Technion—Israel Intitute of Technology
The Erna and Andrew Viterbi Department of Electrical Engineering

Estimation and Identification in Dynamical Systems (048825)
Lecture Notes: Prof. Nahum Shimkin, Spring 2016

4 Derivations of the Discrete-Time Kalman Filter

We derive here the basic equations of the Kalman filter (KF), for discrete-time linear systems. We consider several derivations under different assumptions and viewpoints:

- For the Gaussian case, the KF is the optimal (MMSE) state estimator.
- In the non-Gaussian case, the KF is derived as the best linear (LMMSE) state estimator.
- We also provide a deterministic (least-squares) interpretation.

We start by describing the basic state-space model.

Contents

4.1 The Stochastic State-Space Model

4.2 The KF for the Gaussian Model

4.3 Best Linear Estimator – Innovations Approach

4.4 The KF as a Least-Squares Problem

4.5 The KF Equations – Summary

4.1 The Stochastic State-Space Model

A discrete-time, linear, time-varying state space system is given by:

$$\begin{aligned} x_{k+1} &= F_k x_k + G_k w_k \qquad &\text{(state evolution equation)} \\ z_k &= H_k x_k + v_k \qquad &\text{(measurement equation)} \end{aligned}$$

for $k \geq 0$ (say), and initial conditions x_0.
Here:

- F_k, G_k, H_k are known matrices.
- $x_k \in I\!R^n$ is the state vector.
- $w_k \in I\!R^{n_w}$ is the state noise.
- $z_k \in I\!R^m$ is the observation vector.
- v_k the observation noise.
- The initial conditions are given by x_0, usually a random variable.

The noise sequences (w_k, v_k) and the initial conditions x_0 are stochastic processes with known distributions.

The Markovian model

Recall that a stochastic process $\{X_k\}$ is a *Markov* process if

$$p(X_{k+1}|X_k, X_{k-1}, \dots) = p(X_{k+1}|X_k)\,.$$

For the state x_k to be Markovian, we need the following assumption.

Assumption A1: The state-noise process $\{w_k\}$ is *white in the strict sense*, namely all w_k's are independent of each other. Furthermore, this process is independent of x_0.

The following is then a simple exercise:

Proposition: Under A1, the state process $\{x_k, k \geq 0\}$ is a Markov process.

Note:

- Linearity is not essential: The Markov property follows from A1 also for the nonlinear state equation $x_{k+1} = f(x_k, w_k)$.

- The measurement process z_k is usually *not* Markov.
- The pdf of the state can (in principle) be computed recursively via the following (Chapman-Kolmogorov) equation:

$$p(x_{k+1}) = \int p(x_{k+1}|x_k)p(x_k)dx_k \, .$$

where $p(x_{k+1}|x_k)$ is determined by $p(w_k)$.

The Gaussian model

- Assume that the noise sequences $\{w_k\}$, $\{v_k\}$ and the initial conditions x_0 are jointly Gaussian.
- It easily follows that the processes $\{x_k\}$ and $\{z_k\}$ are (jointly) Gaussian as well.
- If, in addition, A1 is satisfied (namely $\{w_k\}$ is white and independent of x_0), then x_k is a Markov process.

This model is often called the Gauss-Markov Model.

Second-Order Model

We often assume that only the first and second order statistics of the noise is known. Consider our linear system:

$$\begin{aligned} x_{k+1} &= F_k x_k + G_k w_k\,, \quad k \geq 0 \\ z_k &= H_k x_x + v_k\,, \end{aligned}$$

under the following assumptions:

- w_k a 0-mean white noise: $E(w_k) = 0$, $\text{cov}(w_k, w_l) = Q_k \delta_{kl}$.
- v_k a 0-mean white noise: $E(v_k) = 0$, $\text{cov}(v_k, v_l) = R_k \delta_{kl}$.
- $\text{cov}(w_k, v_l) = 0$: uncorrelated noise.
- x_0 is uncorrelated with the other noise sequences.
  denote $\overline{x}_0 = E(x_0)$, $\text{cov}(x_0) = P_0$.

