Another approach for the Kalman filter

Nguyen Duc Thanh

(Spring 2022)

1 Introduction

In section II.3 of the lecture notes, we have seen that the Kalman filter formulas were derived using Bayes' theorem, where the posterior probability (the analysis) was given in terms of the likelihood and the prior probability (the forecast). Here, we will see how these formulas also arise when we apply the Best Linear Unbiased Estimator on the system to obtain a posterior estimate.

2 Model

Let $x^n \in \mathbb{R}^N$ and $y^n \in \mathbb{R}^M$ be the random variables representing the state of the system and the observation at time n. Then, the evolution of the system and the relation between the state and the observation are assumed to be following the set of equations

$$x^{n+1} = F^n x^n + \xi^n, \tag{1}$$
$$y^n = H^n x^n + \varepsilon^n, \tag{2}$$

where $F^n \in M_{NN}(\mathbb{R})$ and $H^n \in M_{MN}(\mathbb{R})$. The errors in the model and the observation are represented by the random variables ξ^n and ε^n, which take values in $\mathbb{R}^N$ and $\mathbb{R}^M$, respectively. For the problem at hand, these errors are assumed to be independent, white-noise processes with Gaussian probability distributions

$$\xi^n \sim \mathcal{N}\left(\mathbf{0}, Q^n\right),$$
$$\epsilon^n \sim \mathcal{N}\left(\mathbf{0}, R^n\right),$$

where $Q^n \in M_{NN}(\mathbb{R})$ and $R^n \in M_{MM}(\mathbb{R})$ are the covariance matrices. We will denote the true state of the system at time n (which is unknown) by $x^{n,t}$. In the subsequent section, the superscripts on F^n, H^n, Q^n, R^n will be omitted for brevity, but we should keep in mind that they are still time-dependent quantities.

3 Derivation

Consider the system starting out at time $n-1$ and in the state x^{n-1}. We let the system evolve according to equation (1) to get the forecast $x^{n,f}$. If the observation is available at this time, meaning that y^n is known, we can do a "correction" step to produce an analysis $x^{n,a}$. Our goal is to find the best analysis based on these forecast and observation.

Let us define the forecast and analysis estimate errors

$$\begin{aligned} e^{n,f} &= x^{n,f} - x^{n,t}, \\ e^{n,a} &= x^{n,a} - x^{n,t}, \end{aligned}$$

as well as their respective error covariance matrices (with the superscripts for the time n being omitted)

$$\begin{aligned} P^f &= \mathbb{E}\left[e^{n,f}(e^{n,f})^T\right], \\ P^a &= \mathbb{E}\left[e^{n,a}(e^{n,a})^T\right]. \end{aligned}$$

For the Kalman filter, we look for an optimal analysis $x^{n,a}$ that is a linear combination of the forecast $x^{n,f}$ and a weighted difference between the observation y^n and the observation prediction $Hx^{n,f}$. In other words, we have

$$x^{n,a} = x^{n,f} + K(y^n - Hx^{n,f}),$$

where $K \in M_{NM}(\mathbb{R})$ is called the Kalman gain matrix (the superscript n has also been omitted). The discrepancy between the actual and the predicted observations is reflected in the difference $d^n = y^n - Hx^{n,f}$ called the innovation. We want to determine an optimal K that "minimizes" the analysis error covariance matrix P^a. In fact, $x^{n,a}$ is the Best Linear Unbiased Estimator (one can check the report on BLUE by Ethan Dowley for more details).

Before defining what does "minimizing" in this case mean, we first rewrite P^a to see its dependence on K

$$\begin{aligned} P^a &= \mathbb{E}\left[e^{n,a}\left(e^{n,a}\right)^T\right] \\ &= \mathbb{E}\left[(x^{n,a} - x^{n,t})(x^{n,a} - x^{n,t})^T\right] \\ &= \mathbb{E}\left[(x^{n,f} + Kd^n - x^{n,t})(x^{n,f} + Kd^n - x^{n,t})^T\right] \\ &= \mathbb{E}\left[(e^{n,f} + Kd^n)(e^{n,f} + Kd^n)^T\right] \\ &= \mathbb{E}\left[e^{n,f}(e^{n,f})^T\right] + \mathbb{E}\left[e^{n,f}(d^n)^TK^T + Kd^n(e^{n,f})^T + Kd^n(d^n)^TK^T\right] \\ &= P^f + \mathbb{E}\left[e^{n,f}(d^n)^TK^T + Kd^n(e^{n,f})^T + Kd^n(d^n)^TK^T\right], \end{aligned}$$

