_t)$, $v_t \sim WN(0, Q_t)$, and $E w_t v_t = 0, \forall s, t \in \mathbf{Z}$; R_t, Q_t, Φ_t and F_t are matrices of the corresponding dimensions. In the model Y_t represents an observed data vector at time t, X_t is an unobserved "state" of the underlying system.

In many important special cases, the matrices R_t, Q_t, Φ_t and F_t will be assumed to be constants, i.e. are invariant in time (independent of time t).

As a conclusion, state space models are an alternative formulation of time series which have a number of the following positive features:

- All ARMA processes may be reformulated as state space models, and this allows a superior treatment of some problems associated with ARMA models, such as exact likelihoods and predictive distributions in short time series, or for a more elegant handling of missing data problems.

- Extension to nonstationary models — for example, a common model in applications of time series analysis is an ARMA with time-varying coefficients, and this situation is handled straightforwardly within a state space formulation.

- Multivariate time series — the standard state space model formulation is multivariate and so automatically suggests models for multivariate time series, which may be easier to handle than multivariate ARMA models (though, of course, such things do exist and there is an extensive theory about them; see Lutkepohl (1993), for example).

- Bayesian approach — the state space model is naturally treated from a Bayesian point of view and so allows one to take advantage of the more general and flexible approach to inference that Bayesian theory provides.

Within the state-space framework, the main problem is the estimation or prediction of the unobserved sequence of states X_t in terms of the observed data points Y_t. The Kalman Filter is a recursive algorithm, first devised a Hungarian mathematician by R.E. Kalman in 1960, designed to solve this problem. It is the centerpiece of all statistical analysis based on state space models.

11.2 ARMA Models in State Space Form

Consider the ARMA(1,1) model

$$y_t = \phi y_{t-1} + \varepsilon_t + \theta \varepsilon_{t-1}. \tag{11.5}$$

Defining $x_t = (x_{1t}, x_{2t})' = (y_t, \theta\varepsilon_t)'$, we can write $y_t = b'x_t$, where $b = (1,0)'$ and

$$\begin{bmatrix} x_{1t} \\ x_{2t} \end{bmatrix} = \begin{bmatrix} \phi & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_{1,t-1} \\ x_{2,t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_t \\ \theta\varepsilon_t \end{bmatrix}. \tag{11.6}$$

Thus the ARMA(1,1) model has a state-space representation.

Notice that state-space representation is not unique.

11.3 The Kalman Filter. The special case: the errors are normal

Denote the vector $(y_1, \ldots, y_t)$ by Y_t. The Kalman filter is a recursive algorithm for producing optimal linear forecasts of x_{t+1} and y_{t+1} from the past history Y_t, assuming that A, b, σ^2, and Σ are known. Define

$$\begin{aligned} \hat{x}_t &= E(x_t|Y_{t-1}) \\ V_t &= \mathrm{Var}(x_t|Y_{t-1}). \end{aligned} \tag{11.7}$$

If the w's and v's are normally distributed, the minimum MSE forecast of x_{t+1} and y_{t+1} at time t is $A\hat{x}_t$ and $b'\hat{x}_{t+1}$ respectively. The key fact is that under normality $\hat{x}_{t+1}$ can be calculated recursively by

$$\begin{aligned} \hat{x}_{t+1} &= A\hat{x}_t + AV_t b \frac{y_t - b'\hat{x}_t}{b'V_t b + \sigma^2} \\ V_{t+1} &= \Sigma + AV_t A' - \frac{AV_t bb'V_t A'}{b'V_t b + \sigma^2} \end{aligned} \tag{11.8}$$

starting with the appropriate initial values $(\hat{x}_1, V_1)$.

In fact, if a scalar random variable Z and a random vector X are jointly normal, then

$$\begin{aligned} E(X|Z) &= E(X) + \frac{\mathrm{Cov}(X,Z)}{\mathrm{Var}(Z)}(Z - EZ), \\ \mathrm{Var}(X|Z) &= \mathrm{Var}(X) - \frac{\mathrm{Cov}(X,Z)\mathrm{Cov}(X,Z)'}{\mathrm{Var}(Z)} \end{aligned} \tag{11.9}$$

Define the random variables $x_t^* = E(x_t|Y_t)$ and $V_t^* = \mathrm{Var}(x_t|Y_t)$. Note that x_t^* and $\hat{x}_t$ are both expectations of the same random variable x_t, the former conditioning on y_t and the latter not. Likewise V_t^* and V_t are both variances of x_t, the former conditioning on y_t and the latter not. Since, conditional on Y_{t-1}, the vector x_t and the scalar y_t are jointly normal, we can use (11.9) to calculate a relationship between x_t^* and $\hat{x}_t$ and between V_t^* and V_t.

