0.1 Kalman Filtering

A common problem in estimation is deciding how to update an estimate after a new measurement. The new estimate should combine information from both the old estimate and the new measurement, each given appropriate weight. An optimal choice of weights is that which most reduces the uncertainty in the updated estimate.

For a simple example, let $\{x_i : i = 1, 2, \ldots\}$ be a sequence of measurements of quantity $x \in \mathbf{R}$, each with an independent zero-mean, unit-variance normal error:

$$x_i = x + r_i, \qquad r_i \sim N(0,1), \quad i = 1, 2, \ldots$$

The Cramér-Rao lower bound implies that after n measurements, the minimum variance of any estimator $\hat{x}_n$ for x is $1/n$. This is attained by the average,

$$\hat{x}_n \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} x_i,$$

which satisfies $\hat{x}_n \sim N(x, 1/n)$ by the Central Limit Theorem. Rewriting gives:

$$\hat{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{n-1}{n}\hat{x}_{n-1} + \frac{1}{n}x_n = \frac{1}{n}\left[\frac{\hat{x}_{n-1}}{1/(n-1)} + \frac{x_n}{1}\right].$$

This decomposition shows how both the new measurement x_n and the prior estimate $\hat{x}_{n-1}$ are normalized by dividing by their respective variances, then averaged to give the updated estimate. But another decomposition suggests a different interpretation:

$$\hat{x}_n = \hat{x}_{n-1} + \frac{1}{n}\left[x_n - \hat{x}_{n-1}\right] \overset{\text{def}}{=} \hat{x}_{n-1} + K_n\left[x_n - \hat{x}_{n-1}\right],$$

where $x_n - \hat{x}_{n-1}$ is the *innovation*, the difference between the new measurement and the prior estimate, and K_n is the *Kalman gain*, a weighting factor applied to the innovation when updating the estimate. The Kalman gain may be written as a ratio of variances:

$$K_n = \frac{\operatorname{Var}(\hat{x}_{n-1})}{\operatorname{Var}(x_n - \hat{x}_{n-1})} = \frac{\operatorname{Var}(\hat{x}_{n-1})}{\operatorname{Var}(r_n) + \operatorname{Var}(\hat{x}_{n-1})} = \frac{1/(n-1)}{1 + 1/(n-1)} = \frac{1}{n}.$$

In this simple example, it is easy to find the optimal updating formula and the Kalman gain because the optimal estimate and its uncertainty are known.

By using matrix algebra and matrix calculus, we may generalize this idea to vectors $\mathbf{x}$. Moreover, we will allow $\mathbf{x}$ to evolve with some randomness, giving a sequence of unknown *state vectors* $\{\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \ldots\} \subset \mathbf{R}^d$ that are each to be estimated. The observations will be, in the generalization, a sequence of vectors $\{\mathbf{y}_1, \mathbf{y}_2, \ldots\} \subset \mathbf{R}^p$ that depend on the state vectors in a known way but that have random measurement errors. Our goal is to construct an optimal estimator $\hat{\mathbf{x}}_n$ of $\mathbf{x}_n$, using the previous estimator $\hat{\mathbf{x}}_{n-1}$ updated by the measurement $\mathbf{y}_n$. This will be done recursively as in the simple example, using an *optimal Kalman gain* K_n that is a matrix version of the ratio of variances.

0.1.1 Covariance matrices

Let $\mathbf{x} \in \mathbf{R}^d$ be a random vector, and suppose that each component $\mathbf{x}(i)$, for $i = 1, \dots, d$, is a random variable with finite mean $E(\mathbf{x}(i))$ and finite variance $\mathrm{Var}(\mathbf{x}(i)) = E([\mathbf{x}(i) - E(\mathbf{x}(i))]^2)$. Denote the vector of these means by $E(\mathbf{x}) \in \mathbf{R}^d$.

Writing $\mathbf{x}' \overset{\mathrm{def}}{=} \mathbf{x} - E(\mathbf{x})$, the zero-mean random vector obtained from $\mathbf{x}$ by subtracting its mean, we have $\mathrm{Var}(\mathbf{x}(i)) = \mathrm{Var}(\mathbf{x}'(i)) = E(\mathbf{x}'(i)^2)$ for every i. Then quadratic functions of $\mathbf{x}$ may be described using *covariances*,

$$\mathrm{cov}(\mathbf{x}(i), \mathbf{x}(j)) \overset{\mathrm{def}}{=} E(\, \mathbf{x}'(i)\mathbf{x}'(j)\,), \qquad i, j = 1, \dots, d.$$

