State-Space Models

Initialization, Estimation and Smoothing of the Kalman Filter
Initialization of the Kalman Filter

The Kalman filter shows how to update past predictors and the corresponding prediction error variances when a new observation becomes available, that is, how to move from time \((t-1)\) to time \(t\). Application of the Kalman filter requires therefore an initial predictor \(\hat{\beta}_0\) and the corresponding covariance matrix, 

\[ P_0 = E[(\hat{\beta}_0 - \beta_0)(\hat{\beta}_0 - \beta_0)'] \].

- Several efficient procedures have been proposed in the literature, but since under certain regularity conditions and for sufficiently long series the current predictors don’t depend on the initialization process, simple procedures can usually be applied.
Initialization procedure

A simple initialization procedure is as follows:

1- Initialize nonstationary components $\beta_{0,j}$ of the state vector by any value (say, $\beta_{0,j} = 0$) and a very large variance, $P_{0,j} = E(\hat{\beta}_{0,j} - \beta_{0,j})^2 = M$ (Large relative to the magnitude of the series.)

2- Initialise stationary components by their (unconditional) mean and variance.

- The use of 1 corresponds to the use of a noninformative (diffuse) prior distribution for the corresponding parameters under a Bayesian paradigm.
Example of Initialization

Consider the BSM but suppose that the observed series \( \{y_t, t = 1, 2, \ldots\} \) consists of survey estimators and hence is subject to sampling errors;

\[ y_t = Y_t + e_t, \]

where \( Y_t \) defines the corresponding population value and \( e_t \) is the sampling error assumed to follow an AR(2) process. The state-space model is therefore,

**Observation Equation:**

\[ y_t = L_t + S_t + e_t + \varepsilon_t \]

**Transition Equation:**

\[ L_t = L_{t-1} + R_{t-1} + \eta_{1t} ; \quad R_t = R_{t-1} + \eta_{2t}, \]

\[ S_t = -S_{t-1} - S_{t-2} - S_{t-3} + \eta_{3t}, \text{ (or trigonometric model)} \]

\[ e_t = \phi_1 e_{t-1} + \phi_2 e_{t-2} + \eta_{4t}. \]
State-space representation of the extended BSM (quarterly series)

Let \( \beta_t = (L_t, R_t, S_t, S_{t-1}, S_{t-2}, e_t, e_{t-1})' \)

**Observation Equation**

\[ y_t = (1, 0, 1, 0, 0, 1, 0) \beta_t + \varepsilon_t \]

**Transition Equation**

\[
\beta_t = \begin{bmatrix}
1 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & -1 & -1 & -1 & 0 & 0 \\
0 & 0 & -1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & -1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 \\
\end{bmatrix}
\begin{bmatrix}
L_{t-1} \\
R_{t-1} \\
S_{t-1} \\
S_{t-2} \\
S_{t-3} \\
e_{t-1} \\
e_{t-2} \\
\end{bmatrix}
+ \begin{bmatrix}
\eta_{1t} \\
\eta_{2t} \\
\eta_{3t} \\
0 \\
0 \\
\eta_{4t} \\
0 \\
\end{bmatrix}
\]

\[
Q = Cov \begin{bmatrix}
\eta_{1t} \\
\eta_{2t} \\
\eta_{3t} \\
0 \\
\eta_{4t} \\
0 \\
0 \\
0 \\
\end{bmatrix}
= \begin{bmatrix}
q_1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & q_2 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & q_3 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & q_4 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\end{bmatrix}
\]
Possible initialization for the extended BSM

\[ \beta_t = (L_t, R_t, S_t, S_{t-1}, S_{t-2}, e_t, e_{t-1})'. \]

Except for \( e_t \) and \( e_{t-1} \), the other 5 component of \( \beta_t \) are nonstationary. Also, as we know by now,

\[ E(e_t) = 0, \quad Var(e_t) = \frac{q_4}{(1 - \rho_1\phi_1 - \rho_2\phi_2)} = V_e. \]

Hence: \( \hat{\beta}_0 = (0,0,0,0,0,0,0)' \), and

\[
P_0 = \begin{bmatrix}
M & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & M & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & M & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & M & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & M & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & V_e & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & V_e & 0 \\
\end{bmatrix}
\]

- \( M \) is a large number.
Estimation of hyper-parameters

So far we assumed that the matrices $X_t$, $T$, $\Sigma$ and $Q$ are known. In practice, the elements of $\Sigma$ and $Q$ are usually unknown, and possibly also some of the elements of the transition matrix $T$. (In the previous example, $T$ contains $\phi_1$ and $\phi_2$.)