We refer to this model as the *standard second-order model.*

It is sometimes useful to allow correlation between v_k and w_k:

$$\text{cov}(w_k, v_l) \equiv E(w_k v_l^T) = S_k \delta_{kl}\,.$$

This gives the *second-order model with correlated noise.*

A short-hand notation for the above correlations:

$$\text{cov}(\begin{bmatrix} w_k \\ v_k \\ x_0 \end{bmatrix}, \begin{bmatrix} w_l \\ v_l \\ x_0 \end{bmatrix}) = \begin{bmatrix} Q_k \delta_{kl} & S_k \delta_{kl} & 0 \\ S_k^T \delta_{kl} & R_k \delta_{kl} & 0 \\ 0 & 0 & P_0 \end{bmatrix}$$

Note that the Gauss-Markov model is a special case of this model.

Mean and covariance propagation

For the standard second-order model, we easily obtain recursive formulas for the mean and covariance of the *state*, when no measurement is given.

- The mean obviously satisfies:

$$\overline{x}_{k+1} = F_k \overline{x}_k + G_k \overline{w}_k = F_k \overline{x}_k$$

- Consider next the covariance:

$$P_k \triangleq E((x_k - \overline{x}_k)(x_k - \overline{x})^T)\,.$$

  Note that $x_{k+1} - \overline{x}_{k+1} = F_k(x_k - \overline{x}_k) + G_k w_k$, and w_k and x_k are uncorrelated (why?). Therefore

$$P_{k+1} = F_k P_k F_k^T + G_k Q_k G_k^T\,.$$

  This equation is in the form of a *Lyapunov difference equation.*

- Since $z_k = H_k x_x + v_k$, it is now easy to compute its covariance:

$$\text{cov}(z_k) = H_k P_k H_k^T + R_k\,,$$

  and similarly for the joint covariances of (x_k, z_k).

- In the Gaussian case, the pdf of x_k is completely specified by the mean and covariance: $x_k \sim N(\overline{x}_k, P_k)$.

4.2 The KF for the Gaussian Model

Consider the linear Gaussian (or Gauss-Markov) model

$$\begin{aligned} x_{k+1} &= F_k x_k + G_k w_k\,, \quad k \geq 0 \\ z_k &= H_k x_x + v_k \end{aligned}$$

where:

- $\{w_k\}$ and $\{v_k\}$ are independent, zero-mean Gaussian white processes with covariances
$$E(v_k v_l^T) = R_k \delta_{kl}\,, \quad E(w_k w_l^T) = Q_k \delta_{kl}$$
- The initial state x_0 is a Gaussian RV, independent of the noise processes, with $x_0 \sim N(\overline{x}_0, P_0)$.

Let $Z_k = (z_0, \ldots, z_k)$. Our goal is to compute recursively the following optimal (MMSE) estimator of x_k:

$$\hat{x}_k^+ \equiv \hat{x}_{k|k} \triangleq E(x_k|Z_k)\,.$$

Also define the *one-step predictor* of x_k:

$$\hat{x}_k^- \equiv \hat{x}_{k|k-1} \triangleq E(x_k|Z_{k-1})$$

and the respective covariance matrices:

$$\begin{aligned} P_k^+ &\equiv P_{k|k} \triangleq E\{x_k - \hat{x}_k^+)(x_k - \hat{x}_k^+)^T|Z_k\} \\ P_k^- &\equiv P_{k|k-1} \triangleq E\{x_k - \hat{x}_k^-)(x_k - \hat{x}_k^-)^T|Z_{k-1}\}\,. \end{aligned}$$

Note that P_k^+ (and similarly P_k^-) can be viewed in two ways:

(i) It is the covariance matrix of the (posterior) estimation error, $e_k = x_k - \hat{x}_k^+$. In particular, $\text{MMSE}(\hat{x}_k^+) = \text{trace}(P_k^+)$.

(ii) It is the covariance matrix of the "conditional RV $(x_k|Z_k)$", namely an RV with distribution $p(x_k|Z_k)$ (since $\hat{x}_k^+$ is its mean).