Thinking of $\mathbf{x}$ as a $d \times 1$ matrix, the (i, j) covariance term will be the expected value of the (i, j) entry in the $d \times d$ matrix $\mathbf{x}'\mathbf{x}'^T$. We may therefore generalize the notion of variance to vector-valued random variables as follows:

$$\begin{aligned}
\mathrm{Var}(\mathbf{x}) \quad &\overset{\mathrm{def}}{=} \quad E(\mathbf{x}'\mathbf{x}'^T) \;=\; E(\begin{pmatrix} \mathbf{x}'(1) \\ \vdots \\ \mathbf{x}'(d) \end{pmatrix} \begin{pmatrix} \mathbf{x}'(1) & \cdots & \mathbf{x}'(d) \end{pmatrix}) \\
&= \quad E \begin{pmatrix} \mathbf{x}'(1)^2 & \cdots & \mathbf{x}'(1)\mathbf{x}'(d) \\ \vdots & \ddots & \vdots \\ \mathbf{x}'(d)\mathbf{x}'(1) & \cdots & \mathbf{x}'(d)^2 \end{pmatrix} \\
&= \quad \begin{pmatrix} E(\mathbf{x}'(1)^2) & \cdots & E(\mathbf{x}'(1)\mathbf{x}'(d)) \\ \vdots & \ddots & \vdots \\ E(\mathbf{x}'(d)\mathbf{x}'(1)) & \cdots & E(\mathbf{x}'(d)^2) \end{pmatrix} \\
&= \quad \begin{pmatrix} \mathrm{Var}(\mathbf{x}(1)) & \cdots & \mathrm{cov}(\mathbf{x}(1), \mathbf{x}(d)) \\ \vdots & \ddots & \vdots \\ \mathrm{cov}(\mathbf{x}(d), \mathbf{x}(1)) & \cdots & \mathrm{Var}(\mathbf{x}(d)) \end{pmatrix}
\end{aligned}$$

Get the diagonal terms using the relation $\mathrm{Var}(y) = \mathrm{cov}(y, y)$. Likewise, $\mathrm{Var}(\mathbf{x})$ collapses to the original definition of $\mathrm{Var}(x)$ for scalar random variables $x \in \mathbf{R}$.

The size of the variance matrix for $\mathbf{x} \in \mathbf{R}^d$ is controlled by its trace:

$$\mathrm{tr}\,\mathrm{Var}(\mathbf{x}) = \sum_{i=1}^{d} \mathrm{Var}(\mathbf{x}(i)) = \sum_{i=1}^{d} E(\mathbf{x}'(i)^2).$$

That is because the diagonal terms dominate the off-diagonal terms:

Theorem 0.1 *If the components of random vector* $\mathbf{x} : \Omega \to \mathbf{R}^d$ *have finite means and variances, then*

$$\|\mathrm{Var}(\mathbf{x})\|_{HS} \le \mathrm{tr}\,\mathrm{Var}(\mathbf{x}),$$

where $\|\cdot\|_{HS}$ *is the* Hilbert-Schmidt *norm.*

Proof: Since $\mathbf{x}(i)$ and $\mathbf{x}(j)$ have finite variances, they are square-integrable over Ω and thus belong to the inner product space $L^2(\Omega)$ with inner product $\langle y, z\rangle \overset{\text{def}}{=} E(yz)$ and derived norm $\|y\| \overset{\text{def}}{=} E(y^2)$. Then the Cauchy–Schwarz inequality implies that

$$|E(\mathbf{x}'(i)\,\mathbf{x}'(j))| = |\,\langle \mathbf{x}'(i), \mathbf{x}'(j)\rangle\,| \leq \|\mathbf{x}'(i)\|\,\|\mathbf{x}'(j)\| = \sqrt{E(\mathbf{x}'(i)^2)}\sqrt{E(\mathbf{x}'(j)^2)}.$$

Summing $E(\mathbf{x}'(i)\,\mathbf{x}'(j))^2$ over $i, j = 1, \ldots, d$ and taking square roots gives $\|\mathrm{Var}(\mathbf{x})\|_{\mathrm{HS}}^2$ on the left and $[\mathrm{tr}\,\mathrm{Var}(\mathbf{x})]^2$ on the right. □

Since $\|\cdot\|_{\mathrm{HS}}$ is comparable to every other norm on the finite-dimensional space $\mathbf{R}^{d\times d}$, every norm on $\mathrm{Var}(\mathbf{x})$ is dominated by a multiple of $\mathrm{tr}\,\mathrm{Var}(\mathbf{x})$.

We may also generalize the notion of covariance to pairs of random vectors $\mathbf{x} \in \mathbf{R}^d$ and $\mathbf{y} \in \mathbf{R}^p$:

