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,

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

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:

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

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:

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

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 $i = 1, \ldots, d$, 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 $\mathbf{x}' \stackrel{\mathrm{def}}{=} \mathbf{x} - E(\mathbf{x})$, 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*,

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

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:

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

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 Cauchy–Schwarz inequality implies that the diagonal terms dominate the off-diagonal terms:

$$E(\mathbf{x}'(i)\,\mathbf{x}'(j)) \le \sqrt{E(\mathbf{x}'(i)^2)}\sqrt{E(\mathbf{x}'(j)^2)}.$$

Summing $E(\mathbf{x}'(i)\,\mathbf{x}'(j))^2$ over $i, j = 1, \ldots, d$ and taking square roots gives

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

where $\|\cdot\|_{\mathrm{HS}}$ is the *Hilbert-Schmidt* norm. 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 $\operatorname{tr}\mathrm{Var}(\mathbf{x})$.

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

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

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.

Recall that two random variables $x, y\in\mathbf{R}$ are called *independent* if and only if $\Pr(x\in I \text{ and } y\in J) = \Pr(x\in I)\Pr(y\in J)$ for any pair of intervals $I, J\subset\mathbf{R}$. That relationship is preserved by affine transformations:

Lemma 0.1 *If $x, y\in\mathbf{R}$ are independent random variables and $a, b\in\mathbf{R}$ are constants, then $ax+b$ and y are independent random variables.*

Proof: Let $S, T\subset\mathbf{R}$ be any intervals.

First consider the case $a=0$. Then the random variable $ax+b=b$ has the probability function

$$\Pr(ax+b\in S) = \Pr(b\in S) = \begin{cases}1, & \text{if } b\in S;\\ 0, & \text{if } b\notin S,\end{cases} \stackrel{\text{def}}{=} \mathbf{1}_S(b),$$

which depends neither on x nor on y. Hence we have the trivial factorization

$$\Pr(ax+b\in S \text{ and } y\in T) = \begin{cases}\Pr(y\in T), & \text{if } b\in S;\\ 0, & \text{if } b\notin S,\end{cases} = \Pr(ax+b\in S)\Pr(y\in T),$$

indicating the independence of $ax+b$ and y.

For $a\neq 0$ we may define $\frac{1}{a}[S-b] \stackrel{\text{def}}{=} \{\frac{1}{a}[w-b] : w\in S\}$. Note that since S is an interval, then so is $\frac{1}{a}[S-b]$ and the two intervals satisfy

$$ax+b\in S \quad\Longleftrightarrow\quad x\in\frac{1}{a}[S-b].$$

But then, by the independence of x and y,

$$\begin{aligned}\Pr(ax+b \in S \text{ and } y \in T) &= \Pr(x \in \frac{1}{a}[S-b] \text{ and } y \in T) \\ &= \Pr(x \in \frac{1}{a}[S-b]) \Pr(y \in T) \\ &= \Pr(ax+b \in S) \Pr(y \in T).\end{aligned}$$

Since S, T were arbitrary, this implies that $ax+b$ and y are independent. □

Lemma 0.2 *If x and y are independent random variables with finite means $E(x), E(y)$, then $E(xy) = E(x)E(y)$.*

Proof: Let Pr be the joint probability function
Write X, Y for the cdfs of x, y, respectively. Then

$$E(x) = \int_{\mathbf{R}} t\, X(dt); \qquad E(y) = \int_{\mathbf{R}} s\, Y(ds).$$

□

Corollary 0.3 *If x and y are independent random variables, then* $\operatorname{cov}(x, y) = 0$.

Proof: The mean-subtracted variables $x' = x - E(x)$ and $y' = y - E(y)$ are independent by two applications of Lemma 0.1, part (ii). Then we may compute $\operatorname{cov}(x, y) = E(x'y') = E(x')E(y') = 0^2 = 0$ by that same lemma's part (i). □

The converse of Corollary 0.3 is false: $\operatorname{cov}(x, y) = 0$ does not imply that x and y are independent. For a counterexample, suppose that x and y are random variables taking values in $[-1, 1] \subset \mathbf{R}$ and related by $y = x^2$. Then the pdf for x is $f_1(x) = \frac{1}{2}$ at each $x \in [-1, 1]$ while the pdf for y is

$$f_2(y) = \begin{cases} 1, & \text{if } y \in [0, 1]; \\ 0, & \text{if } y \in [-1, 0). \end{cases}$$

Definition 1 *Say that random vectors $\mathbf{x} \in \mathbf{R}^d$ and $\mathbf{y} \in \mathbf{R}^p$ are* independent *if all pairs of their coordinates $\mathbf{x}(i), \mathbf{y}(j)$, $i = 1, \ldots, d$, $j = 1, \ldots, p$, are independent random variables.*

Corollary 0.4 *If $\mathbf{x} \in \mathbf{R}^d$ and $\mathbf{y} \in \mathbf{R}^p$ are independent random vectors, then* $\operatorname{cov}(\mathbf{x}, \mathbf{y}) = 0$.