Thus, if we pre-condition x_t and y_t on Y_{t-1} and let x_t play the role of X and y_t the role of Z, we have from (11.9)

$$\begin{aligned} x_t^* &= \hat{x}_t + V_t b \frac{y_t - b'\hat{x}_t}{b'V_t b + \sigma^2} \\ V_t^* &= V_t - \frac{V_t b b' V_t'}{b'V_t b + \sigma^2}. \end{aligned} \tag{11.10}$$

Here we use the following relationships

$$\begin{aligned} \mathrm{Cov(x_t, y_t|Y_{t-1})} &= \mathrm{Cov(x_t, b'x_t|Y_{t-1}) = V_t b}, \\ \mathrm{Var(y_t|Y_{t-1})} &= \mathrm{Var(b'x_t + u_t|Y_{t-1}) = b'V_t b} + \sigma^2, \\ E(x_t|Y_{t-1}) &= \hat{x}_t, \qquad E(y_t|Y_{t-1}) = E(b'x_t + u_t|Y_{t-1}) = b'\hat{x}_t. \end{aligned} \tag{11.11}$$

From (11.2), it follows that

$$\begin{aligned} \hat{x}_{t+1} &= Ax_t^* \\ V_{t+1} &= AV_t^*A' + \Sigma \end{aligned} \tag{11.12}$$

The "updating" equations (11.10) describe how the forecast of the state vector at time t is changed when y_t is observed. Together with the "prediction" equations (11.12), they imply the recursion (11.8).

To forecast y_{t+1} at time t, one needs only the current y_t and the previous forecast of x_t and its variance. Previous values $y_1, \ldots, y_{t-1}$ enter only through $\hat{x}_t$. Note that y_t enters linearly into the calculation of x_t^* and does not enter at all into the calculation of V_t. The forecast of y_t is a linear filter of previous y's.

In practice, one often uses mathematically convenient initial conditions and relies on the fact that, for weakly dependent processes, initial conditions do not matter very much. For more details, see A. Harvey, *Forecasting, Structural Time Series Models and the Kalman Filter* (1989).

Suppose we wish to estimate the unknown parameters of a given state-space model from the observations $y_1, \ldots, y_T$. Let $f(y_t|Y_{t-1})$ represent the conditional density of y_t, given the previous y's. From equations (11.1) and (11.2), it follows that this density has mean $b'a_t$ and variance $\sigma^2 + b'V_t b$. The joint density function for the y's can always be factored as

$$f(y_1)f(y_2|Y_1)f(y_3|Y_2)\ldots(y_T|Y_{T-1}). \tag{11.13}$$

Hence, if the y's are normal, the log likelihood function is (apart from a constant)

$$-\frac{1}{2}\sum_{t=1}^{T}\left[\ln(b'V_t b + \sigma^2) + \frac{(y_t - b'\hat{x}_t)^2}{b'V_t b + \sigma^2}\right]. \tag{11.14}$$

Thus, the same calculations that produce forecasts can be used to numerically evaluate the normal likelihood (for any given values of the parameters). Quasi maximum likelihood estimates can be obtained by iterative methods such as employed in the Newton-Raphson algorithm.

11.4 The Kalman Filter. The general case.

Assume that we observe a random process

$$y_t = \Phi_t x_t + v_t. \tag{11.15}$$

Suppose that we know the matrix Φ_t but Φ_t can evolve in time. The noise v_t is uncorrelated, zero-mean process with variance Q_t, i.e.

$$Ev_t v_s = Q(t)\delta_{t,s}$$

where δ_{ts} is the Kronecker delta ($\delta_{ts} = \ell, t = s$ and $\delta_{ts} = 0, t \neq s$). The variance Q_t can also evolve in time but does not degenerate for any fixed time t.

Our goal is to obtain the estimate $\hat{x}_{t_1}(t_0, t)$ of the random process $\{x_t\}$ at the moment t_1 using observations of $y_s, t_0 \leq s \leq t$.