Denote by $\lambda$ the vector of all the unknown parameters, often referred to as hyper-parameters.

Under the classical (Frequentist) approach, these parameters are estimated from the observed series using Maximum Likelihood (ML) or other standard techniques.

- Unlike in classical statistics, the observations $\{Y_t, t = 1, 2, \ldots N\}$ are not independent, and hence the likelihood is not a simple product of the densities of the individual observations.
**Prediction Error Decomposition**

Denote $Y_0 = \beta_0$ and let $Y_{(t)} = (Y_t, Y_{t-1}, \ldots, Y_0)$ denote the vector of observations until and including time $t$. The following decomposition follows from repeated application of **Bayes Theorem**, where $f(\cdot)$ denotes the corresponding probability density functions (pdf).

\[
f[Y_{(N)}] = f(Y_N, Y_{N-1}, \ldots, Y_1, Y_0) = f[Y_N | Y_{(N-1)}] f[Y_{(N-1)}] = f[Y_N | Y_{(N-1)}] f[Y_{N-1} | Y_{(N-2)}] f[Y_{(N-2)}] = \cdots = \prod_{t=2}^{N} f[Y_t | Y_{(t-1)}] f[Y_1 | \beta_0] f(\beta_0) \ (\beta_0 = Y_0).
\]

Assuming that $\epsilon_t$ and $\eta_t$ are normal and likewise for $\beta_0$, each conditional distribution $f[Y_t | Y_{(t-1)}]$ is also normal:

\[
Y_t | Y_{(t-1)} \sim N(\hat{Y}_{t|t-1}, F_t),
\]

\[
F_t = (X_t P_{t|t-1} X_t' + \Sigma) = Var(Y_t - \hat{Y}_{t|t-1}).
\]
Estimation of hyper-parameters, (cont.)

\[ f(Y_N, Y_{N-1}, \ldots, Y_1, Y_0) = \prod_{t=2}^{N} f[Y_t \mid Y_{(t-1)}] f[Y_1 \mid \beta_0] f(\beta_0) \]

\[ Y_t \mid Y_{(t-1)} \sim N(\hat{Y}_{t|t-1}, F_t) \]

\[ F_t = (X_t P_{t|t-1} X_t' + \Sigma) = \text{Var}(Y_t - \hat{Y}_{t|t-1}) = \text{Var}(e_t). \]

The log-likelihood has therefore the form,

\[ \log[L; \lambda \mid Y(\ldots)] = -\frac{nN}{2} \log(2\pi) - \frac{1}{2} \sum_{t=1}^{N} \log |F_t| \]

\[ -\frac{1}{2} \sum_{t=1}^{N} e_t' F_t^{-1} e_t + \log f(\beta_0) \quad [n=\text{Dim}(Y_t)] \]

• We need to specify \( f(\beta_0) \) which amounts to the Initialization of the Kalman filter.

Possibilities

1- Fix \( \beta_0 \) and add the unknown elements of \( \beta_0 \) to the vector \( \lambda \). This means that we drop the term \( \log f(\beta_0) \) from the likelihood. If \( \beta_0 \) is of high dimension \( [p=\text{dim}(\beta_0) \text{ is large}] \), we end up with many more unknown parameters.
Initialization of the likelihood (cont.)

2- Regress $Y_1...Y_m$ as a function of $\beta_m$ to obtain $\hat{\beta}_m$ and $P_m$, base the likelihood on $\prod_{t=m+1}^{N} f[Y_t | Y_{(t-1)}]$. The value of $m$ is set to the smallest value that yields an estimator with a proper covariance matrix $P_m$.

