8 State space models

8.1 Models with unobserved states

State space models are an alternative formulation of time series with a number of advantages for forecasting.

1. All ARMA models can be written as state space models.
2. Nonstationary models (e.g., ARMA with time varying coefficients) are also state space models.
3. Multivariate time series can be handled more easily.
4. State space models are consistent with Bayesian methods.

In general, the model consists of

$$\begin{array}{ll} \text{observed data:} & X_t = F_t S_t + v_t \\ \text{unobserved state:} & S_t = G_t S_{t-1} + w_t \\ \text{observation noise:} & v_t \sim N(0, V_t) \\ \text{state noise:} & w_t \sim N(0, W_t) \end{array}$$

where v_t, w_t are independent and F_t, G_t are known matrices — often time dependent (e.g., because of seasonality).

Example 8.1
$X_t = S_t + v_t$, $S_t = \phi S_{t-1} + w_t$. Define $Y_t = X_t - \phi X_{t-1} = (S_t + v_t) - \phi(S_{t-1} + v_{t-1}) = w_t + v_t - \phi v_{t-1}$. The autocorrelations of $\{y_t\}$ are zero at all lags greater than 1. So $\{Y_t\}$ is MA(1) and thus $\{X_t\}$ is ARMA(1,1).

Example 8.2
The general ARMA(p,q) model $X_t = \sum_{r=1}^{p} \phi_r X_{t-r} + \sum_{s=0}^{q} \theta_s \epsilon_{t-s}$ is a state space model. We write $X_t = F_t S_t$, where

$$F_t = (\phi_1, \phi_2, \cdots, \phi_p, 1, \theta_1, \cdots, \theta_q), \quad S_t = \begin{pmatrix} X_{t-1} \\ \vdots \\ X_{t-p} \\ \epsilon_t \\ \vdots \\ \epsilon_{t-q} \end{pmatrix} \in \mathbb{R}^{p+q+1}$$

with $v_t = 0$, $V_t = 0$. $S_t = G_t S_{t-1} + w_t$.

$$S_t = \begin{pmatrix} X_{t-1} \\ X_{t-2} \\ X_{t-3} \\ \vdots \\ X_{t-p-1} \\ \epsilon_t \\ \epsilon_{t-1} \\ \epsilon_{t-2} \\ \vdots \\ \epsilon_{t-q} \end{pmatrix} = \begin{pmatrix} \phi_1 & \phi_2 & \cdots & \phi_p & 1 & \theta_1 & \theta_2 & \cdots & \theta_{q-1} & \theta_q \\ 1 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \vdots & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 1 & 0 \end{pmatrix} \begin{pmatrix} X_{t-2} \\ X_{t-3} \\ \vdots \\ X_{t-p-1} \\ \epsilon_{t-1} \\ \epsilon_{t-2} \\ \epsilon_{t-3} \\ \vdots \\ \epsilon_{t-q} \\ \epsilon_{t-q-1} \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ \vdots \\ 0 \\ \epsilon_t \\ 0 \\ \vdots \\ \vdots \\ 0 \end{pmatrix}.$$

8.2 The Kalman filter

Given observed data $X_1, \ldots, X_t$ we want to find the conditional distribution of S_t and a forecast of X_{t+1}.

Recall the following multivariate normal fact: If

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \right) \tag{8.1}$$

then

$$(Y_1, \mid Y_2) \sim N\left(\mu_1 + A_{12}A_{22}^{-1}(Y_2 - \mu_2), A_{11} - A_{12}A_{22}^{-1}A_{21}\right). \tag{8.2}$$

Conversely, if $(Y_1 \mid Y_2)$ satisfies (8.2), and $Y_2 \sim N(\mu_2, A_{22})$ then the joint distribution is as in (8.1).

Now let $\mathcal{F}_{t-1} = (X_1, \ldots, X_{t-1})$ and suppose we know that $(S_{t-1} \mid \mathcal{F}_{t-1}) \sim N\left(\hat{S}_{t-1}, P_{t-1}\right)$. Then

$$S_t = G_t S_{t-1} + w_t,$$

so

$$(S_t \mid \mathcal{F}_{t-1}) \sim N\left(G_t \hat{S}_{t-1}, G_t P_{t-1} G_t^\top + W_t\right),$$

and also $(X_t \mid S_t, \mathcal{F}_{t-1}) \sim N(F_t S_t, V_t)$.

Put $Y_1 = X_t$ and $Y_2 = S_t$. Let $R_t = G_t P_{t-1} G_t^\top + W_t$. Taking all variables conditional on $\mathcal{F}_{t-1}$ we can use the converse of the multivariate normal fact and identify

$$\mu_2 = G_t \hat{S}_{t-1} \quad \text{and} \quad A_{22} = R_t.$$

Since S_t is a random variable,

$$\mu_1 + A_{12}A_{22}^{-1}(S_t - \mu_2) = F_t S_t \Longrightarrow A_{12} = F_t R_t \quad \text{and} \quad \mu_1 = F_t \mu_2 \,.$$

Also

$$A_{11} - A_{12}A_{22}^{-1}A_{21} = V_t \Longrightarrow A_{11} = V_t + F_t R_t R_t^{-1} R_t^\top F_t^\top = V_t + F_t R_t F_t^\top \,.$$