Proof: For each $i = 1, \ldots, d$ and $j = 1, \ldots, p$, the coordinates $\mathbf{x}(i)$ and $\mathbf{y}(j)$ are independent random variables, so $\operatorname{cov}(\mathbf{x}(i), \mathbf{y}(j)) = 0$ by Corollary 0.3. Thus, each element of the $d \times p$ matrix $\operatorname{cov}(\mathbf{x}, \mathbf{y})$ is zero. □

Lemma 0.5 *If $\mathbf{x}, \mathbf{y} \in \mathbf{R}^d$ are independent 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

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

The last two covariances are zero by Corollary 0.4. □

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 independent random vectors, then $A\mathbf{x} + \mathbf{b}$ and $\mathbf{y}$ are also independent random vectors.*

Proof: □

Lemma 0.7 *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*

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

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

Combining Lemmas 0.5 and 0.7 gives us the desired computing tool:

Corollary 0.8 *, If $\mathbf{x} \in \mathbf{R}^d$ and $\mathbf{y} \in \mathbf{R}^p$ are mutually independent random vectors, then*

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

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

0.1.2 Permutations

A *permutation* is a one-to-one correspondence between a finite set and itself. If there are n elements in the set, it may be imagined as a rearrangement of the set of positive integers from 1 to n. Writing $[n] \stackrel{\text{def}}{=} \{1, \ldots, n\}$ for this set, the one-to-one correspondence conditions on $p : [n] \to [n]$ become:

(p1) If $p(i) = p(j)$, then $i = j$, and

(p2) For every $j \in [n]$ there is some $i \in [n]$ such that $p(i) = j$.

There are $n!$ distinct permutations on $[n]$, one for every ordering.

A *transposition* is a permutation that exchanges exactly two elements. Thus $\tau : [n] \to [n]$ is a transposition if and only if

(t1) There are distinct $i, j \in [n]$ such that $\tau(i) = j$, $\tau(j) = i$, and

(t2) $\tau(k) = k$ for all $k \notin \{i, j\}$.

There are $\binom{n}{2}$ transpositions on $[n]$, one for every choice of distinct i, j.

Lemma 0.9 *If $\tau : [n] \to [n]$ is a transposition, then $\tau^2(k) = k$ for all $k \in n$. Thus $\tau^{-1} = \tau$.*

Proof: Check that $\tau^2(i) = \tau(\tau(i)) = \tau(j) = i$ and $\tau^2(j) = \tau(\tau(j)) = \tau(i) = j$, while $\tau^2(k) = \tau(\tau(k)) = \tau(k) = k$ for all $k \notin \{i, j\}$. Thus τ^2 is the identity permutation, so $\tau^{-1} = \tau$. □

Theorem 0.10 *Let $p : [n] \to [n]$ be a permutation. Then p may be written as a composition of transpositions.*

Proof: Use induction on n. The result is evidently true for $n = 1$ where there is only the identity permutation which is a composition of no transpositions. It is also evidently true for $n = 2$ since then there are just two permutations, the identity and the transposition that exchanges 1 and 2.

Now suppose that the result is true for $n - 1$. Given $p : [n] \to [n]$, let $j \in [n]$ be the element satisfying $p(j) = n$. If $j = n$, then p leaves the n^{th} element fixed, so $p : [n - 1] \to [n - 1]$. By the induction hypothesis, p is a composition of transpositions of the first $n - 1$ elements, all of which leave the n^{th} element fixed.

Otherwise, for $j < n$, let τ be the transposition that exchanges elements n and j. Then the composition $p' \stackrel{\text{def}}{=} p\tau$ is a permutation on $[n]$ that satisfies $p'(n) = p(\tau(n)) = p(j) = n$. By the previous case, we know that p' is a composition of transpositions. But then $p = p'\tau^{-1} = p'\tau$ is also composition of transpositions. □

Define the *sign* of a permutation to be

$$\operatorname{sgn} p = \prod_{1 \le i < j \le n} \frac{p(j) - p(i)}{j - i}.$$

Note that for each pair i, j with $1 \le i < j \le n$ we will either have

(+): $p(i) < p(j)$, or

(−): $p(i) > p(j)$.

