Kalman Folding 3: Derivations (WORKING DRAFT)

Extracting Models from Data, One Observation at a Time

Brian Beckman

<2016-05-03 Tue>

Contents

1 Abstract

In *Kalman Folding, Part 1*,[1] we present basic, static Kalman filtering as a functional fold, highlighting the unique advantages of this form for deploying test-hardened code verbatim in harsh, mission-critical environments. The examples in that paper are all static, meaning that the states of the model do not depend on the independent variable, often physical time.

Here, we present mathematical derivations of the basic, static filter. These are semi-formal sketches that leave many details to the reader, but highlight all important points that must be rigorously proved. These derivations have several novel arguments and we strive for much higher clarity and simplicity than is found in most treatments of the topic.

2 Kalman Folding in the Wolfram Language

In this series of papers, we use the Wolfram language[2] because it excels at concise expression of mathematical code. All examples in these papers can be directly transcribed to any modern mainstream language that supports closures. For example, it is easy to write them in C++11 and

[1] B. Beckman, *Kalman Folding, Part 1*, to appear.

[2] `http://reference.wolfram.com/language/`

beyond, Python, any modern Lisp, not to mention Haskell, Scala, Erlang, and OCaml. Many can be written without full closures; function pointers will suffice, so they are easy to write in C. It's also not difficult to add extra arguments to simulate just enough closure-like support in C to write the rest of the examples in that language.

In *Kalman Folding*,[1] we found the following small formulation for the accumulator function of a fold that implements the static Kalman filter:

$$\text{kalmanStatic}\left(\mathbf{Z}\right)\left(\{\boldsymbol{x},\mathbf{P}\},\{\mathbf{A},\boldsymbol{z}\}\right)=\{\boldsymbol{x}+\mathbf{K}\left(\boldsymbol{z}-\mathbf{A}\,\boldsymbol{x}\right),\mathbf{P}-\mathbf{K}\,\mathbf{D}\,\mathbf{K}^{\intercal}\} \tag{1}$$

where

$$\mathbf{K}=\mathbf{P}\,\mathbf{A}^{\intercal}\,\mathbf{D}^{-1} \tag{2}$$

$$\mathbf{D}=\mathbf{Z}+\mathbf{A}\,\mathbf{P}\,\mathbf{A}^{\intercal} \tag{3}$$

and all quantities are matrices:

- $\boldsymbol{z}$ is a $\mathfrak{b}\times 1$ column vector containing one multidimensional observation
- $\boldsymbol{x}$ is an $\mathfrak{n}\times 1$ column vector of *model states*
- $\mathbf{Z}$ is a $\mathfrak{b}\times\mathfrak{b}$ matrix, the covariance of observation noise
- $\mathbf{P}$ is an $\mathfrak{n}\times\mathfrak{n}$ matrix, the theoretical covariance of $\boldsymbol{x}$
- $\mathbf{A}$ is a $\mathfrak{b}\times\mathfrak{n}$ matrix, the *observation partials*
- $\mathbf{D}$ is a $\mathfrak{b}\times\mathfrak{b}$ matrix, the Kalman denominator
- $\mathbf{K}$ is an $\mathfrak{n}\times\mathfrak{b}$ matrix, the Kalman gain

In physical or engineering applications, these quantities carry physical dimensions of units of measure in addition to their matrix dimensions as numbers of rows and columns. If the physical and matrix dimensions of $\boldsymbol{x}$ are $[[\boldsymbol{x}]]\stackrel{\text{def}}{=}(\mathfrak{X},\mathfrak{n}\times 1)$ and of $\boldsymbol{z}$ are $[[\boldsymbol{z}]]\stackrel{\text{def}}{=}(\mathfrak{Z},\mathfrak{b}\times 1)$, then

$$\begin{array}{lcccc}
[[\mathbf{Z}]] & = & (& \mathfrak{Z}^2 & \mathfrak{b}\times\mathfrak{b} \;) \\
[[\mathbf{A}]] & = & (& \mathfrak{Z}/\mathfrak{X} & \mathfrak{b}\times\mathfrak{n} \;) \\
[[\mathbf{P}]] & = & (& \mathfrak{X}^2 & \mathfrak{n}\times\mathfrak{n} \;) \\
[[\mathbf{A}\,\mathbf{P}\,\mathbf{A}^{\intercal}]] & = & (& \mathfrak{Z}^2 & \mathfrak{b}\times\mathfrak{b} \;) \\
[[\mathbf{D}]] & = & (& \mathfrak{Z}^2 & \mathfrak{b}\times\mathfrak{b} \;) \\
[[\mathbf{P}\,\mathbf{A}^{\intercal}]] & = & (& \mathfrak{X}\,\mathfrak{Z} & \mathfrak{n}\times\mathfrak{b} \;) \\
[[\mathbf{K}]] & = & (& \mathfrak{X}/\mathfrak{Z} & \mathfrak{n}\times\mathfrak{b} \;)
\end{array} \tag{4}$$

Dimensional arguments, regarding both matrix dimensions and physical dimensions, are invaluable for checking the derivations that follow.