Finally, denote $P_0^- \triangleq P_0, \quad \hat{x}_0^- \triangleq \overline{x}_0\,.$

Recall the formulas for conditioned Gaussian vectors:

- If X and X are jointly Gaussian, then $p_{x|z} \sim N(m, \Sigma)$, with

$$m = m_x + \Sigma_{xz}\Sigma_{zz}^{-1}(z - m_z)\,,$$
$$\Sigma = \Sigma_{xx} - \Sigma_{xz}\Sigma_{zz}^{-1}\Sigma_{zx}\,.$$

- The same formulas hold when everything is conditioned, in addition, on another random vector.

According to the terminology above, we say in this case that the conditional RV $(X|z)$ is Gaussian.

Proposition: For the model above, all random processes (noises, x_k, z_k) are jointly Gaussian.

Proof: All can be expressed as *linear* combinations of the noise seqeunces, which are jointly Gaussian (why?). □

It follows that $(x_k|Z_m)$ is Gaussian (for any k, m). In particular:

$$(x_k|Z_k) \sim N(\hat{x}_k^+, P_k^+)\,, \quad (x_k|Z_{k-1}) \sim N(\hat{x}_k^-, P_k^-)\,.$$

Filter Derivation

Suppose, at time k, that $(\hat{x}_k^-, P_k^-)$ is given.
We shall compute $(\hat{x}_k^+, P_k^+)$ and $(\hat{x}_{k+1}^-, P_{k+1}^-)$, using the following two steps.

Measurement update step: Since $z_k = H_k x_k + v_k$, then the conditional vector $(\begin{pmatrix} x_k \\ z_k \end{pmatrix} | Z_{k-1})$ is Gaussian, with mean and covariance:

$$\begin{bmatrix} \hat{x}_k^- \\ H_k \hat{x}_k^- \end{bmatrix}, \quad \begin{bmatrix} P_k^- & P_k^- H_k^T \\ H_k P_k^- & M_k \end{bmatrix}$$

where

$$M_k \triangleq H_k P_k^- H_k^T + R_k \,.$$

To compute $(x_k | Z_k) = (x_k | z_k, Z_{k-1})$, we apply the above formula for conditional expectation of Gaussian RVs, with everything pre-conditioned on Z_{k-1}. It follows that $(x_k | Z_k)$ is Gaussian, with mean and covariance:

$$\hat{x}_k^+ \triangleq E(x_k | Z_k) = \hat{x}_k^- + P_k^- H_k^T (M_k)^{-1} (z_k - H_k \hat{x}_k^-)$$

$$P_k^+ \triangleq \text{cov}(x_k | Z_k) = P_k^- - P_k^- H_k^T (M_k)^{-1} H_k P_k^-$$

Time update step Recall that $x_{k+1} = F_k x_k + G_k w_k$. Further, x_k and w_k are independent given Z_k (why?). Therefore,

$$\hat{x}_{k+1}^- \triangleq E(x_{k+1} | Z_k) = F_k \hat{x}_k^+$$

$$P_{k+1}^- \triangleq \text{cov}(x_{k+1} | Z_k) = F_k P_k^+ F_k^T + G_k Q_k G_k^T$$

Remarks:

1. The KF computes both the estimate $\hat{x}_k^+$ and its MSE/covariance P_k^+ (and similarly for $\hat{x}_k^-$).
   Note that the covariance computation is needed as part of the estimator computation. However, it is also of independent importance as is assigns a measure of the uncertainly (or confidence) to the estimate.

2. It is remarkable that the conditional covariance matrices P_k^+ and P_k^- do not depend on the measurements $\{z_k\}$. They can therefore be computed in advance, given the system matrices and the noise covariances.

3. As usual in the Gaussian case, P_k^+ is also the *unconditional* error covariance:

$$P_k^+ = \text{cov}(x_k - \hat{x}_k^+) = E[(x_k - \hat{x}_k^+)(x_k - \hat{x}_k^+)^T] \, .$$

   In the non-Gaussian case, the unconditional covariance will play the central role as we compute the LMMSE estimator.

4. Suppose we need to estimate some $s_k \triangleq C x_k$.
   Then the optimal estimate is $\hat{s}_k = E(s_k|Z_k) = C\hat{x}_k^+$.