Since p is a one-to-one correspondence, the numerator factor $p(j) - p(i)$ will cancel some unique denominator factor of equal magnitude, leaving a quotient of $+1$ in case (+) or -1 in case (−). The product of these $\binom{n}{2}$ cancellation quotients will therefore be either $+1$ or -1. Hence $\operatorname{sgn} p \in \{-1, +1\}$.

Lemma 0.11 *Suppose $n \ge 2$. Let $p : [n] \to [n]$ be a permutation and $\tau : [n] \to [n]$ be a transposition. Then $\operatorname{sgn} p\tau = -\operatorname{sgn} p$.*

Proof: Without loss of generality, by renaming elements if necessary, we may assume that τ exchanges elements 1 and 2. But then we may decompose the product that defines $\operatorname{sgn} p\tau$ into factors that can be simplified:

$$\begin{aligned}
\operatorname{sgn} p\tau &= \prod_{1\le i<j\le n} \frac{p\tau(j)-p\tau(i)}{j-i} \\
&= \frac{p\tau(2)-p\tau(1)}{2-1} \\
&\quad \times \prod_{2<j\le n} \frac{p\tau(j)-p\tau(1)}{j-1} \prod_{2<j\le n} \frac{p\tau(j)-p\tau(2)}{j-2} \\
&\quad \times \prod_{2<i<j\le n} \frac{p\tau(j)-p\tau(i)}{j-i} \\
&= \frac{p(1)-p(2)}{2-1} \\
&\quad \times \prod_{2<j\le n} \frac{p(j)-p(2)}{j-1} \prod_{2<j\le n} \frac{p(j)-p(1)}{j-2} \\
&\quad \times \prod_{2<i<j\le n} \frac{p(j)-p(i)}{j-i} \\
&= \frac{p(1)-p(2)}{2-1} \prod_{\substack{1\le i<j\le n;\\ (i,j)\ne(1,2)}} \frac{p(j)-p(i)}{j-i} \\
&= -\frac{p(2)-p(1)}{2-1} \prod_{\substack{1\le i<j\le n;\\ (i,j)\ne(1,2)}} \frac{p(j)-p(i)}{j-i} \\
&= -\prod_{1\le i<j\le n} \frac{p(j)-p(i)}{j-i} \;=\; -\operatorname{sgn} p,
\end{aligned}$$

since exactly one of the factors in the product has changed sign. □

Lemma 0.12 *Let $p : [n] \to [n]$ be a permutation written as a composition of transpositions:*

$$p = \tau_1 \cdots \tau_m.$$

Then $\operatorname{sgn} p = (-1)^m$.

Proof: Use induction on m. If $m = 0$, then $p = Id$ so

$$\operatorname{sgn} p = \operatorname{sgn} Id = \prod_{1\le i<j\le n} \frac{j-i}{j-i} = +1 = (-1)^0.$$

Now assume that the result holds for m and let

$$p = \tau_1 \cdots \tau_m \tau_{m+1} \stackrel{\text{def}}{=} p' \tau_{m+1}.$$

Then $\operatorname{sgn} p' = (-1)^m$ by the inductive hypothesis, while $\operatorname{sgn} p = -\operatorname{sgn} p'$ by Lemma 0.11, so $\operatorname{sgn} p = (-1)^{m+1}$. □

Corollary 0.13 *Let $p : [n] \to [n]$ be a permutation which can be written as a composition of transpositions in two ways:*

$$p = \tau_1 \cdots \tau_m = \tau'_1 \cdots \tau'_{m'}.$$

Then m, m' are either both even or both odd.

Proof: By Lemma 0.12, $\operatorname{sgn} p = (-1)^m = (-1)^{m'}$. Conclude that m and m' have the same parity. □

Theorem 0.14 *If p, p' are permutations of the same set, then* $\operatorname{sgn} pp' = (\operatorname{sgn} p)(\operatorname{sgn} p')$.

Proof: By Theorem 0.10, write $p = \tau_1 \cdots \tau_m$ and $p' = \tau'_1 \cdots \tau'_{m'}$. Then pp' is a composition of $m + m'$ transpositions. Using Lemma 0.12, compute

$$\operatorname{sgn} p = (-1)^m; \quad \operatorname{sgn} p' = (-1)^{m'}; \quad \operatorname{sgn} pp' = (-1)^{m+m'} = (-1)^m (-1)^{m'} = \operatorname{sgn} p \operatorname{sgn} p',$$

yielding the result. □

Corollary 0.15 *If p is a permutation, then* $\operatorname{sgn} p^{-1} = \operatorname{sgn} p$.