3 Derivations

Here, we derive equations 1, 2 and 3. Again, these derivations are just sketches designed for clarity. These derivations only cover the static Kalman filter, where $\boldsymbol{x}$ are fixed, constant, static states of the model. See Bar-Shalom[3] for derivations of the time-dependent Kalman filter and part 2 of this

[3] Bar-Shalom, Yaakov, *et al.* Estimation with applications to tracking and navigation. New York: Wiley, 2001.

series[4] for intuitive arguments.

The plan is first to develop expressions for the prior estimate $\tilde{\mathbf{x}}$ and covariance $\tilde{\mathbf{P}}$, then expressions for the posterior versions $\hat{\mathbf{x}}$ and $\hat{\mathbf{P}}$, defining the Kalman gain $\mathbf{K}$ and the denominator matrix $\mathbf{D}$ along the way. Finally, we derive the convenient expressions for $\mathbf{K}$ and $\mathbf{D}$ that appear in equations 1, 2, and 3.

3.1 Notation

The word *vector* alone means *column vector* by default. If a quantity is a row vector, we explicitly say so. In general, lower-case boldface symbols like $\mathbf{x}$ denote column vectors. Row vectors include a superscript *transpose* symbol, as in $\mathbf{a}^\intercal$. We write literal vectors in square brackets, as in $[a, b, \ldots]^\intercal$ for a column vector or $[a, b, \ldots]$ for a row vector or when we don't care whether it's a column or row.

Upper-case boldface symbols like $\mathbf{M}$ denote matrices. Because vectors are special cases of matrices, some matrices are also vectors. We may use an upper-case symbol to denote a vector, but we do not use a lower-case symbol to denote a non-vector matrix.

Juxtaposition, as in $\mathbf{A}\,\mathbf{x}$ or $\mathbf{A}\,\mathbf{B}$, means matrix multiplication. We occasionally use a center dot or $\times$ symbol to clarify matrix multiplication, as in $\mathbf{A} \cdot \mathbf{x}$ or $\mathbf{A} \times \mathbf{x}$.

We freely and frequently exploit the following facts without pointing out when we use them.

- For any matrix $\mathbf{M}$, $(\mathbf{M}^\intercal)^\intercal = \mathbf{M}$
- For any invertible matrix $\mathbf{M}$, $\left(\mathbf{M}^{-1}\right)^{-1} = \mathbf{M}$
- For any two matrices $\mathbf{A}$ and $\mathbf{B}$, $(\mathbf{A}\,\mathbf{B})^\intercal = \mathbf{B}^\intercal\mathbf{A}^\intercal$
- $(\mathbf{A}\,\mathbf{B})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$ when the matrices are invertible
- $\mathbf{P}^\intercal = \mathbf{P}$ if and only if $\mathbf{P}$ is symmetric

For any matrix $\mathbf{M}$, $\mathbf{M}^2$ means $\mathbf{M}^\intercal\mathbf{M}$, the transpose of the matrix times the matrix. Such squared matrices are always square and symmetric. This notation pertains to vectors, as well, because they are just special cases of matrices. Thus, $\mathbf{x}^2 = \mathbf{x}^\intercal\mathbf{x}$, the Euclidean 2-norm of $\mathbf{x}$, a scalar; and $(\mathbf{x}^\intercal)^2 = (\mathbf{x}^\intercal)^\intercal \cdot \mathbf{x}^\intercal = \mathbf{x}\,\mathbf{x}^\intercal$ is the outer product of $\mathbf{x}$ with itself. That outer product is an $n \times n$ square, symmetric matrix, where n is the dimensionality of $\mathbf{x}$.

When $\mathbf{M}^2$ is invertible, $\mathbf{M}^{-2}$ means the inverse of $\mathbf{M}^2$, namely $(\mathbf{M}^\intercal\mathbf{M})^{-1}$.

We use the term *tall* to mean a matrix with more rows than columns, that is, an $m \times n$ matrix when $m > n$. When discussing $m \times n$ matrices, we usually assume that $m > n$. We use the term *wide* to mean a matrix with more columns than rows, as in an $n \times m$ matrix. We use the term *small* to mean $n \times n$, and *large* to mean $m \times m$.

3.2 Definitions

t is the independent variable. In many applications, t represents physical time, or an integer index mapped to physical time. It is known and non-random. We treat it as a scalar, here, though it is possible to extend the theory to a vector t.

[4] B. Beckman, *Kalman Folding 2: Tracking and System Dynamics*, to appear.

$\mathbf{x}$ is the (column) vector of n unknown, constant *states* of the model. It's a random variable, and we compute estimates and covariances *via* expectation values over its distribution. This symbol also means an algebraic variable standing for some particular estimate of the states.

$\mathbf{A}\mathbf{x}$ the *model*; it predicts an observation at time t given an estimate of the states $\mathbf{x}$ and a current partials matrix $\mathbf{A}$ that depends on t. The model is a column vector of dimensionality $b \times 1$, the dimensionality of an observation $\mathbf{z}$.