5. The following "output prediction error"

$$\tilde{z}_k \triangleq z_k - H_k \hat{x}_k^- \equiv z_k - E(z_k|Z_{k-1})$$

   is called the *innovation*, and $\{\tilde{z}_k\}$ is the important *innovations process*.
   Note that $M_k = H_k P_k^- H_k^T + R_k$ is just the covariance of $\tilde{z}_k$.

4.3 Best Linear Estimator – Innovations Approach

a. Linear Estimators

Recall that the best linear (or LMMSE) estimator of X given Y is an estimator of the form $\hat{x} = Ay + b$, which minimizes the mean square error $E(\|x - \hat{x}\|^2)$. It is given by:

$$\hat{x} = m_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y - m_y)$$

where Σ_{xy} and Σ_{yy} are the covariance matrices. It easily follows that $\hat{x}$ is unbiased: $E(\hat{x}) = m_x$, and the corresponding (minimal) error covariance is

$$\mathrm{cov}(x - \hat{x}) = E(x - \hat{x})(x - \hat{x})^T = \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{xy}^T$$

We shall find it convenient to denote this estimator $\hat{x}$ as $E^L(x|y)$. Note that this is *not* the standard conditional expectation.

Recall further the orthogonality principle:

$$E((x - E^L(x|y))L(y)) = 0$$

for any *linear* function L(y) of y.

The following property will be most useful. It follows simply by using $y = (y_1; y_2)$ in the formulas above:

- Suppose $\mathrm{cov}(y_1, y_2) = 0$. Then

  $$E^L(x|y_1, y_2) = E^L(x|y_1) + \left[E^L(x|y_2) - E(x)\right].$$

  Furthermore,

  $$\mathrm{cov}(x - E^L(x|y_1, y_2)) = (\Sigma_{xx} - \Sigma_{xy_1}\Sigma_{y_1y_1}^{-1}\Sigma_{xy_1}^T) - \Sigma_{xy_2}\Sigma_{y_2y_2}^{-1}\Sigma_{xy_2}^T.$$

b. The innovations process

Consider a discrete-time stochastic process $\{z_k\}_{k\geq 0}$. The (wide-sense) innovations process is defined as

$$\tilde{z}_k = z_k - E^L(z_k|Z_{k-1}),$$

where $Z_{k-1} = (z_0; \cdots z_{k-1})$. The innovation RV $\tilde{z}_k$ may be regarded as containing only the new statistical information which is not already in Z_{k-1}.

The following properties follow directly from those of the best linear estimator:

(1) $E(\tilde{z}_k) = 0$, and $E(\tilde{z}_k Z_{k-1}^T) = 0$.

(2) $\tilde{z}_k$ is a linear function of Z_k.

(3) Thus, $\mathrm{cov}(\tilde{z}_k, \tilde{z}_l) = E(\tilde{z}_k \tilde{z}_l^T) = 0$ for $k \neq l$.

This implies that the innovations process is a zero-mean *white noise process.*

Denote $\tilde{Z}_k = (\tilde{z}_0; \cdots; \tilde{z}_k)$. It is easily verified that Z_k and $\tilde{Z}_k$ are *linear* functions of each other. This implies that $E^L(x|Z_k) = E^L(x|\tilde{Z}_k)$ for any RV x.

It follows that (taking $E(x) = 0$ for simplicity):

$$\begin{aligned} E^L(x|Z_k) &= E^L(x|\tilde{Z}_k) \\ &= E^L(x|\tilde{Z}_{k-1}) + E^L(x|\tilde{z}_k) = \sum_{l=0}^{k} E^L(x|\tilde{z}_l) \end{aligned}$$

c. Derivation of the KF equations

We proceed to derive the Kalman filter as the best linear estimator for our linear, non-Gaussian model. We slightly generalize the model that was treated so far by *allowing correlation between the state noise and measurement noise.* Thus, we consider the model

$$\begin{aligned} x_{k+1} &= F_k x_k + G_k w_k\,, \quad k \geq 0 \\ z_k &= H_k x_x + v_k\,, \end{aligned}$$

with $[w_k; v_k]$ a zero-mean white noise sequence with covariance

$$E(\begin{bmatrix} w_k \\ v_k \end{bmatrix} [w_l^T, v_l^T]) = \begin{bmatrix} Q_k & S_k \\ S_k^T & R_k \end{bmatrix} \delta_{kl}\,.$$

x_0 has mean $\overline{x}_0$, covariance P_0, and is uncorrelated with the noise sequence.