Example: If $n>p$, we can write,

$$\hat{\beta}_1 = (X_1'\Sigma^{-1}X_1)^{-1}X_1'\Sigma^{-1}Y_1, \quad P_1 = (X_1'\Sigma^{-1}X_1)^{-1}$$

($\Sigma$ is unknown), and base the likelihood on $\prod_{t=2}^{N} f[Y_t | Y_{(t-1)}]$.

If $n\leq p$ we need to sacrifice more observations.

3- Use the Initialization procedure discussed before (and assume a distribution for $\beta_0$). Here again we need to sacrifice the first $m$ observations until we get an estimator $\hat{\beta}_m$ with a proper covariance matrix $P_m$. 
Maximum likelihood estimation (cont.)

Having established the log-likelihood, it is then **maximized** with respect to the elements contained in $\lambda$ (the **hyper-parameters**). The maximization process yields also the asymptotic **covariance** matrix of the estimators via the **Inverse Information** matrix.
Consider the AR(1) model, $Z_t = \phi Z_{t-1} + \varepsilon_t$
where $Z_t = (Y_t - \mu)$.

$$E(Z_t \mid Z_{t-1}) = \phi Z_{t-1}, \ Var(Z_t \mid Z_{t-1}) = \sigma^2 = Var(\varepsilon_t)$$

We assume that $\varepsilon_t \sim N(0, \sigma^2)$.

$$f(Z_N, \ldots, Z_1) = \prod_{t=2}^{N} f(Z_t \mid Z_{t-1}) f(Z_1)$$

$$= \prod_{t=2}^{N} \left[ \frac{1}{\sqrt{2\pi\sigma}} \exp\left[\frac{(Z_t - \phi Z_{t-1})^2}{2\sigma^2}\right] \right] f(Z_1). \text{ Hence,}$$

$$\log(L) = -\frac{N-1}{2} \log(2\pi) - \frac{N-1}{2} \log(\sigma^2)$$

$$+ \sum_{t=2}^{N} \left[ \frac{(Z_t - \phi Z_{t-1})^2}{2\sigma^2} \right] + \log[f(Z_1)].$$

• We need to specify $f(Z_1)$. 

---

**Example: MLE for AR(1) models**
Specification of $f(Z_t)$ for the AR(1) model

**Possibilities**

1- **Fix** $Z_1$ and maximize the likelihood ignoring the last term ($f(Z_1) = \text{const.}$)

2- **Assume** $Z_1 \sim N[0, \frac{\sigma^2}{(1 - \phi^2)}]$ and add

$$\log[f(Z_1)] = \left[ -\frac{1}{2} \log(2\pi\sigma_Z^2) + \frac{Z_1^2}{2\sigma_Z^2} \right]$$

where $\sigma_Z^2 = \sigma^2/(1 - \phi^2)$.

- If we substitute $(Y_t - \mu)$ for $Z_t$ everywhere in the likelihood we can maximize also with respect to $\mu$. 


We considered so far the prediction of the state vectors (i.e., the computation of $\hat{\beta}_{t+k|t-1}$, $k=0,1,…$), and the updating of the current state vector as new data become available, i.e., the computation of $\hat{\beta}_t = \hat{\beta}_{t|t}$.

- Having new data $Y_t$, we may want to modify the previous predictors, $\hat{\beta}_{t-1}$, $\hat{\beta}_{t-2}$..., which is known as smoothing.

For example, modify the estimates of the trend and seasonal effects produced in a previous year, which in X-11 terminology is called revision.

**Note:** for the last time point $N$, $\hat{\beta}_{N|N} = \hat{\beta}_N$. 
Smoothing of the previous state vector

We like to modify (smooth) the estimator $\hat{\beta}_{t-1}$ computed from the observations $Y_1, \ldots, Y_{t-1}$ when a new observation $Y_t$ become available. We again have two independent predictors for $\beta_{t-1}$:

$$\hat{\beta}_{t-1} = \beta_{t-1} + u_{t-1}^* ; \quad E(u_{t-1}^*u_{t-1}'^*) = P_{t-1}$$

$$Y_t = X_t \beta_t + \varepsilon_t = X_t T \beta_{t-1} + X_t \eta_t + \varepsilon_t = X_t T \beta_{t-1} + v_t^* ; \quad E(v_t^*v_t^*) = X_t Q X_t' + \Sigma$$

• $u_{t-1}^*$ and $v_t^*$ are independent, given $\beta_{t-1}$.
Smoothing of previous state vector (cont.)