$\mathbf{A}$ is the *current partials matrix*, the partial derivative of the model with respect to the unknown states $\mathbf{x}$, evaluated at the current value of the independent variable t. We could write $\mathbf{A}$ as $\mathbf{A}(t)$, and perhaps we should; it's an aesthetic judgment not to write the t dependence explicitly because it would make the derivations so much longer and harder to read. Because the model is *linear*, the partials do not depend on $\mathbf{x}$. $\mathbf{A}$ is known, non-random, and depends only on t. Generally, its dimensionality is $b \times n$, where b is the dimensionality of an observation $\mathbf{z}$.

$\tilde{\mathbf{A}}$ is the *prior partials matrix*, a matrix that stacks all the prior rows of $\mathbf{A}$ that precede the current row. It is known, non-random, and $mb \times n$, where m is the number of prior observations, b is the dimensionality of an observation $\mathbf{z}$, and n is the dimensionality of the states $\mathbf{x}$. Thus $\tilde{\mathbf{A}}$ is tall in the typical *overdetermined* case where $m > n$, more observations than states. We do not actually realize $\tilde{\mathbf{A}}$ in computer memory because Kalman keeps *all information* in the running covariance matrix. $\tilde{\mathbf{A}}$ is just a useful abstraction in the derivations below.

$\mathbf{z}$ is the *current observation*. It is known and non-random. Its dimensionality is $b \times 1$, b perhaps suggesting 'bundle.'

$\tilde{\mathbf{z}}$ is a stack or *batch* of all prior observations. It is known, non-random, $mb \times 1$. It's a useful abstraction in the derivations below. It's not necessary to actually realize it in computer memory because we use all its information incrementally by folding.

$\tilde{\mathbf{x}}$ the *prior estimate*, the estimate of $\mathbf{x}$ given all information we have prior to the current observation. It is known, non-random, $n \times 1$.

$\hat{\mathbf{x}}$ the *posterior estimate*, the estimate of $\mathbf{x}$ given (1) the prior estimate $\tilde{\mathbf{x}}$, (2) the current partials $\mathbf{A}$, and (3) the current observation $\mathbf{z}$. It is known, non-random, $n \times 1$. It satisfies *the Kalman update equation*:

$$\hat{\mathbf{x}} = \tilde{\mathbf{x}} + \mathbf{K}(\mathbf{z} - \mathbf{A}\tilde{\mathbf{x}}) \tag{5}$$

which is equivalent to the recurrence $\mathbf{x} \leftarrow \mathbf{x} + \mathbf{K}(z - \mathbf{A}\mathbf{x})$ used in part 1 of this series.

$\tilde{\mathbf{P}}$ *covariance of the priors*, equals $\mathbf{Z}\left(\tilde{\mathbf{A}}^{\intercal}\tilde{\mathbf{A}}\right)^{-1} \stackrel{\text{def}}{=} \mathbf{Z}\tilde{\mathbf{A}}^{-2}$ (proof sketch below). This is called just $\mathbf{P}$ in part one of this series. It is known, non-random, $n \times n$.

$\hat{\mathbf{P}}$ *posterior covariance*, satisfies $\hat{\mathbf{P}}\mathbf{A}^{\intercal} = \mathbf{Z}\mathbf{K} = \mathbf{Z}\tilde{\mathbf{P}}\mathbf{A}^{\intercal}\mathbf{D}^{-1}$ (proof sketch below). We calculate it from the prior covariance $\tilde{\mathbf{P}}$, the observation-noise covariance $\mathbf{Z}$, and the new partials matrix $\mathbf{A}$. It is known, non-random, $n \times n$.

$\mathbf{A}\tilde{\mathbf{x}}$ the *predicted observation* given the prior estimate $\tilde{\mathbf{x}}$ and the current partials matrix $\mathbf{A}$. It is a particular evaluation of the model. It is known, non-random, $b \times 1$.

$z - \mathbf{A}\tilde{x}$ the measurement *residual*, the difference between the current observation and the predicted observation.

ζ *observation noise*, random, column-vector variable with zero mean and unit covariance. Its dimensionality is $b \times 1$, that of the observation z. Its mean is $E[\zeta] = 0$ and its covariance is $E[\zeta\,\zeta^T] = \mathbf{Z}$: known, non-random $b \times b$.

$\mathbf{Z}$ covariance of the observation noise, $E[\zeta\,\zeta^T] = \mathbf{Z}$: known, non-random $b \times b$.

$\tilde{z} = \tilde{\mathbf{A}}\,x + \zeta$ the *observation equation*. $\tilde{z}$ is known, non-random, $mb \times 1$; $\tilde{\mathbf{A}}$ is known, non-random, $mb \times n$; x is unknown, random, $n \times 1$; ζ is unknown, random, $mb \times 1$.

$\mathbf{K}$ *Kalman gain* $= \tilde{\mathbf{P}}\,\mathbf{A}^T\,\mathbf{D}^{-1}$ (proof sketch below). Non-random, $n \times b$.