We use here the following notation:

$$\begin{array}{ll} Z_k = (z_0; \cdots; z_k) & \\ \hat{x}_{k|k-1} = E^L(x_k|Z_{k-1}) & \hat{x}_{k|k} = E^L(x_k|Z_k) \\ \tilde{x}_{k|k-1} = x_k - \hat{x}_{k|k-1} & \tilde{x}_{k|k} = x_k - \hat{x}_{k|k} \\ P_{k|k-1} = \mathrm{cov}(\tilde{x}_{k|k-1}) & P_{k|k} = \mathrm{cov}(\tilde{x}_{k|k}) \end{array}$$

and defne the innovations process

$$\tilde{z}_k \triangleq z_k - E^L(z_k|Z_{k-1}) = z_k - H_k \hat{x}_{k|k-1}.$$

Note that

$$\tilde{z}_k = H_k \tilde{x}_{k|k-1} + v_k\,.$$

Measurement update: From our previous discussion of linear estimation and innovations,

$$
\begin{aligned}
\hat{x}_{k|k} &= E^L(x_k|Z_k) = E^L(x_k|\tilde{Z}_k) \\
&= E^L(x_k|\tilde{Z}_{k-1}) + E^L(x_k|\tilde{z}_k) - E(x_k)
\end{aligned}
$$

This relation is the basis for the innovations approach. The rest follows essentially by direct computations, and some use of the orthogonality principle. First,

$$E^L(x_k|\tilde{z}_k) - E(x_k) = \mathrm{cov}(x_k, \tilde{z}_k)\mathrm{cov}(\tilde{z}_k)^{-1}\tilde{z}_k.$$

The two covariances are next computed:

$$\mathrm{cov}(x_k, \tilde{z}_k) = \mathrm{cov}(x_k, H_k\tilde{x}_{k|k-1} + v_k) = P_{k|k-1}H_k^T,$$

where $E(x_k\tilde{x}_{k|k-1}^T) = P_{k|k-1}$ follows by orthogonality, and we also used the fact that v_k and x_k are not correlated. Similarly,

$$\mathrm{cov}(\tilde{z}_k) = \mathrm{cov}(H_k\tilde{x}_{k|k-1} + v_k) = H_kP_{k|k-1}H_k^T + R_k \triangleq M_k$$

By substituting in the estimator expression we obtain

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + P_{k|k-1}H_k^TM_k^{-1}\tilde{z}_k$$

Time update: This step is less trivial than before due to the correlation between v_k and w_k. We have

$$
\begin{aligned}
\hat{x}_{k+1|k} &= E^L(x_{k+1}|\tilde{Z}_k) = E^L(F_kx_k + G_kw_k|\tilde{Z}_k) \\
&= F_k\hat{x}_{k|k} + G_kE^L(w_k|\tilde{z}_k)
\end{aligned}
$$

In the last equation we used $E^L(w_k|\tilde{Z}_{k-1}) = 0$ since w_k is uncorrelated with $\tilde{Z}_{k-1}$. Thus

$$
\begin{aligned}
\hat{x}_{k+1|k} &= F_k\hat{x}_{k|k} + G_kE(w_k\tilde{z}_k^T)\mathrm{cov}(\tilde{z}_k)^{-1}\tilde{z}_k \\
&= F_k\hat{x}_{k|k} + G_kS_kM_k^{-1}\tilde{z}_k
\end{aligned}
$$

where $E(w_k\tilde{z}_k^T) = E(w_kv_k^T) = S_k$ follows from $\tilde{z}_k = H_k\tilde{x}_{k|k-1} + v_k$.

