Kalman Filter

Overview

1. Summary of Kalman filter
2. Derivations
3. ARMA likelihoods
4. Recursions for the variance

Summary of Kalman filter

Simplifications To make the derivations more direct, assume that the two noise processes are uncorrelated ($S_t = 0$) with constant variance matrices ($Q_t = Q, R_t = R$). In this setting, the natural way to express the model is

\begin{align*}
\text{State:} & \quad X_t = F X_{t-1} + V_t \quad (1) \\
\text{Observation:} & \quad Y_t = H X_t + W_t \quad (2)
\end{align*}

The goal is to find a recursive expression for

$$\hat{X}_{t|t} = \text{projection of } X_t \text{ onto } \{Y_1'\}.$$ 

(n.b. I’ve changed the time lag on the error in the state equation to look more like ARMA models.)

Least squares The optimal estimates associated with these recursions are least squares projections. The least squares predictor of a random variable $Y$ given $X_1, X_2, \ldots$ is the r.v. $\hat{Y}$ that satisfies the orthogonality condition

$$Y - \hat{Y} \perp X_1, X_2, \ldots, X_n \iff \text{Cov}(Y - \hat{Y}, X_j) = 0$$

Note that the space being projected on in the Kalman filter is finite dimensional, namely the space spanned by linear combinations of the prior observed random variables.
Solution  

It is common to express the solution as a two-step procedure (in one of two ways!). Assume that we have observed \( \{Y_1, \ldots, Y_{t-1}\} = Y_{1:t-1} \) and we have our best estimate of the state given this information,

\[
\hat{X}_{t-1|t-1} = \mathbb{E} X_t \mid Y_1, \ldots, Y_{t-1}.
\]

Assume also that we know the variance of this estimator, \( \text{Var}(\hat{X}_{t-1|t-1}) = P_{t-1|t-1} \). The two steps then are

1. Extrapolate, obtaining \( \hat{X}_{t|t-1} \).
2. Update once \( Y_t \) is observed, obtaining \( \hat{X}_{t|t} \).

The first step is easy:

\[
\begin{align*}
\hat{X}_{t|t-1} &= E[X_t \mid Y_{1:t-1}] = F \hat{X}_{t-1|t-1} \\
\tilde{P}_{t|t-1} &= \text{Var}(X_t - \hat{X}_{t|t-1}) = FP_{t|t-1}F' + Q
\end{align*}
\]

From these we obtain the updated filtered estimates

\[
\begin{align*}
\hat{X}_{t|t} &= \hat{X}_{t|t-1} + K_t (Y_t - H \hat{X}_{t|t-1}) \\
\tilde{P}_{t|t} &= \tilde{P}_{t|t-1} - K_t H \tilde{P}_{t|t-1} \tilde{H}
\end{align*}
\]

where the so-called gain of the filter is

\[
K_t = \tilde{P}_{t|t-1} H' (H \tilde{P}_{t|t-1} H' + R)^{-1}.
\]

The term \( Y_t - H \hat{X}_{t|t-1} \) is known as the innovation at time \( t \). It measures the amount of “new information” in the observation \( Y_t \) that was not known before observing \( Y_t \).

Smoothing  

Estimates \( \hat{X}_{t|n} \) based on all of the data \( Y_1, \ldots, Y_n \), \( 1 < t < n \), rather than the data up to \( t \) are known as smoothed estimates of the state (a.k.a., two-sided estimate, interpolation). See S&S, Section 6.2.
Derivations

Summary. Key results come from exploiting orthogonal projection and recursion using the Markovian structure of the state equation:

- Form orthogonal regressors.
- Simplify the orthogonal term.
- Compute the associated regression.

In general, the derivation of the filtering equations works by thinking recursively and continually “splitting” random variables into orthogonal components

\[ X_t = \hat{X}_t + \tilde{X}_t, \quad \hat{X}_t \perp \tilde{X}_t \]

by projecting \( X_t \) onto a subspace. \( \hat{X}_t = X_t - \tilde{X}_t \) are the residuals of this projection.

Benefits of orthogonality It works as in regression: adding an orthogonal variable does not “interfere” with the projection on other variables. In particular, if \( X, Y \) and \( Z \) are normal random variables and \( Y \perp Z \) then

\[
\mathbb{E}(X \mid Y, Z) = \mathbb{E}(X \mid Y) + \mathbb{E}(X \mid Z) - \mathbb{E}X
\]

proof Let \( W = \{Y, Z\} \). Then the variance matrix is block diagonal so that

\[
\mathbb{E}(X \mid W) = \mathbb{E}X + \text{Cov}(X, W) \text{Var}(W)^{-1}(W - \mathbb{E}W)
\]

\[
= (\mathbb{E}X + \text{Cov}(X, Y) \text{Var}(Y)^{-1}(Y - \mathbb{E}Y))
\]

\[
+ (\mathbb{E}X + \text{Cov}(X, Z) \text{Var}(Z)^{-1}(Z - \mathbb{E}Z)) - \mathbb{E}X
\]