$$\begin{aligned}
\mathrm{cov}(\mathbf{x}, \mathbf{y}) \quad &\overset{\text{def}}{=} \quad E(\mathbf{x}'\mathbf{y}'^T) \;=\; E(\begin{pmatrix} \mathbf{x}'(1) \\ \vdots \\ \mathbf{x}'(d) \end{pmatrix} \begin{pmatrix} \mathbf{y}'(1) & \cdots & \mathbf{y}'(p) \end{pmatrix}) \\
&= \quad E\begin{pmatrix} \mathbf{x}'(1)\mathbf{y}'(1) & \cdots & \mathbf{x}'(1)\mathbf{y}'(p) \\ \vdots & \ddots & \vdots \\ \mathbf{x}'(d)\mathbf{y}'(1) & \cdots & \mathbf{x}'(d)\mathbf{y}'(p) \end{pmatrix} \\
&= \quad \begin{pmatrix} E(\mathbf{x}'(1)\mathbf{y}'(1)) & \cdots & E(\mathbf{x}'(1)\mathbf{y}'(p)) \\ \vdots & \ddots & \vdots \\ E(\mathbf{x}'(d)\mathbf{y}'(1)) & \cdots & E(\mathbf{x}'(d)\mathbf{y}'(p)) \end{pmatrix} \\
&= \quad \begin{pmatrix} \mathrm{cov}(\mathbf{x}(1), \mathbf{y}(1)) & \cdots & \mathrm{cov}(\mathbf{x}(1), \mathbf{y}(p)) \\ \vdots & \ddots & \vdots \\ \mathrm{cov}(\mathbf{x}(d), \mathbf{y}(1)) & \cdots & \mathrm{cov}(\mathbf{x}(d), \mathbf{y}(p)) \end{pmatrix},
\end{aligned}$$

which is a $d \times p$ matrix. Argument order matters in the matrix case: $\mathrm{cov}(\mathbf{y}, \mathbf{x}) = \mathrm{cov}(\mathbf{x}, \mathbf{y})^T$ is a $p \times d$ matrix. Note that $\mathrm{Var}(\mathbf{x}) = \mathrm{cov}(\mathbf{x}, \mathbf{x})$, as in the scalar case.

A constant vector $\mathbf{b} \in \mathbf{R}^d$ may be regarded as a random vector with mean $E(\mathbf{b}) = \mathbf{b}$ and variance $\mathrm{Var}(\mathbf{b}) = \mathrm{Var}(\mathbf{0}) = 0$, since $\mathbf{b}' = 0$.

We say that two random variables $x, y \in \mathbf{R}$ are *uncorrelated* if and only if $\mathrm{cov}(x, y) = E(x'y') = 0$. That relationship is preserved by affine transformations:

Lemma 0.2 *If $x, y \in \mathbf{R}$ are uncorrelated random variables and $a, b \in \mathbf{R}$ are constants, then $ax + b$ and y are also uncorrelated random variables.*

Proof: Note that

$$[ax + b]' = ax + b - E(ax + b) = ax + b - [aE(x) + b] = a[x - E(x)] = ax'.$$

Thus $\mathrm{cov}(ax + b, y) = E([ax + b]'y') = E(ax'y') = aE(x'y') = a\,\mathrm{cov}(x, y) = 0.$ □

Lemma 0.3 *If x and y are uncorrelated random variables, then $E(xy) = E(x)E(y)$.*

Proof: Since $x = x' + E(x)$ and $y = y' + E(y)$, compute

$$\begin{aligned} E(xy) &= E([x' + E(x)][y' + E(y)]) \;=\; E(x'y' + y'E(x) + x'E(y) + E(x)E(y)) \\ &= E(x'y') + E(y')E(x) + E(x')E(y) + E(x)E(y) \;=\; E(x)E(y), \end{aligned}$$

since $E(x') = E(y') = E(x'y') = 0$. □

Say that random vectors $\mathbf{x} \in \mathbf{R}^d$ and $\mathbf{y} \in \mathbf{R}^p$ are *uncorrelated* if and only if all pairs of their coordinates $\mathbf{x}(i), \mathbf{y}(j)$, $i = 1, \ldots, d$, $j = 1, \ldots, p$, are uncorrelated random variables. Since $\operatorname{cov}(\mathbf{x}(i), \mathbf{y}(j)) = 0$ for all i, j, this is equivalent to the vanishing of their $d \times p$ and $p \times d$ covariance matrices:

$$\operatorname{cov}(\mathbf{x}, \mathbf{y}) = E(\mathbf{x}'\mathbf{y}'^T) = 0; \qquad \operatorname{cov}(\mathbf{y}, \mathbf{x}) = E(\mathbf{y}'\mathbf{x}'^T) = 0. \tag{1}$$

Lemma 0.4 *If $\mathbf{x}, \mathbf{y} \in \mathbf{R}^d$ are uncorrelated random vectors, then $\operatorname{Var}(\mathbf{x} \pm \mathbf{y}) = \operatorname{Var}(\mathbf{x}) + \operatorname{Var}(\mathbf{y})$.*

Proof: First note that

$$(\mathbf{x} \pm \mathbf{y})' = \mathbf{x} \pm \mathbf{y} - E(\mathbf{x} \pm \mathbf{y}) = [\mathbf{x} - E(\mathbf{x})] \pm [\mathbf{y} - E(\mathbf{y})] = \mathbf{x}' \pm \mathbf{y}'.$$