where we have used $e^{n,f} = x^{n,f} - x^{n,t}$, $\mathbb{E}\left[e^{n,f}(e^{n,f})^T\right] = P^f$ and the linearity of the expectation value. From equation (2), we have

$$\begin{aligned} d^n &= y^n - Hx^{n,f} \\ &= Hx^{n,t} + \varepsilon^n - Hx^{n,f} \\ &= \varepsilon^n - He^{n,f}. \end{aligned}$$

Then, the second term in the expression for P^a can be expanded as

$$\begin{aligned}
&\mathbb{E}\left[e^{n,f}(d^n)^T K^T + K d^n (e^{n,f})^T + K d^n (d^n)^T K^T\right] \\
&= \mathbb{E}\left[e^{n,f}(\varepsilon^n - He^{n,f})^T K^T + K(\varepsilon^n - He^{n,f})(e^{n,f})^T + K(\varepsilon^n - He^{n,f})(\varepsilon^n - He^{n,f})^T K^T\right] \\
&= \mathbb{E}\Big[e^{n,f}(\varepsilon^n)^T K^T - e^{n,f}(e^{n,f})^T H^T K^T + K\varepsilon^n (e^{n,f})^T - KHe^{n,f}(e^{n,f})^T \\
&\quad + K\varepsilon^n(\varepsilon^n)^T K^T - KHe^{n,f}(\varepsilon^n)^T K^T - K\varepsilon^n(e^{n,f})^T H^T K^T + KHe^{n,f}(e^{n,f})^T H^T K^T\Big] \\
&= \mathbb{E}\big[e^{n,f}(\varepsilon^n)^T\big] K^T - P^f H^T K^T + K\mathbb{E}\big[\varepsilon^n (e^{n,f})^T\big] - KHP^f \\
&\quad + KRK^T - KH\mathbb{E}\big[e^{n,f}(\varepsilon^n)^T\big] K^T - K\mathbb{E}\big[\varepsilon^n(e^{n,f})^T\big] H^T K^T + KHP^f H^T K^T \\
&= \Big(- P^f H^T K^T - KHP^f + K(HP^f H^T + R)K^T \Big) \\
&\quad + \Big(\mathbb{E}\big[e^{n,f}(\varepsilon^n)^T\big] K^T + K\mathbb{E}\big[\varepsilon^n(e^{n,f})^T\big] - KH\mathbb{E}\big[e^{n,f}(\varepsilon^n)^T\big] K^T - K\mathbb{E}\big[\varepsilon^n(e^{n,f})^T\big] H^T K^T \Big),
\end{aligned}$$

where we have also used $R = \mathbb{E}\big[\varepsilon^n(\varepsilon^n)^T\big]$ and the fact that K and H are constant matrices (at a given time n) so they go outside the expectation values. Moreover, each component of the observation error ε^n and that of the forecast estimation error $e^{n,f}$ are uncorrelated because the forecast at time n can only depend on the observation made in the preceding time, i.e., for time $\leq n-1$. As a result, the expectation value of their product can be factorized, so we get $\forall i, j$ with $1 \leq i \leq M$ and $1 \leq j \leq N$

$$\mathbb{E}\left[(\varepsilon^n)_i (e^{n,f})_j\right] = \mathbb{E}\left[(\varepsilon^n)_i\right] \mathbb{E}\left[(e^{n,f})_j\right] = 0,$$

from the assumption on the mean of the observation error $\mathbb{E}\left[\varepsilon^n\right] = \mathbf{0}$. Hence, we have the following expression for P^a that is valid for any K