$\mathbf{D}$ *Kalman denominator* $= \mathbf{Z} + \mathbf{A}\,\tilde{\mathbf{P}}\,\mathbf{A}^T$ (proof sketch below). Non-random, $b×{b}$f.

3.3 Demonstration that Prior Covariance $\tilde{\mathbf{P}} = \mathbf{Z}\,\tilde{\mathbf{A}}^{-2}$

The fact that the prior covariance, $\tilde{\mathbf{P}}$, equals $\mathbf{Z}\,\tilde{\mathbf{A}}^{-2}$, which is a tall matrix that stacks all m prior model partial derivatives, means that all the information about the model is carried along in one, small $n \times n$ matrix. This is the secret to Kalman's constant-memory usage.

3.3.1 Covariance of a Random Vector Variable

The covariance of any random column-vector variable y is defined as the expectation value $E[y\,y^T] = E\left[(y^T)^2\right]$ This is the expectation value of an outer product of a column vector y and its transpose, y^T. Therefore, it is a $q \times q$ matrix, where $q \times 1$ is the dimensionality of y.

3.3.2 Prior Estimate $\tilde{x}$

One of our random variables is x, the column vector of unknown states. To calculate its estimate, assume we know the values of all m past partials $\tilde{\mathbf{A}}$ (tall, $mb \times n$) and observations $\tilde{z}$ (tall, $mb \times 1$).

Relate x to the known observations $\tilde{z}$ and the known partials $\tilde{\mathbf{A}}$ through the normally distributed random noise column vector ζ and the observation equation:

$$\tilde{z} = \tilde{\mathbf{A}}\,x + \zeta \tag{6}$$

3.3.3 Sum of Squared Residuals

Consider the following *performance functional*, computed over the population of x.

$$J(x) \stackrel{\text{def}}{=} \zeta^2 = \left(\tilde{z} - \tilde{\mathbf{A}}\,x\right)^2 = \left(\tilde{z} - \tilde{\mathbf{A}}\,x\right)^T \cdot \left(\tilde{z} - \tilde{\mathbf{A}}\,x\right)$$

$J(x)$ is a scalar: the sum of squared residuals. A *residual* is a difference between an actual and a predicted observation. To find the x that minimizes $J(x)$, we could take the classic, school approach of setting to zero the partial derivatives of $J(x)$ with respect to x and solving the resulting equations for x. The following is an easier way. Multiply the residuals across by the wide matrix $\tilde{\mathbf{A}}^T$:

$$\tilde{\mathbf{A}}^T\,\tilde{z} - \tilde{\mathbf{A}}^2\,x$$

producing an n-vector, and then construct a modified performance functional:

$$J'(\mathbf{x}) \stackrel{\text{def}}{=} \left(\tilde{\mathbf{A}}^T \tilde{z} - \tilde{\mathbf{A}}^2 \mathbf{x}\right)^2 = \left(\tilde{\mathbf{A}}^T \tilde{z} - \tilde{\mathbf{A}}^2 \mathbf{x}\right)^T \cdot \left(\tilde{\mathbf{A}}^T \tilde{z} - \tilde{\mathbf{A}}^2 \mathbf{x}\right)$$

$J(\mathbf{x})$ is minimum with respect to $\mathbf{x}$ if and only if (iff) $J'(\mathbf{x})$ is minimum. Because $J'(\mathbf{x})$ is non-negative, when $J'(\mathbf{x})$ *can* be zero, its minimum *must* be zero. $J'(\mathbf{x})$ is zero iff $\tilde{\mathbf{A}}^2$, an $n \times n$ square matrix, is invertible (non-singular) and

$$\mathbf{x} = \tilde{\mathbf{A}}^{-2} \tilde{\mathbf{A}}^T \tilde{z}$$

because then

$$\tilde{\mathbf{A}}^T \tilde{z} = \tilde{\mathbf{A}}^2 \mathbf{x}$$

We call such a solution for $\mathbf{x}$ the *least-squares estimate* of $\mathbf{x}$, the estimate of $\mathbf{x}$ based on all prior observations. From now on, we write it as $\tilde{\mathbf{x}}$

$$\tilde{\mathbf{x}} \stackrel{\text{def}}{=} \tilde{\mathbf{A}}^{-2} \tilde{\mathbf{A}}^T \tilde{z} \tag{7}$$

With this solution, we get a new expression for the performance functional $J(\mathbf{x})$ that is useful below. First note that

$$\begin{aligned}
\tilde{\mathbf{A}}^2 \tilde{\mathbf{A}}^{-2} &= \mathbf{1} \\
\tilde{\mathbf{A}}^2 \tilde{\mathbf{A}}^{-2} \tilde{\mathbf{A}}^T &= \tilde{\mathbf{A}}^T && \text{Multiply on right by } \tilde{\mathbf{A}}^T \\
\tilde{\mathbf{A}}^T \tilde{\mathbf{A}} \tilde{\mathbf{A}}^{-2} \tilde{\mathbf{A}}^T &= \tilde{\mathbf{A}}^T && \text{Expand definition of } \tilde{\mathbf{A}}^2 \\
\therefore \tilde{\mathbf{A}} \tilde{\mathbf{A}}^{-2} \tilde{\mathbf{A}}^T &= \mathbf{1} && \text{Arbitrariness of } \tilde{\mathbf{A}}^T \text{on left}
\end{aligned} \tag{8}$$