Thus

$$\begin{aligned} \operatorname{Var}(\mathbf{x} \pm \mathbf{y}) &= E((\mathbf{x}' \pm \mathbf{y}')(\mathbf{x}' \pm \mathbf{y}')^T) \\ &= E(\mathbf{x}'\mathbf{x}'^T + \mathbf{y}'\mathbf{y}'^T \pm \mathbf{x}'\mathbf{y}'^T \pm \mathbf{y}'\mathbf{x}'^T) \\ &= E(\mathbf{x}'\mathbf{x}'^T) + E(\mathbf{y}'\mathbf{y}'^T) \pm E(\mathbf{x}'\mathbf{y}'^T) \pm E(\mathbf{y}'\mathbf{x}'^T) \\ &= \operatorname{Var}(\mathbf{x}) + \operatorname{Var}(\mathbf{y}) \pm \operatorname{cov}(\mathbf{x}, \mathbf{y}) \pm \operatorname{cov}(\mathbf{y}, \mathbf{x}) \;=\; \operatorname{Var}(\mathbf{x}) + \operatorname{Var}(\mathbf{y}). \end{aligned}$$

The last two covariances are zero by Equation 1. □

Lemma 0.5 *If $\mathbf{x} \in \mathbf{R}^d$ is a random vector, $A \in \mathbf{R}^{p \times d}$ is a fixed $p \times d$ matrix, and $\mathbf{b} \in \mathbf{R}^p$ is a fixed vector, then*

$$\operatorname{Var}(A\mathbf{x} + \mathbf{b}) = A \operatorname{Var}(\mathbf{x})\, A^T.$$

Proof: Since $E(A\mathbf{x} + \mathbf{b}) = AE(\mathbf{x}) + \mathbf{b}$ we have $(A\mathbf{x} + \mathbf{b})' = A\mathbf{x}'$. Thus $\operatorname{Var}(A\mathbf{x} + \mathbf{b}) = E((A\mathbf{x}')(A\mathbf{x}')^T) = E(A\,\mathbf{x}'\mathbf{x}'^T A^T) = A\,E(\mathbf{x}'\mathbf{x}'^T)\,A^T = A \operatorname{Var}(\mathbf{x})\,A^T$. □

Lemma 0.6 *Suppose $A \in \mathbf{R}^{p \times d}$ is a fixed $p \times d$ matrix, and $\mathbf{b} \in \mathbf{R}^p$ is a fixed vector. If $\mathbf{x} \in \mathbf{R}^d$ and $\mathbf{y} \in \mathbf{R}^p$ are uncorrelated random vectors, then $A\mathbf{x} + \mathbf{b}$ and $\mathbf{y}$ are also uncorrelated random vectors.*

Proof: Compute $\operatorname{cov}(A\mathbf{x} + \mathbf{b}, \mathbf{y}) = E(A\mathbf{x}'\mathbf{y}'^T) = A\,E(\mathbf{x}'\mathbf{y}'^T) = A \operatorname{cov}(\mathbf{x}, \mathbf{y}) = 0$. □

Combining Lemmas 0.4 and 0.5 gives us the desired computing tool:

Corollary 0.7 *, If $\mathbf{x} \in \mathbf{R}^d$ and $\mathbf{y} \in \mathbf{R}^p$ are uncorrelated random vectors, then*

$$\operatorname{Var}(\mathbf{y} + A\mathbf{x} + \mathbf{b}) = \operatorname{Var}(\mathbf{y}) + A \operatorname{Var}(\mathbf{x})\, A^T$$

for any fixed matrix $A \in \mathbf{R}^{p \times d}$ and any fixed vector $\mathbf{b} \in \mathbf{R}^p$. □

0.1.2 Matrix calculus

Suppose $X \in \mathbf{R}^{m\times n}$ is a matrix. Then X may be regarded as a list of mn variables $x_{11}, \dots, x_{ij}, \dots, x_{mn}$.

If $X \mapsto F(X) \in \mathbf{R}$ defines a scalar-valued function of the matrix variable X, then its gradient with respect to X may be written as

$$\nabla_X F \quad = \quad \begin{pmatrix} \frac{\partial F}{\partial x_{11}} & \cdots & \frac{\partial F}{\partial x_{1n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial F}{\partial x_{m1}} & \cdots & \frac{\partial F}{\partial x_{mn}} \end{pmatrix},$$

where, for future notational convenience, the partial derivatives are arranged in a matrix of the same dimensions as X.