What this says is that

$$\begin{pmatrix} X_t \\ S_t \end{pmatrix}\Bigg|_{\mathcal{F}_{t-1}} = N\left(\begin{pmatrix} F_t G_t \hat{S}_{t-1} \\ G_t \hat{S}_{t-1} \end{pmatrix}, \begin{pmatrix} V_t + F_t R_t F_t^\top & F_t R_t \\ R_t^\top F_t^\top & R_t \end{pmatrix} \right) .$$

Now apply the multivariate normal fact directly to get $(S_t \mid X_t, \mathcal{F}_{t-1}) = (S_t \mid \mathcal{F}_t) \sim N(\hat{S}_t, P_t)$, where

$$\hat{S}_t = G_t \hat{S}_{t-1} + R_t F_t^\top \left(V_t + F_t R_t F_t^\top\right)^{-1} \left(X_t - F_t G_t \hat{S}_{t-1}\right)$$
$$P_t = R_t - R_t F_t^\top \left(V_t + F_t R_t F_t^\top\right)^{-1} F_t R_t$$

These are the **Kalman filter updating equations**.

Note the form of the right hand side of the expression for $\hat{S}_t$. If contains the term $G_t \hat{S}_{t-1}$, which is simply what we would predict if it were known that $S_{t-1} = \hat{S}_{t-1}$, plus a term that depends on the observed error in forecasting X_t, i.e., $\left(X_t - F_t G_t \hat{S}_{t-1}\right)$. This is similar to the forecast updating expression for simple exponential smoothing in Section 6.4.

All we need to start updating the estimates are the initial values $\hat{S}_0$ and P_0. Three ways are commonly used.

1. Use a Bayesian prior distribution.
2. If F, G, V, W are independent of t the process is stationary. We could use the stationary distribution of S to start.
3. Choosing $S_0 = 0$, $P_0 = kI$ (k large) reflects prior ignorance.

8.3 Prediction

Suppose we want to predict the X_{T+k} given $(X_1, \dots, X_T)$. We already have

$$(X_{T+1} \mid X_1, \dots, X_T) \sim N\left(F_{T+1} G_{T+1} S_t, V_{T+1} + F_{T+1} R_{T+1} F_{T+1}^\top\right)$$

which solves the problem for the case $k = 1$. By induction we can show that

$$(S_{T+k} \mid X_1, \dots, X_T) \sim N\left(\hat{S}_{T+k}, P_{T+k}\right)$$

where

$$\begin{aligned} \hat{S}_{T,0} &= \hat{S}_T \\ P_{T,0} &= P_T \\ \hat{S}_{T,k} &= G_{T+k} \hat{S}_{T,k-1} \\ P_{T,k} &= G_{T+k} P_{T,k-1} G_{T+k}^\top + W_{T+k} \end{aligned}$$

and hence that $(X_{T+k} \mid X_1, \dots, X_T) \sim N\left(F_{T+k} \hat{S}_{T,k}, V_{T+k} + F_{T+k} P_{T,k} F_{T+k}^\top\right)$.

8.4 Parameter estimation revisited

In practice, of course, we may not know the matrices F_t, G_t, V_t, W_t. For example, in ARMA(p, q) they will depend on the parameters $\phi_1, \dots, \phi_p$, $\theta_1, \dots, \theta_q$, σ^2, which we may not know.

We saw that when performing prediction error decomposition that we needed to calculate the distribution of $(X_t \mid X_1, \dots, X_{t-1})$. This we have now done.

Example 8.3

Consider the state space model

$$\begin{array}{ll} \text{observed data} & X_t = S_t + v_t \,, \\ \text{unobserved state} & S_t = S_{t-1} + w_t \,, \end{array}$$

where v_t, w_t are independent errors, $v_t \sim N(0, V)$ and $w_t \sim N(0, W)$.

Then we have $F_t = 1$, $G_t = 1$, $V_t = V$, $W_t = W$. $R_t = P_{t-1} + W$. So if $(S_{t-1} \mid X_1, \dots, X_{t-1}) \sim N\left(\hat{S}_{t-1}, P_{t-1}\right)$ then $(S_t \mid X_1, \dots, X_t) \sim N\left(\hat{S}_t, P_t\right)$, where

$$\hat{S}_t = \hat{S}_{t-1} + R_t (V + R_t)^{-1} (X_t - \hat{S}_{t-1})$$

$$P_t = R_t - \frac{R_t^2}{V + R_t} = \frac{V R_t}{V + R_t} = \frac{V(P_{t-1} + W)}{V + P_{t-1} + W} .$$

Asymptotically, $P_t \to P$, where P is the positive root of $P^2 + WP - WV = 0$ and $\hat{S}_t$ behaves like $\hat{S}_t = (1 - \alpha) \sum_{r=0}^{\infty} \alpha^r X_{t-r}$, where $\alpha = V/(V + W + P)$. Note that this is simple exponential smoothing.

Equally, we can predict S_{T+k} given $(X_1, \dots, X_T)$ as $N\left(\hat{S}_{T,k}, P_{T,k}\right)$ where

$$\begin{aligned} \hat{S}_{T,0} &= S_t \,, \\ P_{T,0} &= P_T \,, \\ \hat{S}_{T,k} &= \hat{S}_T \,, \\ P_{T,k} &= P_T + kW \,. \end{aligned}$$

So $(X_{T+k} \mid X_1, \dots, X_T) \sim N\left(\hat{S}_T, V + P_T + kW\right)$.