Therefore

$$\begin{aligned}
J(\mathbf{x}) &= \left(\tilde{z} - \tilde{\mathbf{A}}\mathbf{x}\right)^T \cdot \left(\tilde{z} - \tilde{\mathbf{A}}\mathbf{x}\right) \\
&= \left(\tilde{z} - \tilde{\mathbf{A}}\mathbf{x}\right)^T \tilde{\mathbf{A}} \tilde{\mathbf{A}}^{-2} \tilde{\mathbf{A}}^T \left(\tilde{z} - \tilde{\mathbf{A}}\mathbf{x}\right) && \text{insert } \mathbf{1} \text{ from equation 8} \\
&= \left(\tilde{z} - \tilde{\mathbf{A}}\mathbf{x}\right)^T \tilde{\mathbf{A}} \tilde{\mathbf{A}}^{-2} \tilde{\mathbf{A}}^2 \tilde{\mathbf{A}}^{-2} \tilde{\mathbf{A}}^T \left(\tilde{z} - \tilde{\mathbf{A}}\mathbf{x}\right) && \text{insert } \mathbf{1} = \tilde{\mathbf{A}}^2 \tilde{\mathbf{A}}^{-2} \\
&= \left[\left(\tilde{z} - \tilde{\mathbf{A}}\mathbf{x}\right)^T \tilde{\mathbf{A}} \tilde{\mathbf{A}}^{-2}\right] \tilde{\mathbf{A}}^2 \left[\tilde{\mathbf{A}}^{-2} \tilde{\mathbf{A}}^T \left(\tilde{z} - \tilde{\mathbf{A}}\mathbf{x}\right)\right] && \text{Regroup} \\
&= (\tilde{\mathbf{x}} - \mathbf{x})^T \tilde{\mathbf{A}}^2 (\tilde{\mathbf{x}} - \mathbf{x}) && \text{Definition of } \tilde{\mathbf{x}}
\end{aligned} \tag{9}$$

using the fact that $\tilde{\mathbf{A}}^2$ is symmetric. This has physical dimensions $\mathcal{Z}^2$ where $\mathcal{Z}$ are the physical dimensions of the observations z.

3.3.4 Prior Covariance $\tilde{\mathbf{P}}$

We now want the covariance of the *residuals*, the differences between our least-squares estimate $\tilde{\mathbf{x}}$ and the random vector $\mathbf{x}$:

$$\tilde{\mathbf{P}} \stackrel{\text{def}}{=} E\left[(\tilde{\mathbf{x}} - x)(\tilde{\mathbf{x}} - x)^T\right] \tag{10}$$

Get $\tilde{\mathbf{x}} - \mathbf{x}$ from the observations and partials at hand as follows:

$$\begin{aligned}
\tilde{\mathbf{z}} &= \tilde{\mathbf{A}}\,\mathbf{x} + \zeta && \text{Equation 6} \\
\tilde{\mathbf{A}}^{-2}\,\tilde{\mathbf{A}}^{\intercal}\,\tilde{\mathbf{z}} &= \mathbf{x} + \tilde{\mathbf{A}}^{-2}\,\tilde{\mathbf{A}}^{\intercal}\,\zeta && \text{Multiply on left by } \tilde{\mathbf{A}}^{-2}\,\tilde{\mathbf{A}}^{\intercal} \\
\tilde{\mathbf{x}} &= \mathbf{x} + \tilde{\mathbf{A}}^{-2}\,\tilde{\mathbf{A}}^{\intercal}\,\zeta && \text{Definition of } \tilde{\mathbf{x}} \\
\therefore \tilde{\mathbf{x}} - \mathbf{x} &= \tilde{\mathbf{A}}^{-2}\tilde{\mathbf{A}}^{\intercal}\zeta
\end{aligned}$$

Now rewrite equation 10:

$$\begin{aligned}
\mathrm{E}\left[(\tilde{\mathbf{x}} - \mathrm{x})(\tilde{\mathbf{x}} - \mathrm{x})^{\intercal}\right] &= \mathrm{E}\left[\tilde{\mathbf{A}}^{-2}\tilde{\mathbf{A}}^{\intercal}\zeta\,\zeta^{\intercal}(\tilde{\mathbf{A}}^{-2}\tilde{\mathbf{A}}^{\intercal}\zeta)^{\intercal}\right] \\
&= \tilde{\mathbf{A}}^{-2}\tilde{\mathbf{A}}^{\intercal}\,\mathrm{E}\left[\zeta\,\zeta^{\intercal}\right](\tilde{\mathbf{A}}^{-2}\tilde{\mathbf{A}}^{\intercal})^{\intercal} \qquad (11)
\end{aligned}$$