For the scalar-valued function $F(X) = \mathrm{tr}\,(AXB)$ with fixed matrices $A = (a_{ij})$ and $B = (b_{ij})$ of appropriate dimensions, we have

$$F(X) = \sum_{r,s,t} a_{rs}\, x_{st}\, b_{tr} \qquad \Rightarrow \frac{\partial F}{\partial x_{ij}} = \sum_r a_{ri}\, b_{jr}.$$

This is the i, j entry in $A^T B^T$, giving

$$\nabla_X \mathrm{tr}\,(AXB) = A^T B^T. \tag{2}$$

Since $\mathrm{tr}\,(M) = \mathrm{tr}\,(M^T)$ for all square matrices M, we also have

$$\nabla_X \mathrm{tr}\,(AX^T B) = BA. \tag{3}$$

For the scalar-valued function $F(X) = \mathrm{tr}\,(XAX^T)$ with fixed square matrix $A = (a_{ij})$ of appropriate dimensions, we have

$$F(X) = \sum_{r,s,t} x_{rs}\, a_{st}\, x_{rt} \qquad \Rightarrow \frac{\partial F}{\partial x_{ij}} = \sum_t a_{jt}\, x_{it} + \sum_s x_{is} a_{sj}.$$

This is the i, j entry in $XA^T + XA$, giving

$$\nabla_X \mathrm{tr}\,(XAX^T) = XA^T + XA. \tag{4}$$

Likewise,

$$\nabla_X \mathrm{tr}\,(X^T AX) = A^T X + AX. \tag{5}$$

In particular, if $A = A^T$ is symmetric, we get

$$\nabla_X \mathrm{tr}\,(XAX^T) = 2XA; \qquad \nabla_X \mathrm{tr}\,(X^T AX) = 2AX.$$

Specializing still further to $A = Id$, for every X we have

$$\nabla_X \mathrm{tr}\,(XX^T) = \nabla_X \mathrm{tr}\,(X^T X) = 2X. \tag{6}$$

0.1.3 Estimating linear systems

The Kalman gain estimation method may be generalized to apply to random vectors that change by an affine process. Namely, suppose that $\mathbf{x}_0 \in \mathbf{R}^d$ is a random vector with known variance matrix $P_0 = \mathrm{Var}(\mathbf{x}_0)$, and define a sequence of random vectors $\{\mathbf{x}_1, \mathbf{x}_2, \ldots\} \subset \mathbf{R}^d$ by the recurrence

$$\mathbf{x}_n = A_n \mathbf{x}_{n-1} + \mathbf{b}_n + \mathbf{q}_n, \qquad n = 1, 2, \ldots,$$

where $\{A_i : i = 1, 2, \ldots\} \subset \mathbf{R}^{d\times d}$ and $\{\mathbf{b}_i : i = 1, 2, \ldots\} \subset \mathbf{R}^d$ are fixed sequences and $\{\mathbf{q}_i : i = 1, 2, \ldots\} \subset \mathbf{R}^d$ is a sequence of mutually uncorrelated random variables with mean zero and variances $\mathrm{Var}(\mathbf{q}_i) = Q_i$, that are also uncorrelated with $\mathbf{x}_0$. Then by Lemma 0.7,

$$P_n \overset{\text{def}}{=} \mathrm{Var}(\mathbf{x}_n) = A_n P_{n-1} A_n^T + Q_n$$

gives the variance of $\mathbf{x}_n$. Such a variance reflects great ignorance about $\mathbf{x}_n$ since no measurements are made. We expect P_n to be large in some sense, and to grow as $n \to \infty$.

Now suppose that p measurements are made of linear combinations of the d coordinates of $\mathbf{x}_n$:

$$\mathbf{y}_n = H_n \mathbf{x}_n + \mathbf{r}_n,$$

where $H_n \in \mathbf{R}^{p\times d}$ is a fixed matrix whose rows give the coefficients of the linear combinations, and the *measurement error* $\mathbf{r}_n \in \mathbf{R}^p$ has variance R_n and is one of a sequence $\{\mathbf{r}_i : i = 1, 2, \ldots\} \subset \mathbf{R}^p$ of mutually uncorrelated mean-zero random variables that are also uncorrelated with $\{\mathbf{x}_i : i = 0, 1, \ldots\}$ and $\{\mathbf{q}_i : i = 1, 2, \ldots\}$. Our goal is to construct an updated estimator $\hat{\mathbf{x}}_n$ for the latest state $\mathbf{x}_n$, using the measurement $\mathbf{y}_n$ and the previous estimator $\hat{\mathbf{x}}_{n-1}$, such that $\hat{P}_n \overset{\text{def}}{=} \mathrm{Var}(\mathbf{x}_n - \hat{\mathbf{x}}_n)$, the latest estimator's variance from the latest state, is minimal in some sense, and in particular is smaller than P_n.

The estimator $\hat{\mathbf{x}}_n$ will be constructed recursively in a manner analogous to updating a running average. Namely, we will define a Kalman gain K_n such that

$$\hat{\mathbf{x}}_n = A_n \hat{\mathbf{x}}_{n-1} + \mathbf{b}_n + K_n \left[\mathbf{y}_n - \hat{\mathbf{y}}_n\right],$$

where $\hat{\mathbf{y}}_n \overset{\text{def}}{=} H_n \left(A_n \hat{\mathbf{x}}_{n-1} + \mathbf{b}_n\right)$ is the expected measurement, given the previous estimate. The difference $\mathbf{y}_n - \hat{\mathbf{y}}_n$ is the innovation. Hence the new estimator is a prediction, namely $A_n \hat{\mathbf{x}}_{n-1} + \mathbf{b}_n$, from the prior estimate $\hat{\mathbf{x}}_{n-1}$, updated by a weighted correction based on the deviation from the predicted.measurement. The variance $\hat{P}_n$ from the true state will also satisfy a recurrence in terms of $\hat{P}_{n-1}$ that can be adjusted, by an optimal choice of K_n, to minimize $\hat{P}_n$ at each n.

