Discrete Kalman Filter

Xiaoming Hu
Optimization and Systems Theory
Royal Institute of Technology
SE 100 44 Stockholm, Sweden

Content

In 1960, Rudolf E. Kalman published his famous paper describing the solution to the discrete-data linear filtering problem.

- Recursive
- Optimal
- First applied to Apollo 11 (navigation computer)

A simple example

Two persons make an observation of something (say the height of a building) each.

- Person1: $y_1,\ \sigma^2_{y_1}$.
- Person2: $y_2,\ \sigma^2_{y_2}$.

We first use

$$\begin{aligned}\hat{x}_1 &= y_1\\ \sigma_1^2 &= \sigma^2_{y_1}\end{aligned}$$

(No priori information about x is available!)

We then update:

$$\begin{aligned}\hat{x}_2 &= \hat{x}_1 + K(y_2 - \hat{x}_1)\\ K &= \frac{\sigma_1^2}{\sigma_1^2+\sigma_{y_2}^2} \Leftarrow \textbf{Kalman gain}\\ \sigma_2^2 &= \frac{\sigma_1^2\sigma_{y_2}^2}{\sigma_1^2+\sigma_{y_2}^2}\end{aligned}$$

$$\Rightarrow \hat{x}_2 = \frac{\sigma_{y_2}^2 y_1 + \sigma_{y_1}^2 y_2}{\sigma_{y_1}^2+\sigma_{y_2}^2} \Leftarrow \textbf{Gauss}-\textbf{Markov}$$

Setup

Consider

$$\begin{aligned}x(t+1) &= Ax(t) + Bv(t)\\ y &= Cx + Dw(t)\\ x(0) &= x_0\ (unknown),\end{aligned}$$

where $v(t)$ and $w(t)$ are white noise with covariance

$$Ev(t)v(t)^T = Q > 0,\ Ew(t)w(t)^T = R > 0.$$

Question 1: Given the measurements $y(0), \cdots, y(t)$, what is the "best" estimation for $x(t)$, the best prediction for $x(t+1)$?

Here "best" is in the sense that the estimation has the least mean square error:

$$E\|x(t) - \hat{x}^*(t)\|^2 \le E\|x(t) - \hat{x}(t)\|^2.$$

Question 2: If $\hat{x}_{t-1}^*(t)$ is the best estimation (prediction) based on $y(0), \cdots, y(t-1)$, can we express the optimal estimation after $y(t)$ is available as

$$\hat{x}^*(t) = \hat{x}_{t-1}^*(t) + K(t)(y(t) - C\hat{x}_{t-1}^*(t))?\ (Recursively)$$

i.e.,

$$E\|x(t) - \hat{x}_{t-1}^*(t) - K^*(t)(y(t) - C\hat{x}_{t-1}^*(t))\|^2 = E\|x(t) - \hat{x}^*(t)\|^2?$$

Least square estimation

Given a linear relation

$$y(t) = \sum_{i=1}^{N} x_i f_i(t),$$

suppose $f_i(t)$ are known and linearly independent, but the coefficients x_i are to be determined. We do M $(M \ge N)$ experiments in order to decide the coefficients:

$$y(t_j) = \sum_{i=1}^{N} x_i f_i(t_j),\ j = 1, \cdots, M.$$

$\Rightarrow$

$$Fx = b.$$

Now find an $\hat{x}$ such that

$$\|F\hat{x} - b\|^2 \to min.$$

Taking the derivative, one get

$$\hat{x} = (F^T F)^{-1} F^T b.$$

For any vector $y = Fx \in Im(F)$, we can easily show

$$\begin{aligned} < F\hat{x} - b, y > &= y^T (F(F^T F)^{-1} F^T - I) b \\ &= x^T (F^T F (F^T F)^{-1} F^T - F^T) b = 0. \end{aligned}$$

$F\hat{x}$ is called the orthogonal projection of b onto the space $Im(F)$, and can be noted as

$$F\hat{x} = E^{Im(F)} b.$$

$\Rightarrow$

Orthogonal projection theorem

Suppose H is a Hilbert space, $b \in H$, and Y a subspace of H. Then $\hat{y} = E^Y b$, or

$$\min_{y \in Y} \|b - y\|^2 = \|b - \hat{y}\|^2$$

if and only if

$$< b - \hat{y}, y >= 0, \ \forall y \in Y.$$

Lemma If Y_1 is orthogonal to Y_2, then

$$E^{Y_1 \oplus Y_2} b = E^{Y_1} b + E^{Y_2} b.$$

Orthogonal projection in function space

Orthogonal projection theorem holds even for infinite-dimensional Hilbert Spaces. For example, the space of square integrable functions $L^2[a, b]$. For the space of random variables with finite second moments, with

$$\|x\|^2 = E\{x^2\}, \ < x, y >= E\{xy\},$$

it is also an infinite-dimensional Hilbert Space.