Combined update: Combining the measurement and time updates, we obtain the one-step update for $\hat{x}_{k|k-1}$:

$$\hat{x}_{k+1|k} = F_k \hat{x}_{k|k-1} + K_k \tilde{z}_k$$

where

$$\begin{aligned} K_k &\triangleq (F_k P_{k|k-1} H_k + G_k S_k) M_k^{-1} \\ \tilde{z}_k &= z_k - H_k \hat{x}_{k|k-1} \\ M_k &= H_k P_{k|k-1} H_k^T + R_k \,. \end{aligned}$$

Covariance update: The relation between $P_{k|k}$ and $P_{k|k-1}$ is exactly as before. The recursion for $P_{k+1|k}$ is most conveniently obtained in terms of $P_{k|k-1}$ directly. From the previous relations we obtain

$$\tilde{x}_{k+1|k} = (F_k - K_k H_k) \tilde{x}_{k|k-1} + G_k w_k - K_k v_k$$

Since $\tilde{x}_k$ is uncorrelated with w_k and v_k,

$$\begin{aligned} P_{k+1|k} = \; & (F_k - K_k H_k) P_{k|k-1} (F_k - K_k H_k)^T + G_k Q_k G_k^T \\ & + K_k R_k K_k^T - (G_k S_k K_k^T + K_k S_k^T G_k^T) \end{aligned}$$

This completes the filter equations for this case.

Addendum: A Hilbert space interpretation

The definitions and results concerning linear estimators can be nicely interpreted in terms of a Hilbert space formulation.

Consider for simplicity all RVs in this section to have 0 mean.

Recall that a Hilbert space is a (complete) inner-product space. That is, it is a linear vector space V, with a real-valued inner product operation $\langle v_1, v_2 \rangle$ which is bi-linear, symmetric, and non-degenerate ($\langle v, v \rangle = 0$ iff $v = 0$). (Completeness means that every Cauchy sequence has a limit.) The derived norm is defined as $\|v\|^2 = \langle v, v \rangle$. The following facts are standard:

1. A subspace S is a linearly-closed subset of V. Alternatively, it is the linear span of some set of vectors $\{v_\alpha\}$.
2. The *orthogonal projection* $\Pi_S v$ of a vector v unto the subspace S is the closest element to v in S, i.e., the vector $v' \in S$ which minimizes $\|v - v'\|$. Such a vector exists and is unique, and satisfies $(v - \Pi_S v) \perp S$, i.e., $\langle v - \Pi_S v, s \rangle = 0$ for $s \in S$.
3. If $S = \text{span}\{s_1, \ldots, s_k\}$, then $\Pi_S v = \sum_{i=1}^k \alpha_i s_i$, where
$$[\alpha_1, \ldots, \alpha_k] = [\langle v, s_1 \rangle, \ldots, \langle v, s_k \rangle][\langle s_i, s_j \rangle_{i,j=1\ldots k}]^{-1}$$
4. If $S = S_1 \oplus S_2$ (S is the direct sum of two orthogonal subspaces S_1 and S_2), then
$$\Pi_S v = \Pi_{S_1} v + \Pi_{S_2} v\,.$$
If $\{s_1, \ldots, s_k\}$ is an *orthogonal basis* of S, then
$$\Pi_S v = \sum_{i=1}^k \langle v, s_i \rangle \langle s_i, s_i \rangle^{-1} s_i$$
5. Given a set of (independent) vectors $\{v_1, v_2 \ldots\}$, the following *Gram-Schmidt* procedure provides an orthogonal basis:
$$\begin{aligned} \tilde{v}_k &= v_k - \Pi_{span\{v_1 \ldots v_{k-1}\}} v_k \\ &= v_k - \sum_{i=1}^{k-1} \langle v_k, \tilde{v}_i \rangle \langle \tilde{v}_i, \tilde{v}_i \rangle^{-1} v_i \end{aligned}$$

We can fit the previous results on linear estimation to this framework by noting the following correspondence:

- Our Hilbert space is the space of all zero-mean random variables X (on a given probability space) which are square-integrable: $E(X^2) = 0$. The inner product in defined as $\langle X, Y \rangle = E(XY)$.
- The optimal linear estimator $E^L(x_k|Z_k)$, with $Z_k = (z_0, \ldots, z_k)$, is the orthogonal projection of the vector x_k on the subspace spanned by Z_k. (If x_k is vector-valued, we simply consider the projection of each element separately.)
- The innovations process $\{z_k\}$ is an orthogonalized version of $\{z_k\}$.