The Kalman estimate $\hat{\mathbf{x}}$ is computed in several steps:

- Predicted (*a priori*) state vector from the previous state estimate:

$$\hat{\mathbf{x}}_{n|n-1} \overset{\text{def}}{=} A_n \hat{\mathbf{x}}_{n-1} + \mathbf{b}_n$$

- Predicted (*a priori*) measurement:

$$\hat{\mathbf{y}}_{n|n-1} \overset{\text{def}}{=} H_n \hat{\mathbf{x}}_{n|n-1}$$

- Innovation:

$$\hat{\mathbf{y}}_n \overset{\text{def}}{=} \mathbf{y}_n - \hat{\mathbf{y}}_{n|n-1}$$

- Updated (*a posteriori*) state vector estimate:

$$\hat{\mathbf{x}}_n \overset{\text{def}}{=} \hat{\mathbf{x}}_{n|n-1} + K_n \hat{\mathbf{y}}_n$$

We will choose K_n to minimize $\operatorname{tr} \hat{P}_n$, defined below.

- Predicted (*a priori*) state estimate variance from the previous state estimate variance:

$$\begin{aligned} \hat{P}_{n|n-1} &\overset{\text{def}}{=} \operatorname{Var}(\mathbf{x}_n - \hat{\mathbf{x}}_{n|n-1}) \;=\; \operatorname{Var}(A_n(\mathbf{x}_{n-1} - \hat{\mathbf{x}}_{n-1}) + \mathbf{q}_n) \\ &= A_n \operatorname{Var}(\mathbf{x}_{n-1} - \hat{\mathbf{x}}_{n-1})\, A_n^T + \operatorname{Var}(\mathbf{q}_n) \;=\; A_n \hat{P}_{n-1} A_n^T + Q_n. \end{aligned}$$

- Innovation variance:

$$\begin{aligned} S_n &\overset{\text{def}}{=} \operatorname{Var}(\hat{\mathbf{y}}_n) \;=\; \operatorname{Var}(H_n(\mathbf{x}_n - \hat{\mathbf{x}}_{n|n-1}) + \mathbf{r}_n) \\ &= H_n \operatorname{Var}(\mathbf{x}_n - \hat{\mathbf{x}}_{n|n-1})\, H_n^T + \operatorname{Var}(\mathbf{r}_n) \;=\; H_n \hat{P}_{n|n-1} H_n^T + R_n. \end{aligned}$$

Update the (*a posteriori*) state estimate variance:

$$\begin{aligned} \hat{P}_n &= \operatorname{Var}(\mathbf{x}_n - \hat{\mathbf{x}}_n) \;=\; \operatorname{Var}(\mathbf{x}_n - [\hat{\mathbf{x}}_{n|n-1} + K_n \hat{\mathbf{y}}_n]) \\ &= \operatorname{Var}(\mathbf{x}_n - [\hat{\mathbf{x}}_{n|n-1} + K_n(\mathbf{y}_n - H_n \hat{\mathbf{x}}_{n|n-1})]) \\ &= \operatorname{Var}(\mathbf{x}_n - [\hat{\mathbf{x}}_{n|n-1} + K_n([H_n \mathbf{x}_n + \mathbf{r}_n] - H_n \hat{\mathbf{x}}_{n|n-1})]) \\ &= \operatorname{Var}([I - K_n H_n](\mathbf{x}_n - \hat{\mathbf{x}}_{n|n-1}) - K_n \mathbf{r}_n) \\ &= (I - K_n H_n)\, \hat{P}_{n|n-1}\, (I - K_n H_n)^T + K_n\, R_n\, K_n^T, \end{aligned}$$

since $\mathbf{r}_n$ and $\mathbf{x}_n - \hat{\mathbf{x}}_{n|n-1}$ are uncorrelated. To derive the full recurrence formula $\hat{P}_{n-1} \to \hat{P}_n$ takes one more substitution:

$$\hat{P}_n = (I - K_n H_n)[A_n \hat{P}_{n-1} A_n^T + Q_n](I - K_n H_n)^T + K_n\, R_n\, K_n^T.$$

Now minimize $\hat{P}_n$ by setting $\nabla_K(\operatorname{tr} \hat{P}_n) = 0$. It is enough to use the recurrence formula $\hat{P}_{n|n-1} \to \hat{P}_n$ since $\hat{P}_{n|n-1}$ has no K_n dependence. Begin by expanding:

$$\begin{aligned} \hat{P}_n &= \hat{P}_{n|n-1} - K_n H_n \hat{P}_{n|n-1} - \hat{P}_{n|n-1} H_n^T K_n^T + K_n[H_n \hat{P}_{n|n-1} H_n^T + R_n] K_n^T \\ &= \hat{P}_{n|n-1} - K_n H_n \hat{P}_{n|n-1} - \hat{P}_{n|n-1} H_n^T K_n^T + K_n S_n K_n^T. \end{aligned}$$