$$P^a = P^f - P^f H^T K^T - KHP^f + K(HP^f H^T + R)K^T.$$

Since P^a is symmetric, we can always find an eigenbasis that diagonalizes the matrix. Working in these (principal) axes, the diagonal components of P^a represent the estimation error variances which we want to minimize. Letting P^a in this basis be denoted by P^a_{diag}, our goal is then to minimize its trace which is the sum of the diagonal components. However, P^a is symmetric so there exists a change of basis matrix $S \in M_{NN}(\mathbb{R})$ that satisfies

$$P^a_{diag} = S^{-1} P^a S,$$

and is also orthogonal ($S^{-1} = S^T$). Then, we have

$$\begin{aligned}
Tr(P^a_{diag}) &= Tr(S^{-1}P^a S) \\
&= \sum_{i=1}^{N} (S^{-1}P^a S)_{ii} \\
&= \sum_{i=1}^{N} \left[\sum_{j=1}^{N} \sum_{k=1}^{N} S^{-1}_{ij} P^a_{jk} S_{ki} \right]
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{k=1}^{N}\Big[\sum_{i=1}^{N}\sum_{j=1}^{N} S_{ki}S^{-1}_{ij}P^a_{jk}\Big] \\
&= \sum_{k=1}^{N}(SS^{-1}P^a)_{kk} \\
&= \sum_{k=1}^{N} P^a_{kk} \\
&= Tr(P^a).
\end{aligned}
$$

This means that the trace is invariant under this change of basis (in fact, this is true for general P^a and S) so we need not to minimize the trace in its eigenbasis and instead, work directly with $Tr(P^a)$.

Next, we will consider an infinitesimal change in K and expand $Tr(P^a)$ to first order

$$
Tr(P^a)[K+\Delta K] \approx Tr(P^a)[K] + \sum_{i=1}^{N}\sum_{j=1}^{M} \frac{dTr(P^a)}{dK_{ij}}\bigg|_K (\Delta K)_{ij}.
$$

When K is the (local) minimum point, the value of $Tr(P^a)$ increases for any small change ΔK, so we have

$$
\sum_{i=1}^{N}\sum_{j=1}^{M} \frac{dTr(P^a)}{dK_{ij}}\bigg|_K (\Delta K)_{ij} \geq 0.
$$

Since we take arbitrary ΔK, $(\Delta K)_{ij}$ can either be smaller, equal, or greater than 0. The only way for the above inequality to hold is

$$
\frac{dTr(P^a)}{dK_{ij}}\bigg|_K = 0 \qquad \forall i, j.
$$

We will now prove two results from matrix differentiation that will be useful for our problem. Let us consider the matrices $A \in M_{ab}(\mathbb{R})$ and $B \in M_{ba}(\mathbb{R})$, and a symmetric matrix $C \in M_{bb}(\mathbb{R})$. This means that AB, ACA^T are all square matrices so we can define their traces. We will show that $\forall i, j$ with $1 \leq i \leq a$ and $1 \leq j \leq b$, we get

$$
\begin{aligned}
&\frac{d}{dA_{ij}}Tr(AB) = B_{ji}, \\
&\frac{d}{dA_{ij}}Tr(ACA^T) = 2(AC)_{ij}.
\end{aligned}
$$

For the first result, we have

$$
\frac{d}{dA_{ij}}Tr(AB) = \frac{d}{dA_{ij}}\sum_{k=1}^{a}\sum_{l=1}^{b} A_{kl}B_{lk} = \sum_{k=1}^{a}\sum_{l=1}^{b} \frac{dA_{kl}}{dA_{ij}}B_{lk},
$$

where each component of A is considered as an independent variable. By using the Kronecker delta notation

$$
\delta_{\alpha\beta} = \begin{cases} 1 & \text{if } \alpha = \beta, \\ 0 & \text{otherwise,} \end{cases}
$$

we get

$$\frac{dA_{kl}}{dA_{ij}} = \delta_{ki}\delta_{lj}.$$

Then, the above equation reads

$$\frac{d}{dA_{ij}}Tr(AB) = \sum_{k=1}^{a}\sum_{l=1}^{b}\delta_{ki}\delta_{lj}B_{lk} = B_{ji}.$$