Now suppose for a variable (function) x in H, several independent observations (functions) $y_1, \cdots, y_m$ in H are given. Let $Y = span\{y_1, \cdots, y_m\} \in H$. Obviously the best approximation of x by the observations is $E^Y x$.

Let us first consider x as a scalar.

$$E^Y x = \min_k \|x - ky\|^2 = \min_k \|x - y^T k^T\|^2,$$

where $y = [y_1 \cdots y_m]^T, \ k = [k_1 \cdots k_m]$.

Following the previous discussion, we have

$$k^* y = [y^T (y \cdot y^T)^{-1} y \cdot x]^T = x \cdot y^T (y \cdot y^T)^{-1} y,$$

where "$\cdot$" denotes component-wise inner product.

If $x = [x_1, \cdots, x_n]^T$ has n components, we just do the projection component-wise. $\Rightarrow$

$$E^Y x = K^* y = x \cdot y^T (y \cdot y^T)^{-1} y.$$

Note: For $L^2[a, b]$, $x \cdot y = \int_a^b x(t) y(t) dt$. For the space of random variables, $x \cdot y = E\{xy\}$.

Kalman gain

Now let's go back to the Kalman filter problem.

Suppose $y_1(0), \cdots, y_m(0), \cdots, y_1(t), \cdots, y_m(t)$ are the observations available at t (remember each of these is a random variable). Let H_t denote the space spanned by these variables. Apparently

$$\hat{x}_t(t) = E^{H_t} x(t),$$

where the subscript indicates the latest time an observation is made. Similarly

$$\hat{x}_{t-1}(t) = E^{H_{t-1}} x(t).$$

Now we show we can calculate $\hat{x}_t(t)$ recursively!

Lemma Let H be a **finite** dimensional Hilbert space, then $E^H Px = PE^H x$.

Proof: Let $\{y_1, \cdots, y_N\}$ be a basis of H. Then,

$$E^H Px = (Px) \cdot y^T (y \cdot y^T)^{-1} y.$$

Since $Px = [\sum_{j=1}^n p_{1j} x_j, \cdots, \sum_{j=1}^n p_{mj} x_j]^T$,

$$(Px) \cdot y^T = [(\sum_{j=1}^n p_{1j} x_j \cdot y^T)^T, \cdots, (\sum_{j=1}^n p_{mj} x_j \cdot y^T)^T]^T = P(x \cdot y^T)$$

Thus, $E^H Px = PE^H x$.

Let $\tilde{y}(t) = y(t) - E^{H_{t-1}} y(t)$, then by the orthogonal projection theorem, $\tilde{y}(t)$ is orthogonal to H_{t-1}. Thus,

$$H_t = H_{t-1} \oplus [\tilde{y}(t)],$$

where $[\tilde{y}(t)]$ denotes the space spanned by components of $\tilde{y}(t)$.

$\Rightarrow$

Kalman filter

$$\begin{aligned}
\hat{x}_t(t) &= E^{H_t} x(t) = E^{H_{t-1}} x(t) + E^{[\tilde{y}(t)]} x(t) \\
&= \hat{x}_{t-1}(t) + K(t)\tilde{y}(t) \\
&= \hat{x}_{t-1}(t) + K(t)(y(t) - C\hat{x}_{t-1}(t)),
\end{aligned}$$

since $\tilde{y}(t) = y(t) - E^{H_{t-1}}(Cx(t) + Dw(t)) = y(t) - CE^{H_{t-1}} x(t)$.

Now we have finally shown (rigorously) that the optimal estimation can be obtained by linear recursion!

From this point on, there are several methods for deriving the optimal $K(t)$, besides the orthogonal projection method originally used by Kalman.