Noise ζ is Gaussian, normal, with variance $\mathbf{Z}$. Equation 11 collapses to

$$\begin{aligned}
\tilde{\mathbf{P}} = \tilde{\mathbf{A}}^{-2}\tilde{\mathbf{A}}^{\intercal}\,\mathrm{E}\left[\zeta\,\zeta^{\intercal}\right](\tilde{\mathbf{A}}^{-2}\tilde{\mathbf{A}}^{\intercal})^{\intercal} &= \tilde{\mathbf{A}}^{-2}\tilde{\mathbf{A}}^{\intercal}\,\mathbf{Z}\,(\tilde{\mathbf{A}}^{-2}\tilde{\mathbf{A}}^{\intercal})^{\intercal} \\
&= \tilde{\mathbf{A}}^{-2}\tilde{\mathbf{A}}^{\intercal}\,\mathbf{Z}\,\tilde{\mathbf{A}}(\tilde{\mathbf{A}}^{-2})^{\intercal} \\
&= \mathbf{Z}\,\tilde{\mathbf{A}}^{-2}\tilde{\mathbf{A}}^{2}(\tilde{\mathbf{A}}^{-2})^{\intercal} \\
&= \mathbf{Z}\,(\tilde{\mathbf{A}}^{-2})^{\intercal} \\
&= \mathbf{Z}\,\tilde{\mathbf{A}}^{-2}
\end{aligned}$$

because $\tilde{\mathbf{A}}^{-2}$ is symmetric and because $\mathbf{Z}$ is diagonal and thus commutes with all other matrix products of compatible matrix dimension. We can now rewrite the definition of the least squares estimate in equation 7:

$$\tilde{\mathbf{x}} = \mathbf{Z}^{-1}\,\tilde{\mathbf{P}}\,\tilde{\mathbf{A}}^{\intercal}\,\tilde{\mathbf{z}} \qquad (12)$$

3.4 Posterior Estimate $\hat{\mathbf{x}}$ and Covariance $\hat{\mathbf{P}}$

To effect incremental updates of $\mathbf{x}$ and $\mathbf{P}$, we need the posterior estimate $\hat{\mathbf{x}}$ and covariance $\hat{\mathbf{P}}$ in terms of the priors $\tilde{\mathbf{x}}$, $\tilde{\mathbf{P}}$, and the new partials $\mathbf{A}$ and observation z. This is exactly what our *kalmanStatic* function from equation 1 does, of course, in functional form, but we derive the posteriors from scratch to seek opportunities to define $\mathbf{K}$ and $\mathbf{D}$ and radically shorten the expressions.

First, define a new performance functional $\mathrm{J}_1(\mathbf{x})$ as the sum of the performance of the priors $\tilde{\mathrm{J}}(\mathbf{x})$ from equation 9, now written with tildes overhead, and a new term $\mathrm{J}_2(\mathbf{x})$ for the performance of the new data:

$$\mathrm{J}_1(\mathbf{x}) \stackrel{\text{def}}{=} \tilde{\mathrm{J}}(\mathbf{x}) + \mathrm{J}_2(\mathbf{x}) \qquad (13)$$

$$\tilde{\mathrm{J}}(\mathbf{x}) \stackrel{\text{def}}{=} (\tilde{\mathbf{x}} - \mathbf{x})^{\intercal}\,\tilde{\mathbf{A}}^{2}\,(\tilde{\mathbf{x}} - \mathbf{x}) \qquad \text{Equation 9}$$

$$\begin{aligned}
\mathrm{J}_2(\mathbf{x}) &\stackrel{\text{def}}{=} (z - \mathbf{A}\,\mathbf{x})^2 \qquad (14) \\
&= (z - \mathbf{A}\,\mathbf{x})^{\intercal} \cdot (z - \mathbf{A}\,\mathbf{x}) \\
&= z^2 - 2\,z\,\mathbf{A}\,\mathbf{x} + (\mathbf{A}\,\mathbf{x})^2
\end{aligned}$$

This time, I don't have a handy trick for minimizing the performance functional. Let's find the minimizing $\mathbf{x}$ the classic way: by solving $\mathrm{d}\,\mathrm{J}_1(\mathbf{x})/\mathrm{d}\,\mathbf{x} = 0$. The usual way to write a vector derivative is with the *nabla* operator ∇, which produces *gradient* vectors from scalar functions.

$$\nabla\, \mathrm{f}(\mathbf{x}) \stackrel{\text{def}}{=} \begin{bmatrix} \mathrm{df}(\mathbf{x})/\mathrm{dx}_0 \\ \mathrm{df}(\mathbf{x})/\mathrm{dx}_1 \\ \vdots \\ \mathrm{df}(\mathbf{x})/\mathrm{dx}_{n-1} \end{bmatrix}$$