We seek the estimate $\hat{x}_{t_1}(t_0, t)$ in the class of linear estimates

$$\hat{x}_{t_1}(t_0, t) = \sum_{K=t_0}^{t} h(t_1, k) y_k. \tag{11.16}$$

We require $\hat{x}_{t_1}(t_0, t)$ to minimize the cost function

$$L\left[t_0, t; t_1\right] = E|x_{t_1} - \hat{x}_{t_1}(t_0, t)|^2. \tag{11.17}$$

If so, we call $\hat{x}_{t_1}(t_0, t)$ an optimal estimate. Here $h(\cdot, \cdot)$ is a deterministic function that is called **weight or impulse function** of the filter (model) 11.16.

- If the time point t_1 belongs to $[t_0, t)$, then we have **the interpolation or smoothing problem**.
- If the time point t_1 coincides with t, then we have **the filtering problem**.
- If the time point $t_1$does not belong to $[t_0, t]$, then we have **the forecasting (extrapolation) problem**.

It is difficult to find the solution of such a problem in so general statement. In our previous lectures we restrict ourselves to stationary processes. However, the Kalman-Bucy filter can handle non-stationary cases as well. We assume that the random process $\{x_t\}$ is generated by

$$x_{t+1} = F_t x_t + w_{t+1}, \tag{11.18}$$

where F_t is the known (matrix) function of time, w_t is uncorrelated zero-mean random process ($Ew_t w_s = R(t)\delta_{ts}$) and is also uncorrelated with v_t.

Hence, the Kalman-Bucy filters can handle non-stationary random processes but these processes should be generated by linear equations.

The necessary and sufficient condition of optimality of $\hat{x}_{t_1}(t_0, t)$ is

$$E\left[\left(x_{t_1} - \hat{x}_{t_1}(t_0, t)\right) y_s^T\right] = 0, \qquad t_0 \leq s \leq t, \tag{11.19}$$

or with the account of 11.16

$$Ex_{t_1}y_s^T = \sum_{k=t_0}^{t} h(t_1,k)Ey_ky_s^T. \tag{11.20}$$

The equation 11.19 is called **the Wiener-Hopf (WH) equation**. The Wiener-Hopf equation implies that the optimal estimate $\hat{x}_{t_1}$ is a strictly orthogonal projection of x_{t_1} onto the linear span of $y_s, t_0 \leq s \leq t$.

There exists many ways to derive the Kalman-Bucy filter. We will consider a version of recurrent form of $h(\cdot,\cdot)$. Without loss of generality we consider a case of one-step ahead prediction of $t_1 = t+1$. Let us re-write the WH equation 11.20 for time points $t_1 = t$ (we then use only $y_{t_0},\dots,y_{t-1}$ to construct prediction) and $t_1 = t+1$ (we then use only $y_{t_0},\dots,y_{t-1},y_t$ to construct prediction)

$$Ex_ty_s^T = \sum_{k=t_0}^{t-1} h(t,k)Ey_ky_s^T, \qquad s = t_0,\dots,t-1 \tag{11.21}$$

$$Ex_{t+1}y_s^T = \sum_{k=t_0}^{t} h\left(t+1,k\right)Ey_ky_s^T, \qquad s = t_0,\dots,t-1,t. \tag{11.22}$$

Now subtract 11.21 from 11.22 and obtain

$$E\left(x_{t+1}-x_t\right)y_s^T = \sum_{k=t_0}^{t-1}\left(h\left(t+1,k\right)-h(t,k)\right)r_{ks} + h(t+1,t)r_{ts}, \qquad s = t_0,\dots,t-1. \tag{11.23}$$

Here $r_{ks} = Ey_ky_s^T$. (Note that here we take into account that 11.22 is valid also for $s = t_0,\dots,t-1$.)

In view of 11.18 we have for $s = t_0,\dots,t-1$

$$E\left[\left(x_{t+1}-x_t\right)y_s^T\right] = E\left[\left(F_tx_t - x_t + w_{t+1}\right)y_s^T\right] = \left(F_t - I\right)Ex_ty_s^T = \left(F_t - I\right)\sum_{k=t_0}^{t-1} h(t,k)r_{ks}$$

Taking into account 11.15, we get

$$r_{ts} = Ey_ty_s^T = E\left[\left(\Phi_tx_t + v_t\right)y_s^T\right] = \Phi_tEx_ty_s^T = \Phi_t\sum_{k=t_0}^{t-1} h(t,k)r_{ks}.$$