Let $e(t) = x(t) - \hat{x}_{t-1}(t) - K(t)(y(t) - C\hat{x}_{t-1})$, one method is to solve

$$\min_{K(t)} E\|e(t)\|^2 = \min_{K(t)} tr\ E\{e(t)e^T(t)\} = \min_{K(t)} tr\ P_t(t).$$

Denote $E\{(x(t) - \hat{x}_{t-1}(t))(x(t) - \hat{x}_{t-1}(t))^T\}$ by $P_{t-1}(t)$. With considerable hindsight, let

$$K(t) = P_{t-1}(t)C^T(CP_{t-1}(t)C^T + DRD^T)^{-1} + \tilde{K},$$

we have

$$\begin{aligned} trP_t(t) &= tr[P_{t-1}(t) - P_{t-1}(t)C^T(CP_{t-1}(t)C^T + DRD^T)^{-1}CP_{t-1}(t) \\ &+ \tilde{K}(CP_{t-1}(t)C^T + DRD^T)\tilde{K}^T]. \end{aligned}$$

$\Rightarrow \tilde{K} = 0$ gives the optimal K!

Once $\hat{x}_t(t)$ is obtained, the best prediction for $x(t+1)$ based on observations up to t can be derived as

$$\hat{x}_t(t+1) = E^{H_t}x(t+1) = E^{H_t}(Ax(t) + Bw(t)) = A\hat{x}_t(t).$$

Accordingly,

$$e_t(t+1) = x(t+1) - \hat{x}_t(t+1) = Ae_t(t) + Bw(t)$$

and

$$P_t(t+1) = E(e_t(t+1)e_t^T(t+1)) = AP_t(t)A^T + BQB^T.$$

Kalman filter in summary

Kalman filter consists of two phases.

Measurement update (correction):

$$\begin{aligned} x_t(t) &= \hat{x}_{t-1}(t) + K(t)(y(t) - C\hat{x}_{t-1}(t)) \\ P_t(t) &= P_{t-1}(t) - K(t)CP_{t-1}(t) \\ K(t) &= P_{t-1}(t)C^T(CP_{t-1}(t)C^T + DRD^T)^{-1}. \end{aligned}$$

After the update, we always have

$$P_t(t) \leq P_{t-1}(t).$$

We note that $K(t)$ is defined slightly different from the compendium.

Time update (prediction):

$$\begin{aligned} \hat{x}_t(t+1) &= A\hat{x}_t(t) \\ P_t(t+1) &= AP_t(t)A^T + BQB^T. \end{aligned}$$

Kalman filter and classical parameter estimation

Consider the problem of estimating parameter x from the observations

$$y = Ax + v,$$

where $E\{vv^T\} = V$. We wish to find the *linear, unbiased, minimum variance* estimator $\hat{x}^*$. Namely, in the class of $\hat{x} = Ky$, and $E\{\hat{x}\} = E\{x\}$, we have

$$E\{(x - \hat{x}^*)^T(x - \hat{x}^*)\} \rightarrow \ \min.$$

The Gauss-Markov theorem (see also the notes by Trygger) tells us

$$\hat{x}^* = \mathcal{I}^{-1}A^TV^{-1}y,$$

where $\mathcal{I} = A^TV^{-1}A$ is called the *information matrix*.

Now the interesting question is how this compares with Kalman filter:

$$\hat{x}_{k+1} = \hat{x}_k + P_k A^T (A P_k A^T + V)^{-1} (y - A\hat{x}_k).$$

We can view x_k and P_k as the priori information we have on x. Rewrite

$$\begin{aligned} \hat{x}_{k+1} &= P_k A^T (A P_k A^T + V)^{-1} y + [I - P_k A^T (A P_k A^T + V)^{-1} A] \hat{x}_k \\ &= [P_k^{-1} + \mathcal{I}]^{-1} A^T V^{-1} y + [P_k^{-1} + \mathcal{I}]^{-1} P_K^{-1} \hat{x}_k. \end{aligned}$$

Here we have used the equalities
$PA^T[APA^T + V]^{-1} = [I + PV^{-1}A]^{-1} PA^T V^{-1}$ and
$I - PA^T(APA^T + V)^{-1}A = [I + PA^T V^{-1} A]^{-1}$

Conclusion: When $P_0^{-1} = 0$, Kalman filter is the same as Gauss-Markov estimation!