The particular scalar function we're differentiating is, of course, the new performance functional $\mathrm{J}_1(\mathbf{x}) = \tilde{\mathrm{J}}(\mathbf{x}) + \mathrm{J}_2(\mathbf{x})$. Because $\tilde{\mathbf{A}}^2$ is symmetric,

$$\begin{aligned} \nabla\, \tilde{\mathrm{J}}(\mathbf{x}) &= \nabla \left((\tilde{\mathbf{x}} - \mathbf{x})^\intercal\, \tilde{\mathbf{A}}^2\, (\tilde{\mathbf{x}} - \mathbf{x}) \right) \\ &= -2\, \tilde{\mathbf{A}}^2\, (\tilde{\mathbf{x}} - \mathbf{x}) \end{aligned}$$

and we similarly compute the gradient of $\mathrm{J}_2(\mathbf{x})$, which contains the new observation and partials:

$$\begin{aligned} \nabla\, \mathrm{J}_2(\mathbf{x}) &= \nabla \left(z^2 - 2\, z\, \mathbf{A}\, \mathbf{x} + (\mathbf{A}\, \mathbf{x})^2 \right) \\ &= 2\, \mathbf{A}^\intercal\, (\mathbf{A}\, \mathbf{x} - z) \\ &= 2\, \left(\mathbf{A}^2\, \mathbf{x} - \mathbf{A}^\intercal\, z \right) \end{aligned}$$

We can solve the resulting equation on sight, writing the new estimate with an overhat. We skip many intermediate steps that become obvious if you reproduce the derivation by hand. Be aware that $\mathbf{A}^2$ is an outer product, thus a matrix, in the common case of scalar observations, where $b = 1$ and $\mathbf{A}$ is a row.

$$\begin{aligned} \nabla\, \mathrm{J}_1(\mathbf{x}) &= \nabla\, \tilde{\mathrm{J}}(\mathbf{x}) + \nabla\, \mathrm{J}_2(\mathbf{x}) = 0 \\ &= \tilde{\mathbf{A}}^2\, \mathbf{x} - \tilde{\mathbf{A}}^2\, \tilde{\mathbf{x}} + \mathbf{A}^2\, \mathbf{x} - \mathbf{A}^\intercal\, z \\ &\Leftrightarrow x = \hat{x} \stackrel{\text{def}}{=} \left(\tilde{\mathbf{A}}^2 + \mathbf{A}^2 \right)^{-1} \cdot \left(\mathbf{A}^\intercal\, z + \tilde{\mathbf{A}}^2\, \tilde{\mathbf{x}} \right) \end{aligned}$$

Look how pretty this is. Equation 12 for the priors gave us the form $\tilde{\mathbf{x}} = \mathbf{Z}^{-1}\, \tilde{\mathbf{P}}\, \tilde{\mathbf{A}}^\intercal\, z$, a scaled covariance times a transform of the observations by the partials, transposed. The new estimate has exactly the same form if we regard the first matrix factor $\left(\tilde{\mathbf{A}}^2 + \mathbf{A}^2 \right)^{-1}$ as $\mathbf{Z}^{-1}$ times a covariance and if we regard *all* the priors $\tilde{\mathbf{A}}\, \tilde{\mathbf{x}}$ as a *single* additional observation to add to the current z. This is really close to the recurrent form we want. We get there by some rewrites. First, define the new covariance as the inverse of the sum of the old inverse covariance $\tilde{\mathbf{P}}^{-1} = \mathbf{Z}^{-1}\, \tilde{\mathbf{A}}^2$ and the new inverse covariance $\mathbf{Z}^{-1}\, \mathbf{A}^2$:

$$\hat{\mathbf{P}} \stackrel{\text{def}}{=} \mathbf{Z}\, \left(\tilde{\mathbf{A}}^2 + \mathbf{A}^2 \right)^{-1} \tag{15}$$

We can write this as a reciprocal because $\mathbf{Z}$ is diagonal, and see that it looks just like the classic 'sum of resistors' formula:

$$\frac{1}{\hat{\mathbf{P}}} = \frac{\tilde{\mathbf{A}}^2}{\mathbf{Z}} + \frac{\mathbf{A}^2}{\mathbf{Z}} = \frac{1}{\tilde{\mathbf{P}}} + \frac{\mathbf{A}^2}{\mathbf{Z}}$$

or

$$\frac{1}{\hat{\mathbf{P}}} - \frac{\mathbf{A}^2}{\mathbf{Z}} = \frac{1}{\tilde{\mathbf{P}}}$$

but, defining

$$\mathbf{K} \stackrel{\text{def}}{=} \mathbf{Z}^{-1}\,\hat{\mathbf{P}}\,\mathbf{A}^\intercal \tag{16}$$

we have

$$\mathbf{Z}\,\mathbf{K} = \hat{\mathbf{P}}\,\mathbf{A}^\intercal$$

so

$$\mathbf{Z}\,\mathbf{K}\,\mathbf{A} = \hat{\mathbf{P}}\,\mathbf{A}^2$$
$$\hat{\mathbf{P}}^{-1}\,\mathbf{K}\,\mathbf{A} = \frac{\mathbf{A}^2}{\mathbf{Z}}$$

