Kalman Filter

Overview

1. Summary of filtering equations
2. Derivations
3. Application to ARMA likelihoods
4. Further recursions

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

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

The goal is to find a recursive expression for

$$\hat{X}_{t|t} = \text{projection of } X_t \text{ onto } \{Y_1^t\}.$$

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 defined as 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 upon is finite dimensional in the Kalman filter.

Solution It is common to express the solution as a two-step procedure. Assume that we have observed $Y_1, \ldots, Y_{t-1}$ and we have our best estimate of the state given this information, $\hat{X}_{t-1|t-1}$. We also 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 $\hat{X}_{t|t-1}$. This is easy.
2. Update once Y_t is observed, obtaining $\hat{X}_{t|t}$.

The first step is handled as

$$\begin{aligned} \hat{X}_{t|t-1} &= E[X_t|Y_1^{t-1}] = F\hat{X}_{t-1|t-1} \\ P_{t|t-1} &= \text{Var}(X_t - \hat{X}_{t|t-1}) = FP_{t|t}F' + Q \end{aligned}$$

and from these we obtain the updated filtered estimates

$$\begin{aligned} \hat{X}_{t|t} &= \hat{X}_{t|t-1} + K_t(Y_t - H\hat{X}_{t|t-1}) \\ P_{t|t} &= P_{t|t-1} - K_t H P_{t|t-1} \end{aligned}$$

where the so-called *gain* of the filter is

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

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. See S&S, Section 6.2.

Derivations

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

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

In general, most of the filtering works by thinking recursively and continually "splitting" random variable into orthogonal components

$$X_t = \widehat{X}_t + \tilde{X}_t, \quad \widehat{X}_t \perp \tilde{X}_t$$

by projecting X_t onto a subspace. $\tilde{X}_t$ 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

$$\begin{aligned}
\mathbb{E}\,(X \mid W) &= \mathbb{E}\,X + \operatorname{Cov}(X, W)\operatorname{Var}(W)^{-}1(W - \mathbb{E}\,W) \\
&= (\mathbb{E}\,X + \operatorname{Cov}(X, Y)\operatorname{Var}(Y)^{-1}(Y - \mathbb{E}\,Y)) \\
&\quad +(\mathbb{E}\,X + \operatorname{Cov}(X, Z)\operatorname{Var}(Z)^{-1}(Z - \mathbb{E}\,Z)) - \mathbb{E}\,X
\end{aligned}$$

Othogonalize regressors Develop a recursion for the estimate of the state at time t given $Y_1^t = \{Y_1, \ldots, Y_t\}$. The idea is to split Y_1^t into two orthogonal subspaces $\tilde{Y}_{t|t-1}$ and Y_1^{t-1}, so that the projection is the sum of two simpler projections. Without defining $\tilde{Y}_{t|t-1}$, we obtain (assume as usual that the mean of Y_t and X_t is zero)

$$\begin{aligned}
\widehat{X}_{t|t} &= \mathbb{E}\,[X_t|Y_t, \ldots, Y_1] \\
&= \mathbb{E}\,[X_t|\tilde{Y}_{t|t-1}, Y_1^{t-1}] \\
&= \mathbb{E}\,[X_t|\tilde{Y}_{t|t-1}] + \mathbb{E}\,[X_t|Y_1^{t-1}] \\
&= K_t\tilde{Y}_{t|t-1} + \widehat{X}_{t|t-1} && (3) \\
&= K_t\tilde{Y}_{t|t-1} + \mathbb{E}\,[F\,X_{t-1} + V_t|Y_1^{t-1}] \\
&= K_t\tilde{Y}_{t|t-1} + F\widehat{X}_{t-1|t-1} && (4)
\end{aligned}$$

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

$$\begin{aligned}
\tilde{Y}_{t|t-1} &= Y_t - \mathbb{E}\,[Y_t|Y_1^{t-1}] \\
&= Y_t - \mathbb{E}\,[H\,X_t + W_t|Y_1^{t-1}] \\
&= (HX_t + W_t) - H\widehat{X}_{t|t-1} \\
&= H\tilde{X}_{t|t-1} + W_t && (5) \\
&= H(X_t - \widehat{X}_{t|t-1}) + W_t
\end{aligned}$$

$$
\begin{aligned}
&= H(FX_{t-1} + V_t - F\widehat{X}_{t-1|t-1}) + W_t \\
&= HF\tilde{X}_{t-1|t-1} + HV_t + W_t \qquad (6)
\end{aligned}
$$

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

$$
\begin{aligned}
\widehat{X}_{t|t} &= F\widehat{X}_{t-1|t-1} + K_t(Y_t - HF\widehat{X}_{t-1|t-1}) \\
&= (I - K_t H)F\widehat{X}_{t-1|t-1} + K_t Y_t \qquad (7)
\end{aligned}
$$