Continuous time Kalman filter

Now consider

$$\begin{aligned} \dot{x} &= Ax(t) + Bv(t) \\ y &= Cx + Dw(t) \\ x(0) &= x_0 \, (unknown), \end{aligned}$$

Here we assume all matrices are time invariant. $x_0, \ w, \ v$ are pairwise uncorrelated with zero mean and further more $w, \ v$ are white noises with

$$Ev(t)v(s)^T = Q\delta(t-s), \ Ew(t)w(s)^T = R\delta(t-s).$$

Heuristically (we do not intend to be very rigorous!), we can understand the white noises as the derivatives of some *Brownian motion* (although this derivative does not exist in a conventional sense). For example,

$$\int_s^t w(r)dr = \beta(t) - \beta(s),$$

where

$$E(\beta(t) - \beta(s)) = 0, \ E(\beta(t) - \beta(s))(\beta(t) - \beta(s))^T = R(t-s), \ t > s.$$

Thus,

$$\int_s^t R dr = E \int_s^t w(r)dr \int_s^t w^T(\tau)d\tau.$$

Then,

$$\int_s^t (\int_s^t E\{w(r)w^T(\tau)d\tau - R\})dr = 0.$$

Since this is true for any interval, we have

$$\int_s^t E\{w(r)w^T(\tau)\}d\tau = R, \ \forall r \in [t, s].$$

Thus,

$$E\{w(r)w^T(\tau)\} = R\delta(r - \tau).$$

We can derive $E\{w(t)\} = 0$ similarly.

Now we use the discrete Kalman filter to derive the continuous one (by letting $\Delta t \to 0$.)

Let $x(t+1) = x(t+\Delta t)$, when Δt is very small, we have ("$\approx$" means equal up to $\mathcal{O}(\Delta t^2)$)

$$A_d = e^{A\Delta t} \approx I + A\Delta t, \ \ C_d = C.$$

Since $v_d(t) = \int_t^{t+\Delta t} e^{A(t+\Delta t - s)} v(s) ds \approx \beta(t+\Delta t) - \beta(t)$,

$$Q_d \approx Q\Delta t, \ \ R_d \approx R/\Delta t.$$

Then, $K_d(t) \approx P(t)C^T(DRD^T)^{-1}\Delta t$. We have

$$\hat{x}(t+1) \approx (I + A\Delta t)(\hat{x}(t) + K_d(t)(y(t) - C\hat{x}(t)),$$

or,

$$\hat{x}(t+1) - \hat{x}(t) \approx A\hat{x}(t)\Delta t + P(t)C^T(DRD^T)^{-1}(y(t) - C\hat{x}(t))\Delta t.$$

Thus, by dividing both sides with Δt and taking the limit, we have

$$\dot{\hat{x}}(t) = A\hat{x}(t) + K(t)(y(t) - C\hat{x}(t)),$$

where $K(t) = P(t)C^T(DRD^T)^{-1}$.

Since

$$\begin{aligned} & P(t+\Delta t) \approx (I + A\Delta t)(I - K_d C)P(t)(I + A\Delta t)^T + BQ\Delta t B^T \\ \approx \ \ & P(t) + (AP(t) + P(t)A^T)\Delta t - P(t)C^T(DRD^T)^{-1}CP(t)\Delta t + BQB^T\Delta t \end{aligned}$$

Similarly, we obtain

$$\dot{P}(t) = AP(t) + P(t)A^T - P(t)C^T(DRD^T)^{-1}CP(t) + BQB^T,$$

and we assume $P(0) = P_0$ is known.

Steady-state Kalman filter

When (A, B) is controllable and (C, A) is observable, we know that $P(t)$ has a limit as $t \to \infty$ (remember $DRD^T > 0, \ Q > 0$):

$$AP_\infty + P_\infty A^T - P_\infty C^T(DRD^T)^{-1}CP_\infty + BQB^T = 0.$$

What this implies is that the rate the information comes in, $PC^T(DRD^T)^{-1}CP$ (less noise means better quality), is just balanced by the rate the information diffuses from the system, BQB^T (smaller diffusion means less loss), and by any damping or amplification the system may have.

In practice, we may use P_∞ to compute the Kalman gain K.