Orthogonalize regressors Develop a recursion for the estimate of the state at time \( t \) given \( Y_{1:t} \). The idea is to split \( Y_{1:t} \) into two orthogonal subspaces \( \hat{Y}_{t|t-1} \) and \( Y_{1:t-1} \), so that the projection is the sum of two simpler projections. Without defining \( \hat{Y}_{t|t-1} \) (yet), we obtain (assume as usual that the mean of \( Y_t \) and \( X_t \) is zero)

\[
\hat{X}_{t|t} = \mathbb{E}[X_t | Y_t, \ldots, Y_1]
\]

\[
= \mathbb{E}[X_t | \hat{Y}_{t|t-1}, Y_{1:t-1}]
\]

\[
= \mathbb{E}[X_t | \hat{Y}_{t|t-1}] + \mathbb{E}[X_t | Y_{1:t-1}]
\]
\[ K_t \hat{Y}_{t|t-1} + \hat{X}_{t|t-1} \quad (3) \]
\[ = K_t \hat{Y}_{t|t-1} + \mathbb{E} [FX_{t-1} + V_t | Y_{1:t-1}] \]
\[ = K_t \hat{Y}_{t|t-1} + F \hat{X}_{t-1|t-1} \quad (4) \]

Note:

- The (as yet unknown) coefficient \( K_t \) is the gain of the filter at time \( t \).
- The term \( \tilde{Y}_{t|t-1} \) of \( Y_t \) orthogonal to the past \( Y_{1:t-1} \) is known as the innovation at time \( t \).

**Structure of innovation** Using the linearity of conditional expectations (or projections), write the innovation as

\[
\tilde{Y}_{t|t-1} = Y_t - \mathbb{E} [Y_t | Y_{1:t-1}]
= Y_t - \mathbb{E} [H X_t + W_t | Y_{1:t-1}]
= (H X_t + W_t) - H \hat{X}_{t|t-1}
= H \tilde{X}_{t|t-1} + W_t
= H (X_t - \hat{X}_{t|t-1}) + W_t
= H (FX_{t-1} + V_t - F \hat{X}_{t-1|t-1}) + W_t
= H F \tilde{X}_{t-1|t-1} + HV_t + W_t \quad (5)
\]

The expression (5) leads to an important form of the recursion. Substituting (5) into (4) gives

\[
\tilde{X}_{t|t} = F \hat{X}_{t-1|t-1} + K_t (Y_t - H F \tilde{X}_{t-1|t-1})
= (I - K_t H) F \hat{X}_{t-1|t-1} + K_t Y_t \quad (7)
\]

The form in the first line of (7) is generally preferred since it focuses attention upon the innovation rather than the actual observation \( Y_t \).

**Compute the gain** \( K_t \) This part is easy if we remember the fundamentals of regression. We need to regress \( X_t \) on the innovation \( \tilde{Y}_{t|t-1} \). The orthogonality condition

\[
0 = \text{Cov}(X_t - K_t \tilde{Y}_{t|t-1}, \tilde{Y}_{t|t-1}) = \mathbb{E} [(X_t - K_t \tilde{Y}_{t|t-1}) \tilde{Y}_{t|t-1}']
\]

implies

\[
\text{Cov}(X_t, \tilde{Y}_{t|t-1}) = K_t \text{Var}(\tilde{Y}_{t|t-1}).
\]
Splitting $X_t$ into orthogonal parts and using (5), we find the gain matrix via regression:

$$K_t = \text{Cov}(X_t, \tilde{Y}_{t|t-1}) \text{Var}(\tilde{Y}_{t|t-1})^{-1}$$

$$= \text{Cov}(\tilde{X}_{t|t-1} + \tilde{X}_{t|t-1}, H\tilde{X}_{t|t-1} + W_t) \text{Var}(H\tilde{X}_{t|t-1} + W_t)^{-1}$$

$$= \text{Cov}(\tilde{X}_{t|t-1}, H\tilde{X}_{t|t-1})(HP_{t|t-1}H' + R)^{-1}$$

$$= P_{t|t-1}H'(HP_{t|t-1}H' + R)^{-1}$$

Variance matrices The matrices $P_t$ and $P_{t|t-1}$ which are both variance matrices of the error in estimating the state.

$$P_t = P_{t|t} = \text{Var}(\tilde{X}_{t|t}) = (I - K_t H)P_{t|t-1}. \quad (9)$$

The matrix $P_{t|t-1}$ also has nice interpretation, namely as the conditional variance of the one-step-ahead prediction error,

$$P_{t|t-1} = FP_{t-1}F' + Q = \text{Var}(\tilde{X}_{t|t-1}).$$

ARMA likelihood

Akaike representation The canonical representation (minimal dimension state) requires correlated errors, so use the larger formulation with uncorrelated errors and dimension $d = \max(p, q + 1)$ and state coefficients arranged as

$$F = \begin{pmatrix} 0_{d-1} & I_{d-1} \\ \bar{\phi}' \end{pmatrix}$$

with the reversed coefficients in the last row. Then

$$X_t = FX_{t-1} + w_t \begin{pmatrix} 1 \\ \psi_1 \\ \vdots \\ \psi_d \end{pmatrix}$$