The Hilbert space formulation provides a nice insight, and can also provide useful technical results, especially in the continuous-time case. However, we shall not go deeper into this topic.

4.4 The Kalman Filter as a Least-Squares Problem

Consider the following deterministic optimization problem.

The cost function (to be minimized):

$$\begin{aligned} J_k &= \frac{1}{2}\,(x_0 - \overline{x}_0)^T P_0^{-1}(x_0 - \overline{x}_0) \\ &\quad + \frac{1}{2}\,\sum_{l=0}^{k}(z_l - H_l x_l)^T R_l^{-1}(z_l - H_l x_l) \\ &\quad + \frac{1}{2}\,\sum_{l=0}^{k-1} w_l^T Q_l^{-1} w_l \end{aligned}$$

under constraints:

$$x_{l+1} = F_l x_l + G_l w_l\,, \quad l = 0, 1, \ldots, k-1$$

The variables are:

$$x_0, \ldots x_k;\; w_0, \ldots w_{k-1}\,.$$

Here $\overline{x}_0$, $\{z_l\}$ are given vectors, and P_0, R_l, Q_l symmetric positive-definite matrices.

Let $(x_o^{(k)}, \ldots, x_k^{(k)})$ denote the optimal solution of this problem. We claim that $x_k^{(k)}$ is given by exactly the same equations as $\hat{x}_{k|k}$ in the corresponding KF problem.

This claim can be established by writing explicitly the least-squares solution for $k-1$ and k, and manipulating the matrix expressions. We will take here a quicker route, using the Gaussian insight.

Theorem The minimizing solution $(x_o^{(k)}, \dots, x_k^{(k)})$ of the above LS problem is the maximizer of the conditional probability (that is, the MAP estimator):

$$p(x_0, \dots, x_k | Z_k)\,, \quad w.r.t.(x_o, \dots, x_k)$$

related to the Gaussian model:

$$\begin{aligned} x_{k+1} &= F_k x_k + G_k w_k\,, \quad x_0 \sim N(\overline{x}_0, P_0) \\ z_k &= H_k x_k + v_k\,, \quad w_k \sim N(0, Q_k),\; v_k \sim N(0, P_k) \end{aligned}$$

with w_k, v_k white and independent of x_0.

Proof: Write down the distribution $p(x_0 \dots x_k, Z_k)$.

Immediate Consequence: Since for Gaussian RV's *MAP*=*MMSE*, then $(x_0, \dots, x_k)^{(k)}$ are equivalent to the expected means: In particular, $x_k^{(k)} = x_k^+$.

Remark: The above theorem (but not the last consequence) holds true even for the non-linear model: $x_{k+1} = F_k(x_k) + G_k w_k$.

4.5 The KF Equations – Summary

a. The basic equations

Initial Conditions:

$$\hat{x}_0^- = \overline{x}_0 \triangleq E(x_0), \quad P_0^- = P_0 \triangleq \text{cov}(x_0)\,.$$

Measurement update:

$$\hat{x}_k^+ = \hat{x}_k^- + K_k(z_k - H_k\hat{x}_k^-)$$
$$P_k^+ = P_k^- - K_kH_kP_k^-$$

where K_k is the *Kalman gain* matrix:

$$K_k = P_k^- H_k^T (H_k P_k^- H_k^T + R_k)^{-1}\,.$$

Time update:

$$\hat{x}_{k+1}^- = F_k\hat{x}_k^+ \;\; [+B_ku_k]$$
$$P_{k+1}^- = F_kP_k^+F_k^T + G_kQ_kG_k^T$$

Note: The measurement update above is valid for any gain matrix K_k, not just the optimal one.

b. One-step iterations

The two-step equations may obviously be combined into a one-step update which computes $\hat{x}_{k+1}^+$ from $\hat{x}_k^+$ (or $\hat{x}_{k+1}^-$ from $\hat{x}_k^-$).
For example,

$$\begin{aligned}\hat{x}_{k+1}^- &= F_k\hat{x}_k^- + F_kK_k(z_k - H_k\hat{x}_k^-)\\ P_{k+1}^- &= F_k(P_k^- - K_kH_kP_k^-)F_k^T + G_kQ_kG_k^T\,.\end{aligned}$$

$L_k \triangleq F_kK_k$ is also known as the Kalman gain.

