7 Discounting

7.1 Stochastic shifting

Consider the generic DLM

$$y_t = F_t^\top \theta_t + \nu_t \qquad \nu_t \sim \langle 0,\, V_t \rangle$$
$$\theta_t = G_t\, \theta_{t-1} + \omega_t \quad \omega_t \sim \langle \mathbf{0},\, W_t \rangle$$

in the case where n, the length of the state vector, is greater than 1. This is not as unambiguous as it looks. The interpretation of the state vector θ_t and the state equation error variance W_t are intimately linked.

Suppose we have a random vector quantity

$$\epsilon_t \sim \langle 0,\, U_t \rangle$$

where ϵ_t is uncorrelated with all other disturbances (i.e. all the ν and ω terms) and U_t is chosen to have the property that $\mathsf{Var}\,[F_t{}^\top \epsilon_t] = F_t^\top U_t\, F_t = 0$. By construction we must have $F_t^\top \epsilon_t = 0$ with probability one.

Now take our DLM and add a bit of extra variance to the state equation in the form of $G_t\, \epsilon_{t-1}$, to give

$$y_t = F_t^\top \theta_t + \nu_t$$
$$\theta_t = G_t\, \theta_{t-1} + (\omega_t + G_t\, \epsilon_{t-1})$$

or, if we define a new state vector $\psi_t := \theta_t + \epsilon_t$,

$$y_t = F_t^\top \psi_t + \nu_t$$
$$\psi_t = G_t\, \psi_{t-1} + \omega'_t$$

where $\omega'_t = \omega_t + \epsilon_t \sim \langle 0,\, W_t + U_t \rangle$. Note that we have added $F_t^\top \epsilon_t$ to the observation equation as well, but this doesn't change anything by construction. From this you an see that a model described by the set $\{F_t, V_t, G_t, W'_t\}$ can be represented an uncountable number of different ways *with different state vectors* by making different choices for W_t and U_t, where $W'_t = W_t + U_t$.

So perhaps we can ignore this ambiguity in practice, if the chances of finding a U_t satisfying $\mathsf{Var}\,[F_t^\top \epsilon_t] = 0$ are small. Unfortunately we can always find a U_t that fits the bill, providing $n \geq 2$. Here's one way, where we drop the t subscripts for convenience. Partition ϵ as $\epsilon = [\epsilon' \mid \epsilon_n]$, where $\epsilon' := (\epsilon_1, \ldots, \epsilon_{n-1}) \sim \langle \mathbf{0},\, U_1 \rangle$ for arbitrary $(n-1) \times (n-1)$ variance matrix U_1. Define

$$\epsilon_n = -(f_1\, \epsilon_1 + \cdots + f_{n-1}\, \epsilon_{n-1})/f_n =: F_1^\top \epsilon'$$

where $F = (f_1, \ldots, f_n)^\top$. This ensures that $\mathsf{Var}\,[F^\top \epsilon] = 0$, as required. We see that $\mathsf{E}[\epsilon] = \mathbf{0}$ and

$$U := \mathsf{Var}\,[\epsilon] = \begin{pmatrix} U_1 & U_1\, F_1 \\ F_1^\top U_1 & F_1^\top U_1\, F_1 \end{pmatrix}.$$

For example, if $F = (1, 1)^\top$ then $F_1 = -1$. Setting $U_1 = u > 0$, we must have

$$U = u \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}.$$

As u is an arbitrary positive scalar there are any number of U matrices that fit the bill, and likewise any number of state vectors that give rise to the same DLM.

7.2 Discounting

The moral of the previous subsection is that W_t is tricky. It is also the most difficult quantity to elicit in practice. For these and other reasons (see West and Harrison, 1997, pp. 194-5), a practical strategy for avoiding the complications of specifying W_t has been developed. This is known 'discounting'.

At time $t-1$ we have $\theta_{t-1} \mid D_{t-1} \sim \langle m_{t-1}, C_{t-1} \rangle$. The value W_t enters into the determination of $R_t = \mathsf{Var}\,[\theta_t \mid D_{t-1}]$:

$$R_t = G_t\, C_{t-1}\, (G_t)^\intercal + W_t = P_t + W_t \quad \text{where } P_t := G_t\, C_{t-1}\, (G_t)^\intercal.$$

Clearly $R_t \geq P_t$.[1] The bigger that W_t is, the more that R_t exceeds P_t. This has lead to the 'short-cut' in which W_t is dropped and the relation between R_t and P_t is specified directly through a discount factor $\delta \in (0, 1]$:

$$R_t = \frac{1}{\delta} P_t. \tag{1}$$

Small values of δ indicate that there is a lot of uncertainty about the evolution of the state vector. We can solve for the implicit value of W_t as

$$W_t = \frac{1-\delta}{\delta} P_t. \tag{2}$$

Typically we choose values for δ in the range $[0.9, 0.99]$, although we should pay close attention to diagnostics to alert us to a poor choice.

From a given starting point C_0 we can compute $W_1, W_2, \ldots$ directly, without reference to the data $y_1, y_2, \ldots$, because the updating equations for C_t do not depend upon the data. Therefore with this modification we do not violate the structure of our graph of the DLM, on which all our analysis is based. For adjusting by new data we simply use (1) to compute R_t for a given choice of δ (which we might choose to allow to vary in time as well). For filtering we use the implicit value for W_t given in (2).

For forecasting we have to be a little more subtle. Suppose, for simplicity, that the DLM is time-invariant. In this case we can easily show that

$$R_t(k) := \mathsf{Var}\,[\theta_{t+k} \mid D_t] = \sum_{i=0}^{k-1} G^i W (G^i)^\intercal + G^k C_t (G^k)^\intercal. \tag{3}$$

What happens instead if we use discounting? Setting

$$P_{t+k} = G R_t(k-1) G^\intercal,$$

as seems natural, we find that

$$R_t(k) = G R_t(k-1) G^\intercal + W_{t+k} = P_{t+k} + W_{t+k} = \frac{1}{\delta} P_{t+k}.$$

By successive back-substitution we can get down to

$$R_t(k) = \frac{1}{\delta}\big(G R_t(k-1) G^\intercal\big) = \cdots = \frac{1}{\delta^k} G^k C_t\, (G^k)^\intercal. \tag{4}$$

This shows that the information in C_t decays exponentially as we go forwards in the forecast. Clearly (4) is not consistent with the usual behaviour of the DLM forecast variance, as given in (3). Therefore we usually define 'future' values for W to be the same as the one-step-ahead value at time t:

$$\mathsf{Var}\,[\omega_{t+k} \mid D_t] = W_{t+1} = \frac{1-\delta}{\delta} P_{t+1} \quad \text{for all } k \geq 1.$$

[1] If A and B are both non-negative definite symmetric (NNDS) matrices then we say that $A \geq B$ if $A = B + C$ for some C where C is also NNDS.