$\psi = (1, \psi_1, \psi_2, \ldots, \psi_{d-1})'$ are the weights from the infinite moving average representation. The observation equation picks off the first element of the state,

$$y_t = (1 0 \cdots 0)'X_t.$$
The state vector is
\[ X_t = (y_t, \mathbb{E}(y_{t+1}|t), \ldots, \mathbb{E}(y_{t+d-1}|t))'. \]

**Gaussian likelihood** Let \( y_1, \ldots, y_n \) denote a partial realization from a Gaussian ARMA process. Then the log likelihood has the form
\[ \ell(\phi, \theta) = \sum_t \log f(y_t|y_{t-1}, \ldots, y_1). \]

Since each conditional density is normal (assumed to have mean zero), the likelihood may be evaluated by knowing the sequence of conditional means and variances,
\[ \mathbb{E}(y_1) = 1, \text{Var}(y_1), \mathbb{E}[y_2|y_1], \text{Var}(y_2|y_1), \mathbb{E}[y_3|y_2, y_1], \text{Var}(y_3|y_2, y_1), \ldots, \mathbb{E}[y_n|y_{n-1}, \ldots, y_1], \text{Var}(y_n|y_{n-1}, \ldots, y_1). \]

**Kalman recursions** give both of these. The first element in \( \hat{X}_t|t-1 \) is \( \mathbb{E}[y_t|y_{t-1}, \ldots, y_1] \) and the associated conditional variance is the leading diagonal element of \( P_{t|t-1} \). The only messy issue is initializing the variance of the state at time 0 before observations. (R cites Jones, 1980, *Technometrics*)

**Recursions for the variance**

**Notation** Let \( P_t X \) denote the projection of \( X \) onto \( \{Y_t, Y_{t-1}, \ldots, Y_1\} \) (not probability), \( \langle X, Y \rangle \) denote \( \text{Cov}(X,Y) \), and \( \|x\|^2 = \text{Var}(X) \).

**Filtering equations** The Kalman filter defines the one-step-ahead estimates
\[
\begin{align*}
\hat{X}_{t|t-1} &= P_{t-1}X_t = F\hat{X}_{t-1|t-1} \\
P_{t|t-1} &= \text{Var}(X_t - \hat{X}_{t|t-1}) = FP_{t|t-1}F' + Q.
\end{align*}
\]

The updated filtered estimates are
\[
\begin{align*}
\hat{X}_{t|t} &= \hat{X}_{t|t-1} + K_t(Y_t - H\hat{X}_{t|t-1}) \\
P_{t|t} &= P_{t|t-1} - K_tHP_{t|t-1}
\end{align*}
\]

where the gain (the regression coefficient) is
\[ K_t = P_{t|t-1}H'(HP_{t|t-1}H' + R)^{-1}. \]
Recursions 1. Expression for $P_{t|t-1}$ is immediate. For $P_{t|t}$,

$$P_{t|t} = \|X_t - \hat{X}_{t|t}\|^2 = \|X_t - \hat{X}_{t|t-1} - K_t(Y_t - H\hat{X}_{t|t-1})\|^2 = \|-K_tW_t + (I - K_tH)\hat{X}_{t|t-1}\|^2 = K_tRK_t' + (I - K_tH)P_{t|t-1}(I - K_tH)'$$

While correct (and avoiding any matrix inversions), this expression for $P_{t|t}$ conceals the evolution of the recursion... After all, shouldn’t $P_{t|t}$ be “smaller” than $P_{t|t-1}$?

Regression analogy Notice the form for the residual SS in a regression equation,

$$(Y - X\hat{\beta})'(Y - X\hat{\beta}) = Y'Y - \hat{\beta}'X'Y - Y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta} = Y'Y - \hat{\beta}'X'Y$$

Recursions 2. For $P_{t|t}$,

$$P_{t|t} = \|(X_t - \hat{X}_{t|t-1}) - K_t\hat{Y}_{t|t-1}\|^2 = \|\hat{X}_{t|t-1}\|^2 - \langle\hat{X}_{t|t-1}, K_t\hat{Y}_{t|t-1}\rangle - \langle K_t\hat{Y}_{t|t-1}, \hat{X}_{t|t-1}\rangle + \|K_t\hat{Y}_{t|t-1}\|^2 = P_{t|t-1} - \text{Cov}(\hat{X}_{t|t-1}, K_tH\hat{X}_{t|t-1}) - \text{Cov}(K_tH\hat{X}_{t|t-1}, \hat{X}_{t|t-1}) + K_t\text{Var}(\hat{Y}_{t|t-1})K_t' = P_{t|t-1} - \text{Cov}(\hat{X}_{t|t-1}, H'K_t' - K_tH\text{Cov}(\hat{X}_{t|t-1}, \hat{X}_{t|t-1}) + \text{Cov}(\hat{X}_{t|t-1}, \hat{Y}_{t|t-1})K_t' = (I - K_tH)P_{t|t-1}$$

where the terms cancel as in regression. Clearly, the gain controls the rate at which the information accumulates with new observations.