Thus, we can re-write 11.23 as

$$\begin{aligned} E\left[\left(x_{t+1}-x_t\right)y_s^T\right] &= \left(F_t - I\right)\sum_{k=t_0}^{t-1} h(t,k)r_{ks} \qquad (11.24)\\ &= \sum_{k=t_0}^{t-1}\left[h(t+1,k)-h(t,k)\right]r_{ks} + h(t+1,t)\Phi_t\sum_{k=t_0}^{t-1} h(t,k)r_{ks}. \end{aligned}$$

Hence, we get the equation

$$\begin{aligned} &\left(F_t - I\right)\sum_{k=t_0}^{t-1} h(t,k)r_{ks} - \sum_{k=t_0}^{t-1} h(t+1,k)r_{ks} \qquad (11.25)\\ &+\sum_{k=t_0}^{t-1} h(t,k)r_{ks} - h(t+1,t)\Phi_t\sum_{k=t_0}^{t-1} h(t,k)r_{ks}\\ &=\sum_{k=t_0}^{t-1}\left[F_th(t,k) - h(t+1,t)\Phi_th(t,k) - h(t+1,k)\right]r_{ks} = \sum_{k=t_0}^{t-1}\Delta(t,k)r_{ks} = 0. \end{aligned}$$

Now we want to show that $\Delta(t,k) = 0$. Notice that in view of 11.25, the estimate $\check{x}_t(t_0, t-1) = \sum\limits_{k=t_0}^{t-1} [h(t,k) - \Delta(t,k)]y_k$ also satisfies the Wiener-Hopf equation and thus is an optimal predictor of x_t given $y_{t_0}, \ldots, y_{t-1}$.

Hence

$$E\left|\check{x}_t\left(t_0, t-1\right) - \hat{x}_t\left(t_0, t-1\right)\right|^2 = 0,$$

or

$$E\left|\sum_{k=t_0}^{t-1} \Delta(t,k)\Phi_k x_k\right|^2 + \sum_{k=t_0}^{t-1} \Delta(t,k)^T Q(k)\Delta(t,k) = 0. \tag{11.26}$$

However, since $R(t)$ is non-negative definite, 11.26 can hold only if $\Delta(t,k) = 0$, which implies that

$$F_t h(t,k) - h(t+1,t)\Phi_t h(t,k) = h(t+1,k). \tag{11.27}$$

This is the recursive equation our filter $h(\cdot,\cdot)$ should satisfy in order to get the optimality of forecasts.

Using 11.27, we get

$$\begin{aligned}
\hat{x}_{t+1}\left(t_0, t\right) - \hat{x}_t\left(t_0, t-1\right) &= h(t+1,t)y_t + \sum_{k=t_0}^{t-1}\left[h(t+1,k) - h(t,k)\right] y_k \\
&= h(t+1,t)y_t + \sum_{k=t_0}^{t-1}\{F_t h(t,k) - h(t+1,t)\Phi_t h(t,k) - h(t,k)]\, y_k \\
&= h(t+1,t)y_t + \{F_t - h(t+1,t)\Phi_t - I\}\sum_{k=t_0}^{t-1} h(t,k)y_k \\
&= h(t+1,t)y_t + \{F_t - h(t+1,t)\Phi_t - I\}\,\hat{x}_t\left(t_0, t-1\right).
\end{aligned} \tag{11.28}$$

Now, if we denote $K_t = h(t+1,t)$ that is called the Kalman coefficient or Kalman gain, we get the Kalman filter (11.29 - 11.31).

$$\hat{x}_{t+1}(t_0, t) = F_t\hat{x}_t(t_0, t-1) + K_t\left[y_t - \Phi_t\hat{x}_t(t_0, t-1)\right]. \tag{11.29}$$

The Kalman gain K_t is constructed by

$$K_t = F_t P_t \Phi_L^T\left[Q(t) + \Phi_t P_t \Phi_t^T\right]^{-1}, \tag{11.30}$$

$$\begin{aligned}
P_t &= \mathrm{Cov}\left[\hat{x}_t\left(t_0, t-1\right) - x_t\right] \\
&= \left[F_{t-1} - K_{t-1}Q_{t-1}\right] P_{t-1}\left[F_{t-1} - K_{t-1}Q_{t-1}\right]^T + K_{t-1}Q(t-1)K_t^T + R(t).
\end{aligned} \tag{11.31}$$