All of the K dependence is explicit in this formula, so we may compute the gradient with respect to K by matrix calculus. To find the optimal Kalman gain K_n, set the K-gradient of the trace of $\hat{P}_n$ to zero:

$$\begin{aligned} 0 = \nabla_K \mathrm{tr}\, \hat{P}_n &= \nabla_K \mathrm{tr}\, (\hat{P}_{n|n-1}) - \nabla_K \mathrm{tr}\, (K_n H_n \hat{P}_{n|n-1}) \\ &\qquad -\nabla_K \mathrm{tr}\, (\hat{P}_{n|n-1} H_n^T K_n^T) + \nabla_K \mathrm{tr}\, (K_n S_n K_n^T) \\ &= 0 - (H_n \hat{P}_{n|n-1})^T - \hat{P}_{n|n-1} H_n^T + K_n S_n^T + K_n S_n \\ &= -2(H_n \hat{P}_{n|n-1})^T + 2K_n S_n, \end{aligned}$$

since $S_n^T = S_n$ and $\hat{P}_{n|n-1}^T = \hat{P}_{n|n-1}$. Conclude that the optimal Kalman gain is

$$K_n = \hat{P}_{n|n-1} H_n^T S_n^{-1}. \tag{7}$$

For this optimal gain, the *a posteriori* state estimate variance recurrence simplifies:

$$\begin{aligned} \hat{P}_n &= \hat{P}_{n|n-1} - K_n H_n \hat{P}_{n|n-1} - \hat{P}_{n|n-1} H_n^T K_n^T + K_n S_n K_n^T \\ &= (I - K_n H_n)\hat{P}_{n|n-1} - \hat{P}_{n|n-1} H_n^T K_n^T + (\hat{P}_{n|n-1} H_n^T S_n^{-1}) S_n K_n^T \\ &= (I - K_n H_n)\hat{P}_{n|n-1}. \end{aligned}$$

Summarizing, we start with an unknown prior state vector $\mathbf{x}_{n-1}$ that we estimate by $\hat{\mathbf{x}}_{n-1}$ with variance $\hat{P}_{n-1}$. This evolves to an unknown current state vector $\mathbf{x}_n = A_n \mathbf{x}_{n-1} + \mathbf{b}_n + \mathbf{q}_n$. We make one measurement of vector $\mathbf{y}_n = H_n \mathbf{x}_n + \mathbf{r}_n$ and then update the estimate and its variance as follows:

- Predicted current state estimate variance: $\hat{P}_{n|n-1} = A_n \hat{P}_{n-1} A_n^T + Q_n$;
- Innovation variance $S_n = H_n \hat{P}_{n|n-1} H_n^T + R_n$;
- Optimal Kalman gain $K_n = \hat{P}_{n|n-1} H_n^T S_n^{-1}$;
- Predicted state vector estimate $\hat{\mathbf{x}}_{n|n-1} = A_n \hat{\mathbf{x}}_{n-1} + \mathbf{b}_n$;
- Predicted measurement $\hat{\mathbf{y}}_{n|n-1} = H_n \hat{\mathbf{x}}_{n|n-1}$;
- Updated state vector estimate $\hat{\mathbf{x}}_n = \hat{\mathbf{x}}_{n|n-1} + K_n(\mathbf{y}_n - \hat{\mathbf{y}}_{n|n-1})$.
- Updated state estimate variance $\hat{P}_n = (I - K_n H_n)\hat{P}_{n|n-1}$.

By assumption, the initial state vector $\mathbf{x}_0$ and all the noise vectors $\mathbf{q}_k$ and $\mathbf{r}_k$ are mutually uncorrelated random vectors with known means and variances. The state estimate and its variance may be initialized by

$$\hat{\mathbf{x}}_0 \overset{\text{def}}{=} E(\mathbf{x}_0); \qquad \hat{P}_0 \overset{\text{def}}{=} \mathrm{Var}(\mathbf{x}_0),$$

while the noise vectors all have mean zero and variances Q_k and R_k, respectively.

0.2 Exercises

1. For fixed $h > 0$ and $n = 0, 1, 2, \ldots$, let $t_n = t_0 + nh$ be a grid of time steps. Write $\mathbf{x}_n = (x_n, x'_n, x''_n)$ for the vector describing the position, velocity, and acceleration, respectively, of a particle at time $t_n = t_0 + nh$. Assuming that the particle's trajectory is smooth, use Taylor's theorem to write a recurrence for $\mathbf{x}_n$ in terms of $\mathbf{x}_{n-1}$, treating any unknown terms as errors.