Proof: Write $p = \tau_1 \cdots \tau_m$ as a product of m transpositions. By Lemma 0.9, $p^{-1} = \tau_m^{-1} \cdots \tau_1^{-1} = \tau_m \cdots \tau_1$ is composed of the same number of transpositions. By Lemma 0.12, $\operatorname{sgn} p^{-1} = (-1)^m = \operatorname{sgn} p$. □

0.1.3 Matrix calculus

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

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

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

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

$$\nabla_X \mathrm{tr}\,(AXB) = A^T B^T. \tag{1}$$

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

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

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

$$\nabla_X \mathrm{tr}\,(XAX^T) = XA^T + XA. \tag{3}$$

Likewise,

$$\nabla_X \mathrm{tr}\,(X^T AX) = A^T X + AX. \tag{4}$$

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

$$\nabla_X \mathrm{tr}\,(XX^T) = \nabla_X \mathrm{tr}\,(X^T X) = 2X. \tag{5}$$

0.1.4 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 independent random variables with mean zero and variances $\mathrm{Var}(\mathbf{q}_i) = Q_i$, that are also independent of $\mathbf{x}_0$. Then by Lemma 0.8,

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

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 independent mean-zero random variables that are also independent of $\{\mathbf{x}_i : i = 0, 1, \ldots\}$ 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 $\hat{P}_n \stackrel{\text{def}}{=} \operatorname{Var}(\mathbf{x}_n - \hat{\mathbf{x}}_n)$, 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 $\hat{\mathbf{y}}_n \stackrel{\text{def}}{=} H_n \left(A_n \hat{\mathbf{x}}_{n-1} + \mathbf{b}_n \right)$ 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:

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

- Predicted (*a priori*) measurement:

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

- Innovation:

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

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

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

We will choose that K_n that maximizes the decrease in estimator variance $\hat{P}_{n-1} \to \hat{P}_n$.

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

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

- Innovation variance:

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

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 independent. 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:

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

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

$$K_n = \hat{P}_{n|n-1} H_n^T S_n^{-1}. \tag{6}$$

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 $\hat{\mathbf{y}}_n = H_n \hat{\mathbf{x}}_{n|n-1}$;
- 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 independent random vectors with known means and variances. The state estimate and its variance may be initialized by

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

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

0.2 Exercises

For the following problems, write $[n] \stackrel{\text{def}}{=} \{1, \ldots, n\}$ for the set of positive integers from 1 to n.

1. One way to define the determinant of a square matrix $A = (a_{ij}) \in \mathbf{R}^{n \times n}$ is

$$\det A = \sum_{p \in P(n)} (\operatorname{sgn} p) \prod_{i=1}^{n} a_{ip(i)}.$$

   Prove from this definition that $\det A^T = \det A$.

2. The *permanent* of a square matrix $A = (a_{ij}) \in \mathbf{R}^{n\times n}$ is

$$\operatorname{perm} A = \sum_{p\in P(n)} \prod_{i=1}^{n} a_{ip(i)}.$$

Prove from this definition that $\operatorname{perm} A^T = \operatorname{perm} A$.

3. Let X be a square matrix. Find $\nabla_X(\operatorname{perm} X)$.

4. Let X be a square matrix. Find $\nabla_X(\det X)$.

5. Define the *minor matrix* A_{ij} of a matrix $A \in \mathbf{R}^{n\times n}$ to be the $(n-1)\times(n-1)$ matrix obtained from A by removing row i and column j. Prove *Laplace's formula* for the determinant: for fixed row i or fixed column j,

$$\begin{aligned} \det A &= a_{i1}(-1)^{i+1}\det A_{i1} + \cdots + a_{in}(-1)^{i+n}\det A_{in} \\ &= a_{1j}(-1)^{1+j}\det A_{1j} + \cdots + a_{nj}(-1)^{n+j}\det A_{nj}. \end{aligned}$$

6. Compute $\nabla_X \det(X) = \det(X)X^{-1}$ whenever X is invertible. Hint: write $\det(X)$ as an expansion in minors.

0.3 Solutions to the Exercises

For the following problems, write $[n] \stackrel{\text{def}}{=} \{1,\ldots,n\}$ for the set of positive integers from 1 to n.

1. One way to define the determinant of a square matrix $A = (a_{ij}) \in \mathbf{R}^{n\times n}$ is

$$\det A = \sum_{p\in P(n)} (\operatorname{sgn} p) \prod_{i=1}^{n} a_{ip(i)}.$$

Prove from this definition that $\det A^T = \det A$.