Similarly, for the second one, we have

$$\begin{aligned}
\frac{d}{dA_{ij}}Tr(ACA^T) &= \frac{d}{dA_{ij}}\sum_{k=1}^{a}\sum_{l=1}^{b}\sum_{m=1}^{b}A_{kl}C_{lm}(A^T)_{mk} \\
&= \frac{d}{dA_{ij}}\sum_{k=1}^{a}\sum_{l=1}^{b}\sum_{m=1}^{b}A_{kl}C_{lm}A_{km} \\
&= \sum_{k=1}^{a}\sum_{l=1}^{b}\sum_{m=1}^{b}\frac{d}{dA_{ij}}\big(A_{kl}C_{lm}A_{km}\big) \\
&= \sum_{k=1}^{a}\sum_{l=1}^{b}\sum_{m=1}^{b}\left(\frac{dA_{kl}}{dA_{ij}}C_{lm}A_{km} + A_{kl}C_{lm}\frac{dA_{km}}{dA_{ij}}\right) \\
&= \sum_{k=1}^{a}\sum_{l=1}^{b}\sum_{m=1}^{b}\Big(\delta_{ki}\delta_{lj}C_{lm}A_{km} + A_{kl}C_{lm}\delta_{ki}\delta_{mj}\Big) \\
&= \sum_{m=1}^{b}A_{im}C_{jm} + \sum_{l=1}^{b}A_{il}C_{lj}.
\end{aligned}$$

Since C is assumed to be symmetric, $C_{jm} = C_{mj}$. Furthermore, m and l are dummy indices so we can combine them as

$$\frac{d}{dA_{ij}}Tr(ACA^T) = \sum_{m=1}^{b}A_{im}C_{mj} + \sum_{l=1}^{b}A_{il}C_{lj} = 2\sum_{m=1}^{b}A_{im}C_{mj} = 2(AC)_{ij}.$$

These two results give us $2 \times a \times b$ equations which can summarized in matrix form as

$$\begin{aligned}
\frac{d}{dA}Tr(AB) &= B^T, \\
\frac{d}{dA}Tr(ACA^T) &= 2AC.
\end{aligned}$$

We then return to our problem

$$\begin{aligned}\frac{dTr(P^a)}{dK} &= \frac{dTr(P^f)}{dK} - \frac{dTr(P^fH^TK^T)}{dK} - \frac{dTr(KHP^f)}{dK} + \frac{dTr\Big(K(HP^fH^T+R)K^T\Big)}{dK} \\ &= 0 - \frac{dTr(P^fH^TK^T)}{dK} - \frac{dTr(KHP^f)}{dK} + \frac{dTr\Big(K(HP^fH^T+R)K^T\Big)}{dK},\end{aligned}$$

where we have used the linearity of the trace due to it being a sum of the diagonal components. Another consequence of the fact that only the diagonal components are involved is that the trace of a (square) matrix is equal to the trace of its transpose. We also note that P^f and R are symmetric, so $(P^f)^T = P^f$, $R^T = R$, and, hence, $(HP^fH^T + R)^T = (H^T)^T(P^f)^TH^T + R^T = HP^fH^T + R$. Therefore, we have

$$\begin{aligned}\frac{dTr(P^a)}{dK} &= -\frac{dTr\big((P^fH^TK^T)^T\big)}{dK} - \frac{dTr(KHP^f)}{dK} + \frac{dTr\Big(K(HP^fH^T+R)K^T\Big)}{dK} \\ &= -2\frac{d}{dK}Tr(KHP^f) + \frac{d}{dK}Tr\Big(K(HP^fH^T+R)K^T\Big) \\ &= -2(HP^f)^T + 2K(HP^fH^T+R),\end{aligned}$$

where the two previously mentioned results were used in the last step. Setting this expression to zero gives us

$$-2(HP^f)^T + 2K(HP^fH^T+R) = -2P^fH^T + 2K(HP^fH^T+R) = \mathbf{0} \in M_{NM}(\mathbb{R}).$$