Set the ‘regression’ model,

\[
\begin{bmatrix}
\hat{\beta}_{t-1} \\
Y_t
\end{bmatrix} = \begin{bmatrix}
I \\
X_tT
\end{bmatrix} \beta_{t-1} + \begin{bmatrix}
u_{t-1}^* \\
v_t^*
\end{bmatrix}, \text{ where}
\]

\[
\text{Var} \begin{bmatrix}
u_{t-1}^* \\
v_t^*
\end{bmatrix} = \begin{bmatrix}
P_{t-1} & 0 \\
0 & X_tQX_t' + \Sigma
\end{bmatrix} = V.
\]

The generalised least square predictor (GLS) is,

\[
\hat{\beta}_{t-1|t} = \left[ [I, T'X_t']V^{-1} \left[ \begin{bmatrix}
I \\
X_tT
\end{bmatrix} \right]^{-1} [I, T'X_t']V^{-1} \right] \begin{bmatrix}
\hat{\beta}_{t-1} \\
Y_t
\end{bmatrix}
\]

\[= \hat{\beta}_{t-1} + P_{t-1}^*(\hat{\beta}_t - T\hat{\beta}_{t-1}) \quad ; \quad P_{t-1}^* = P_{t-1}T'P_{t|t-1}^{-1}.
\]

- The predictor \( \hat{\beta}_{t-1} \) is updated based on the prediction error of \( (\hat{\beta}_t - T\hat{\beta}_{t-1}) = (\hat{\beta}_t - \hat{\beta}_{t|t-1}). \)
**Prediction Error Variance**

\[ P_{t-1|t} = E[(\hat{\beta}_{t-1|t} - \beta_{t-1})(\hat{\beta}_{t-1|t} - \beta_{t-1})'] \]

\[ = P_{t-1} - P_{t-1}^* (P_{t|t-1} - P_t) P_{t-1}^* \] \[ \Rightarrow P_{t-1|t} \leq P_{t-1}. \]

\[ P_{t-1|t} \leq P_{t-1} \] means that,

\[ Var(a' \hat{\beta}_{t-1|t}) \leq Var(a' \hat{\beta}_{t-1}) \] for all vectors \( a \neq 0 \).

- As expected, smoothing **reduces** the variance of the predictors.

**Recursive smoothing algorithm**

\[ \hat{\beta}_{t-1|N} = \hat{\beta}_{t-1} + P_{t-1}^* (\hat{\beta}_{t|N} - T \hat{\beta}_{t-1}); \quad P_{t-1}^* = P_{t-1}T'P_{t|t-1}^{-1} \]

\[ P_{t-1|N} = E[(\hat{\beta}_{t-1|N} - \beta_{t-1})(\hat{\beta}_{t-1|N} - \beta_{t-1})'] \]

\[ = P_{t-1} - P_{t-1}^* (P_{t|t-1} - P_{t|N}) P_{t-1}^*. \]

- The algorithm starts with, \( \hat{\beta}_{N|N} = \hat{\beta}_N, \quad P_{N|N} = P_N \).
Example

Random walk + noise

\[ y_t = 1 \times \theta_t + \varepsilon_t \quad , \quad \text{Var}(\varepsilon_t) = \sigma^2 \quad (X_t = 1) \]

\[ \theta_t = 1 \times \theta_{t-1} + \eta_t \quad , \quad \text{Var}(\eta_t) = q \quad (T_t = 1) \]

For this model, \( p_t^* = p_t / (p_t + q) \), and

\[ \hat{\theta}_{t|N} = (1 - p_t^*) \hat{\theta}_t + p_t^* \hat{\theta}_{t+1|N}. \]

- As \( t \to \infty \), \( p_t \to \text{const} \Rightarrow p_t / (p_t + q) \to c \)

\[ \text{and} \quad \hat{\theta}_{t|N} = (1 - c) \hat{\theta}_t + c \hat{\theta}_{t+1|N} = \hat{\theta}_t + c(\hat{\theta}_{t+1|N} - \hat{\theta}_t). \]

(See exercise).