Solution: Since $A^T = (a^T_{ij})$ satisfies $a^T_{ij} = a_{ji}$, write

$$\begin{aligned} \det A^T &= \sum_{p\in P(n)} (\operatorname{sgn} p) \prod_{i=1}^{n} a^T_{ip(i)} \\ &= \sum_{p\in P(n)} (\operatorname{sgn} p) \prod_{i=1}^{n} a_{p(i)i} \\ &= \sum_{p\in P(n)} (\operatorname{sgn} p) \prod_{i=1}^{n} a_{ip^{-1}(i)} \end{aligned}$$

$$= \sum_{p^{-1}\in P(n)} (\operatorname{sgn} p^{-1}) \prod_{i=1}^{n} a_{ip(i)}$$

$$= \sum_{p^{-1}\in P(n)} (\operatorname{sgn} p) \prod_{i=1}^{n} a_{ip(i)} \;=\; \det A,$$

since the sum ranges over all $n!$ permutations whether we label them p or p^{-1}. □

2. The *permanent* of a square matrix $A = (a_{ij}) \in \mathbf{R}^{n\times n}$ is

$$\operatorname{perm} A = \sum_{p\in P(n)} \prod_{i=1}^{n} a_{ip(i)}.$$

Prove from this definition that $\operatorname{perm} A^T = \operatorname{perm} A$.

Solution: Since $A^T = (a_{ij}^T)$ satisfies $a_{ij}^T = a_{ji}$, write

$$\begin{aligned}\operatorname{perm} A^T &= \sum_{p\in P(n)} \prod_{i=1}^{n} a_{ip(i)}^T \;=\; \sum_{p\in P(n)} \prod_{i=1}^{n} a_{p(i)i} \\ &= \sum_{p\in P(n)} \prod_{i=1}^{n} a_{ip^{-1}(i)} \;=\; \sum_{p^{-1}\in P(n)} \prod_{i=1}^{n} a_{ip(i)} \\ &= \sum_{p^{-1}\in P(n)} \prod_{i=1}^{n} a_{ip(i)} \;=\; \operatorname{perm} A,\end{aligned}$$

since the sum ranges over all $n!$ permutations whether we label them p or p^{-1}. □

3. Let X be a square matrix. Find $\nabla_X(\operatorname{perm} X)$.

Solution: Writing $X = (x_{ij}) \in \mathbf{R}^{n\times n}$, the k,l element of $\nabla_X(\operatorname{perm} X)$ will be:

$$\frac{\partial}{\partial x_{kl}} \operatorname{perm} X = \sum_{\{p\in P(n):p(k)=l\}} \prod_{i\neq k} x_{ip(i)} = \operatorname{perm} X_{kl},$$

where X_{kl} is the $(n-1)\times(n-1)$ *minor matrix* obtained from X by removing row k and column l. □

4. Let X be a square matrix. Find $\nabla_X(\det X)$.

Solution: Writing $X = (x_{ij}) \in \mathbf{R}^{n\times n}$, the k,l element of $\nabla_X(\det X)$ will be:

$$\frac{\partial}{\partial x_{kl}} \det X = \sum_{\{p\in P(n):p(k)=l\}} (\operatorname{sgn} p) \prod_{i\neq k} x_{ip(i)} = \det X_{kl},$$

where X_{kl} is the $(n-1)\times(n-1)$ *minor matrix* obtained from X by removing row k and column l. □

5. Define the *minor matrix* A_{ij} of a matrix $A \in \mathbf{R}^{n\times n}$ to be the $(n-1)\times(n-1)$ matrix obtained from A by removing row i and column j. Prove *Laplace's formula* for the determinant: for fixed row i or fixed column j,

$$\begin{aligned}\det A &= a_{i1}(-1)^{i+1}\det A_{i1} + \cdots + a_{in}(-1)^{i+n}\det A_{in} \\ &= a_{1j}(-1)^{1+j}\det A_{1j} + \cdots + a_{nj}(-1)^{n+j}\det A_{nj}.\end{aligned}$$

Solution: □

6. Compute $\nabla_X \det(X) = \det(X)X^{-1}$ whenever X is invertible. Hint: write $\det(X)$ as an expansion in minors.

Solution: Following the hint, for square matrix $X \in \mathbf{R}^{n\times n}$ write

$$\begin{aligned}\det X &= x_{i1}M_{i1} + \cdots + x_{ij}M_{ij} + \cdots + x_{in}M_{in} \\ &= x_{1j}M_{1j} + \cdots + x_{ij}M_{ij} + \cdots + x_{nj}M_{nj},\end{aligned}$$

where i, j are any fixed row or column indices, respectively, and M_{ij} is the signed minor of X, namely $(-1)^{i+j}\det X_{ij}$ where X_{ij} is the $(n-1)\times(n-1)$ matrix obtained from X by removing row i and column j. □