Equivalently, this reads

$$\boxed{K = P^fH^T(HP^fH^T+R)^{-1}}\,.$$

We make a small remark that $HP^fH^T + R$ is the covariance matrix of the innovation d^n. This can be shown as follows

$$\begin{aligned}\mathbb{E}\big[d^n(d^n)^T\big] &= \mathbb{E}\big[(\varepsilon^n - He^{n,f})(\varepsilon^n - He^{n,f})^T\big] \\ &= \mathbb{E}\big[\varepsilon^n(\varepsilon^n)^T - He^{n,f}(\varepsilon^n)^T - \varepsilon^n(e^{n,f})^TH^T + He^{n,f}(e^{n,f})^TH^T\big] \\ &= \mathbb{E}\big[\varepsilon^n(\varepsilon^n)^T\big] - H\mathbb{E}\big[e^{n,f}(\varepsilon^n)^T\big] - \mathbb{E}\big[\varepsilon^n(e^{n,f})^T\big]H^T + H\mathbb{E}\big[e^{n,f}(e^{n,f})^T\big]H^T \\ &= R - 0 - 0 + HP^fH^T \\ &= HP^fH^T + R.\end{aligned}$$

To understand how the gain matrix affects the analysis, we consider the two limits: $R \to 0$ and $P^f \to 0$. For the case of $R \to 0$, we assume that H is square and invertible, so there exists H^{-1} such that $H^{-1}H = I$; then, H^{-1} gives us the state from the observation. This is a strong assumption because in general, we cannot infer the whole state just from the available information (i.e., the observation), but since we only look for the qualitative effect of K, this assumption is justified. We get these limits for K as

$$
\begin{aligned}
\lim_{R\to 0} K &= P^f H^T (H P^f H^T)^{-1} \\
&= (H^{-1}H) P^f H^T (H P^f H^T)^{-1} \\
&= H^{-1}(H P^f H^T)(H P^f H^T)^{-1} \\
&= H^{-1}, \\
\lim_{P^f\to 0} K &= 0.
\end{aligned}
$$

The analysis then becomes

$$
\begin{aligned}
\lim_{R\to 0} x^{n,a} &= x^{n,f} + H^{-1}(y^n - Hx^{n,f}) \\
&= x^{n,f} + H^{-1}y^n - x^{n,f} \\
&= H^{-1}y^n \\
\lim_{P^f\to 0} x^{n,a} &= x^{n,f}.
\end{aligned}
$$

When the error in the observation represented by R approaches zero, the observation is trusted more, whereas if the forecast error estimate covariance matrix P^f approaches zero, the prediction from the model is trusted more; hence, we get the above results.

With the optimal gain derived, we can also calculate the analysis error estimate covariance matrix as follows

$$
\begin{aligned}
P^a &= P^f - P^f H^T K^T - K H P^f + K(H P^f H^T + R)K^T \\
&= P^f - P^f H^T \big[P^f H^T (H P^f H^T + R)^{-1}\big]^T - \big[P^f H^T (H P^f H^T + R)^{-1}\big] H P^f \\
&\quad + \big[P^f H^T (H P^f H^T + R)^{-1}\big](H P^f H^T + R)\big[P^f H^T (H P^f H^T + R)^{-1}\big]^T \\
&= P^f - P^f H^T \big[P^f H^T (H P^f H^T + R)^{-1}\big]^T - P^f H^T (H P^f H^T + R)^{-1} H P^f \\
&\quad + P^f H^T \big[P^f H^T (H P^f H^T + R)^{-1}\big]^T \\
&= P^f - P^f H^T (H P^f H^T + R)^{-1} H P^f.
\end{aligned}
$$

This is an explicit expression for the covariance matrix, but we could equally well use the simpler form that includes the optimal gain

$$
\boxed{P^a = P^f - K H P^f}\ .
$$

To finish our derivation, we will calculate how the forecast is updated given the previous analysis. In other words, if $x^{n,a}$ and $P^{n,a}$ are known, what will $x^{n+1,f}$ and $P^{n+1,f}$ be? The true state is unknown in general, so we can only estimate it to be the mean of the analysis (since the true state is not a random variable), i.e. $x^{n,t} = \bar{x}^{n,a}$ where we have also denoted the mean $\mathbb{E}[x]$ by $\bar{x}$. From equation (1), we obtain the forecast and the estimated true state at time $n+1$

$$
\begin{aligned}
x^{n+1,f} &= F^n x^{n,a} + \xi^n, \\
x^{n+1,t} &= F^n \bar{x}^{n,a}.
\end{aligned}
$$