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

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

implies

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

Splitting X_t into orthogonal parts and using (5), we find the gain matrix via regression:

$$
\begin{aligned}
K_t &= \mathrm{Cov}(X_t, \tilde{Y}_{t|t-1})\,\mathrm{Var}(\tilde{Y}_{t|t-1})^{-1} \\
&= \mathrm{Cov}(\widehat{X}_{t|t-1} + \tilde{X}_{t|t-1}, H\tilde{X}_{t|t-1} + W_t)\,\mathrm{Var}(H\tilde{X}_{t|t-1} + W_t)^{-1} \\
&= \mathrm{Cov}(\tilde{X}_{t|t-1},\ H\tilde{X}_{t|t-1})(HP_{t|t-1}H' + R)^{-1} \\
&= P_{t|t-1}H'(HP_{t|t-1}H' + R)^{-1} \qquad (8)
\end{aligned}
$$

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

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

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

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

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

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

with the reversed coefficients in the last row. Then

$$\boldsymbol{X}_t = F\boldsymbol{X}_{t-1} + w_t \begin{pmatrix} 1 \\ \psi_1 \\ \vdots \\ \psi_d \end{pmatrix}$$

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

$$y_t = (1\ 0\ \cdots\ 0)'\boldsymbol{X}_t\ .$$

The state vector is

$$\boldsymbol{X}_t = (\widehat{X}_{t|t} = y_t, \widehat{X}_{t+1|t}, \ldots, \hat{X}_{t+d-1|t})'.$$

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,

$$\mathbb{E}\,(y_1) = 1,\ \mathrm{Var}(y_1),\quad \mathbb{E}\,[y_2|y_1],\quad \mathrm{Var}(y_2|y_1),\ \mathbb{E}\,[y_3|y_2, y_1],\ \mathrm{Var}(y_3|y_2, y_1),$$
$$\ldots,\qquad \mathbb{E}\,[y_n|y_{n-1}, \ldots, y_1],\ \mathrm{Var}(y_n|y_{n-1}, \ldots, y_1)\ .$$

Kalman recursions give both of these. The first element in $\widehat{X}_{t|t-1}$ 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

$$\begin{aligned}\widehat{X}_{t|t-1} &= P_{t-1}X_t = F\hat{X}_{t-1|t-1}\\ P_{t|t-1} &= \text{Var}(X_t - \widehat{X}_{t|t-1}) = FP_{t|t}F' + Q\ .\end{aligned}$$

The updated filtered estimates are

$$\begin{aligned}\widehat{X}_{t|t} &= \widehat{X}_{t|t-1} + K_t(Y_t - H\widehat{X}_{t|t-1})\\ P_{t|t} &= P_{t|t-1} - K_tHP_{t|t-1}\end{aligned}$$

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 $P_{t|t}$,

$$\begin{aligned}P_{t|t} &= \|\boldsymbol{X}_t - \widehat{X}_{t|t}\|^2\\ &= \|\boldsymbol{X}_t - \widehat{X}_{t|t-1} - K_t(Y_t - H\widehat{X}_{t|t-1})\|^2\\ &= \| - K_tW_t + (I - K_tH)\tilde{X}_{t|t-1}\|^2\\ &= K_tRK_t' + (I - K_tH)P_{t|t-1}(I - K_tH)'\end{aligned}$$

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,

$$\begin{aligned}(Y - X\hat{\beta})'(Y - X\hat{\beta}) &= Y'Y - \hat{\beta}'X'Y - Y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta} \\ &= Y'Y - \hat{\beta}'X'Y\end{aligned}$$

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

$$\begin{aligned}P_{t|t} &= \|(\boldsymbol{X}_t - \widehat{X}_{t|t-1}) - K_t\tilde{Y}_{t|t-1}\|^2 \\ &= \|\tilde{X}_{t|t-1}\|^2 - \langle\tilde{X}_{t|t-1},\, K_t\tilde{Y}_{t|t-1}\rangle - \langle K_t\tilde{Y}_{t|t-1},\, \tilde{X}_{t|t-1}\rangle + \|K_t\tilde{Y}_{t|t-1}\|^2 \\ &= P_{t|t-1} - \text{Cov}(\tilde{X}_{t|t-1}, K_tH\tilde{X}_{t|t-1}) - \text{Cov}(K_tH\tilde{X}_{t|t-1}, \tilde{X}_{t|t-1}) + K_t\,\text{Var}(\tilde{Y}_{t|t-1})K_t' \\ &= P_{t|t-1} - \text{Cov}(\tilde{X}_{t|t-1}, \tilde{X}_{t|t-1})H'K_t' - K_tH\,\text{Cov}(\tilde{X}_{t|t-1}, \tilde{X}_{t|t-1}) + \text{Cov}(\tilde{X}_{t|t-1}, \tilde{Y}_{t|t-1})K_t' \\ &= P_{t|t-1} - K_tHP_{t|t-1} \\ &= (I - K_tH)P_{t|t-1}\ ,\end{aligned}$$

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