The iterative equation for P_k^- is called the (discrete-time, time-varying) *Matrix Riccati Equation.*

c. Other important quantities

The measurement prediction, the innovations process, and the innovations covariance are given by

$$\begin{aligned}
\hat{z}_k &\triangleq E(z_k|Z_{k-1}) = H_k\hat{x}_k^- \ (+B_k u_k) \\
\tilde{z}_k &\triangleq z_k - \hat{z}_k = H_k\tilde{x}_k^- \\
M_k &\triangleq \mathrm{cov}(\tilde{z}_k) = H_k P_k^- H_k^T + R_k
\end{aligned}$$

Note that $K_k = P_k^- H_k^T M_k^{-1}$.

d. Alternative Forms for the measurement covariance update (general gain)

Evidently,

$$P_k^+ = P_k^- - K_k H_k P_k^- = (I - K_k H_k)P_k^- \,.$$

A more significant variation is given by the **Joseph form.**
Noting that

$$x_k - \hat{x}_k^+ = (I - K_k H_k)(x_k - \hat{x}_k^-) - K_k v_k$$

it follows immediately that

$$P_k^+ = (I - K_k H_k)P_k^-(I - K_k H_k)^T + K_k R_k K_k^T$$

This form may be more computationally expensive, but has the following advantages:

- Numerically, it is guaranteed to preserve positive-definiteness ($P_k^+ > 0$).
- As mentioned, it holds for any gain K_k (not just the optimal) that may be used in the measurement update $\hat{x}_k^+ = \hat{x}_k^- + K_k\tilde{z}_k$.

e. Alternative Forms for the measurement covariance update (optimal gain)

Substituting the *optimal* gain K_k in the expression for P_k^+, we obtain a symmetric expressions:

$$\begin{aligned}
P_k^+ &= P_k^- - P_k^- H_k^T M_k^{-1} H_k P_k^- \\
&= P_k^- - K_k M_k K_k^T
\end{aligned}$$

It can also be verified that $K_k = P_k^+ H_k^T R_k^{-1}$.

The Information Form: The covariance matrix update for the optimal measurement update also satisfies

$$(P_k^+)^{-1} = (P_k^-)^{-1} + H_k^T R_k^{-1} H_k$$

This equivalence may obtained via the useful *Matrix Inversion Lemma*:

$$(A + BCD)^{-1} = A^{-1} - A^{-1}B(DA^{-1}B + C^{-1})^{-1}DA^{-1}$$

where A, C are square nonsingular matrices (possibly of different size).

P^{-1} is called the *Information Matrix.* It forms the basis for the "information filter", which only computes the inverse covariances.

f. Relation to Deterministic Observers

The one-step recursion for $\hat{x}_k^-$ is similar in form to the algebraic *state observer* from control theory.
Given a (deterministic) system:

$$\begin{aligned} x_{k+1} &= F_k x_k + B_k u_k \\ z_k &= H_k x_k \end{aligned}$$

a state observer is defined by

$$\hat{x}_{k+1} = F_k \hat{x}_k + B_k u_k + L_k(z_k - H_k \hat{x}_k)$$

where L_k are gain matrices to be chosen, with the goal of obtaining $\tilde{x}_k \triangleq (x_k - \hat{x}_k) \to 0$ as $k \to \infty$.

Since

$$\tilde{x}_{k+1} = (F_k - L_k H_k)\tilde{x}_k \,,$$

we need to choose L_k so that the linear system defined by $A_k = (F_k - L_k H_k)$ is asymptotically stable.
This is possible when the original system is *detectable.*

The Kalman gain automatically satisfies this stability requirement (whenever the detectability condition is satisfied).