Taking the expectation value of the first equation and using the assumption that $\mathbb{E}[\xi^n] = \mathbf{0}$, we get

$$\bar{x}^{n+1,f} = x^{n+1,t} = F^n \bar{x}^{n,a} = F^n x^{n,t}.$$

The forecast estimate covariance matrix can then be determined

$$\begin{aligned}
P^{n+1,f} &= \mathbb{E}\big[e^{n+1,f}(e^{n+1,f})^T\big] \\
&= \mathbb{E}\big[(x^{n+1,f} - x^{n+1,t})(x^{n+1,f} - x^{n+1,t})^T\big] \\
&= \mathbb{E}\big[(F^n x^{n,a} + \xi^n - F^n x^{n,t})(F^n x^{n,a} + \xi^n - F^n x^{n,t})^T\big] \\
&= \mathbb{E}\big[(F^n e^{n,a} + \xi^n)(F^n e^{n,a} + \xi^n)^T\big] \\
&= F^n \mathbb{E}\big[(e^{n,a})(e^{n,a})^T\big](F^n)^T + \mathbb{E}\big[\xi^n (e^{n,a})^T\big](F^n)^T + F^n \mathbb{E}\big[e^{n,a}(\xi^n)^T\big] + \mathbb{E}\big[\xi^n(\xi^n)^T\big] \\
&= F^n P^{n,a}(F^n)^T + \mathbb{E}\big[\xi^n (e^{n,a})^T\big](F^n)^T + F^n \mathbb{E}\big[e^{n,a}(\xi^n)^T\big] + Q^n,
\end{aligned}$$

where we have used $e^{n,a} = x^{n,a} - x^{n,t}$ and $\mathbb{E}\big[\xi^n(\xi^n)^T\big] = Q^n$. We notice that the analysis error $e^{n,a}$ depends on the forecast at time n, which itself can only depend on the model errors ξ^t for $t \leq n-1$ (as shown by the dynamical equations). This implies that $e^{n,a}$ and ξ^n are uncorrelated, so we again have

$$\mathbb{E}\big[\xi^n(e^{n,a})^T\big] = \mathbf{0} \in M_{NN}(\mathbb{R}) \quad \text{and} \quad \mathbb{E}\big[e^{n,a}(\xi^n)^T\big] = \mathbf{0} \in M_{NN}(\mathbb{R}).$$

Therefore, the update for the forecast estimate covariance matrix is given by

$$P^{n+1,f} = F^n P^{n,a}(F^n)^T + Q^n.$$

4 Algorithm

We conclude our problem by presenting the algorithm for the Kalman filter as follows: starting from a the previous analysis $\bar{x}^{n,a}$ and $P^{n,a}$, the state of the system is then updated in two steps

1. Forecast step:

$$\begin{aligned}
\bar{x}^{n+1,f} &= F^n \bar{x}^{n,a}, \\
P^{n+1,f} &= F^n P^{n,a}(F^n)^T + Q^n
\end{aligned}$$

2. Analysis step:

$$\begin{aligned}
K^{n+1} &= P^{n+1,f}(H^{n+1})^T\big[H^{n+1}P^{n+1,f}(H^{n+1})^T + R^{n+1}\big]^{-1}, \\
\bar{x}^{n+1,a} &= \bar{x}^{n+1,f} - K^{n+1}(y^{n+1} - H^{n+1}\bar{x}^{n+1,f}) \\
P^{n+1,a} &= P^{n+1,f} - K^{n+1}H^{n+1}P^{n+1,f}
\end{aligned}$$

5 Reference

1. Introduction to data assimilation, lecture notes by Serge Richard.
1. [ABN] Data assimilation: methods, algorithms, and applications, by M. Asch, M. Bocquet, and M. Nodet.
2. Best Linear Unbiased Estimator, report by Ethan Dowley.