2. For fixed $h > 0$ and $n = 0, 1, 2, \ldots$, let $t_n = t_0 + nh$ be a grid of time steps. Write $\mathbf{x}_n = (x_n, x'_n, x''_n)$ for the vector describing the position, velocity, and acceleration, respectively, of a particle at time $t_n = t_0 + nh$. Assuming that the position and velocity of the particle can be measured with errors $\mathbf{r}_n = (r_n, r'_n)$ at each time step, write the linear equation for measurement $\mathbf{y}_n$ that should be used in Kalman filtering for $\mathbf{x}_n$.

3. Implement the Kalman filtering algorithm for the system described in the problems above. Assume that $\mathbf{x}_0 = (0, 0, 0)$ with $P_0 = \mathrm{Var}(\mathbf{x}_0) = Id$, and for every $n = 1, 2, 3, \ldots$, put

$$Q_n = \mathrm{Var}(\mathbf{q}_n) = \begin{pmatrix} h^6/36 & 0 & 0 \\ 0 & h^4/4 & 0 \\ 0 & 0 & h^2 \end{pmatrix}; \qquad R_n = \mathrm{Var}(\mathbf{r}_n) = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \tau^2 \end{pmatrix},$$

with values of h, σ, τ to be specified at run time.

0.3 Solutions to the Exercises

1. For fixed $h > 0$ and $n = 0, 1, 2, \ldots$, let $t_n = t_0 + nh$ be a grid of time steps. Write $\mathbf{x}_n = (x_n, x'_n, x''_n)$ for the vector describing the position, velocity, and acceleration, respectively, of a particle at time $t_n = t_0 + nh$. Assuming that the particle's trajectory is smooth, use Taylor's theorem to write a recurrence for $\mathbf{x}_n$ in terms of $\mathbf{x}_{n-1}$, treating any unknown terms as errors.

Solution: By Taylor's theorem,

$$\begin{aligned} x(t+h) &= x(t) + h\,x'(t) + \frac{h^2}{2}\,x''(t) + q(t,h), \\ x'(t+h) &= x'(t) + h\,x''(t) + q'(t,h), \\ x''(t+h) &= x''(t) + q''(t,h), \end{aligned}$$

with errors $q(t,h) = x'''(c)h^3/6$, $q'(t,h) = x'''(c')h^2/2$, and $q''(t,h) = x'''(c'')h$ that depend on the values of x''' at unknown points $c, c', c'' \in [t, t+h]$. Putting $t = t_{n-1}$ so that $t + h = t_n$, we may write

$$\mathbf{x}_n = \begin{pmatrix} 1 & h & h^2/2 \\ 0 & 1 & h \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_{n-1} \\ x'_{n-1} \\ x''_{n-1} \end{pmatrix} + \begin{pmatrix} q_n \\ q'_n \\ q''_n \end{pmatrix} \overset{\text{def}}{=} A\mathbf{x}_{n-1} + \mathbf{q}_n,$$

where $\mathbf{q}_n = \mathbf{q}(t_n, h) = (q_n, q'_n, q''_n)$ is the vector of process errors. □

2. For fixed $h > 0$ and $n = 0, 1, 2, \ldots$, let $t_n = t_0 + nh$ be a grid of time steps. Write $\mathbf{x}_n = (x_n, x'_n, x''_n)$ for the vector describing the position, velocity, and acceleration, respectively, of a particle at time $t_n = t_0 + nh$. Assuming that the position and velocity of the particle can be measured with errors $\mathbf{r}_n = (r_n, r'_n)$ at each time step, write the linear equation for measurement $\mathbf{y}_n$ that should be used in Kalman filtering for $\mathbf{x}_n$.

   Solution: Putting $\mathbf{y}_n = (y_n, y'_n)$ for the vector of position and velocity measurements, we may write

$$\mathbf{y}_n = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} x_n \\ x'_n \\ x''_n \end{pmatrix} + \begin{pmatrix} r_n \\ r'_n \end{pmatrix} \overset{\text{def}}{=} H\mathbf{x}_n + \mathbf{r}_n,$$

   where $\mathbf{r}_n = \mathbf{r}(t_n) = (r_n, r'_n)$ is the vector of measurement errors. □

3. Implement the Kalman filtering algorithm for the system described in the problems above. Assume that $\mathbf{x}_0 = (0, 0, 0)$ with $P_0 = \mathrm{Var}(\mathbf{x}_0) = Id$, and for every $n = 1, 2, 3, \ldots$, put

$$Q_n = \mathrm{Var}(\mathbf{q}_n) = \begin{pmatrix} h^6/36 & 0 & 0 \\ 0 & h^4/4 & 0 \\ 0 & 0 & h^2 \end{pmatrix}; \qquad R_n = \mathrm{Var}(\mathbf{r}_n) = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \tau^2 \end{pmatrix},$$

   with values of h, σ, τ to be specified at run time.

   Solution: Modify the codes `kalmanf.m` and `taylor3.m` on the class web site. □