Therefore

$$\hat{\mathbf{P}}^{-1}\,(1 - \mathbf{K}\,\mathbf{A}) = \tilde{\mathbf{P}}^{-1}$$
$$\hat{\mathbf{P}} = \mathbf{L}\,\tilde{\mathbf{P}} \tag{17}$$

where

$$\mathbf{L} \stackrel{\text{def}}{=} 1 - \mathbf{K}\,\mathbf{A} \tag{18}$$

We have one of our three equivalent expressions for the posterior covariance, which we can write as a recurrence:

$$\mathbf{P} \leftarrow \mathbf{L}\,\mathbf{P} \tag{19}$$

Note the following identity for the future:

$$\hat{\mathbf{P}}\,\tilde{\mathbf{A}}^2 + \hat{\mathbf{P}}\,\mathbf{A}^2 = \mathbf{Z} \tag{20}$$

Now rewrite $\hat{\mathbf{x}}$, noting that equation 20 implies that $\mathbf{Z}\,\mathbf{L} = \mathbf{Z}\,(1 - \mathbf{K}\,\mathbf{A}) = (\mathbf{Z} - \hat{\mathbf{P}}\,\mathbf{A}^2) = \hat{\mathbf{P}}\,\tilde{\mathbf{A}}^2$.

$$\begin{aligned}
\hat{\mathbf{x}} &= \left(\tilde{\mathbf{A}}^2 + \mathbf{A}^2\right)^{-1} \cdot \left(\mathbf{A}^\intercal\, z + \tilde{\mathbf{A}}^2\,\tilde{\mathbf{x}}\right) \\
&= \mathbf{Z}^{-1}\left(\hat{\mathbf{P}}\,\mathbf{A}^\intercal\, z + \hat{\mathbf{P}}\,\tilde{\mathbf{A}}^2\,\tilde{\mathbf{x}}\right) \\
&= \mathbf{K}\,z + (1 - \mathbf{K}\,\mathbf{A})\,\tilde{\mathbf{x}} \\
\therefore \hat{\mathbf{x}} &= \tilde{\mathbf{x}} + \mathbf{K}\,(z - \mathbf{A}\,\tilde{\mathbf{x}})
\end{aligned}$$

We have the update recurrence for the vector estimate $\mathbf{x}$. There remain two more covariance formulas to derive, namely

$$\mathbf{P} \leftarrow \mathbf{L}\,\mathbf{P}\,\mathbf{L}^\intercal + \mathbf{K}\,\mathbf{Z}\,\mathbf{K}^\intercal \tag{21}$$

and the canonical form,

$$\mathbf{P} \leftarrow \mathbf{P} - \mathbf{K}\,\mathbf{D}\,\mathbf{K}^\intercal \tag{22}$$

3.4.1 Minimizing $J_1(\mathbf{x})$

The new covariance is defined as

$$\hat{\mathbf{P}} = \mathrm{E}\left[(\hat{\mathbf{x}} - \mathbf{x})(\hat{\mathbf{x}} - \mathbf{x})^\intercal\right] \tag{23}$$

Get a new expression for $\hat{\mathbf{x}}$:

$$\hat{\mathbf{x}} = \tilde{\mathbf{x}} + \mathbf{K}\,(\mathbf{z} - \mathbf{A}\,\tilde{\mathbf{x}}) = \mathbf{K}\,\mathbf{z} + \mathbf{L}\,\tilde{\mathbf{x}} \tag{24}$$

where, again

$$\mathbf{L} = (\mathbf{1} - \mathbf{K}\,\mathbf{A}) = \mathbf{Z}^{-1}\hat{\mathbf{P}}\,\tilde{\mathbf{A}}^2 \tag{25}$$

Remembering the observation equation (6), write a single instance of it $\mathbf{z} = \mathbf{A}\,\mathbf{x} + \boldsymbol{\zeta}$ and find

$$\begin{aligned}
\hat{\mathbf{x}} &= \mathbf{K}\,\mathbf{A}\,\mathbf{x} + \mathbf{K}\,\boldsymbol{\zeta} + \mathbf{L}\,\tilde{\mathbf{x}} \\
&= (\mathbf{1} - \mathbf{L})\;\mathbf{x} + \mathbf{K}\,\boldsymbol{\zeta} + \mathbf{L}\,\tilde{\mathbf{x}} \\
&\Rightarrow (\hat{\mathbf{x}} - \mathbf{x}) = \mathbf{L}\;(\tilde{\mathbf{x}} - \mathbf{x}) + \mathbf{K}\,\boldsymbol{\zeta}
\end{aligned} \tag{26}$$

Remembering that $\mathrm{E}[\boldsymbol{\zeta}] = \mathbf{0}$, $\mathrm{E}[\boldsymbol{\zeta}\,\boldsymbol{\zeta}^\intercal] = \mathbf{Z}$ and skipping intermediate steps, find that

$$\hat{\mathbf{P}} = \mathbf{L}\,\tilde{\mathbf{P}}\,\mathbf{L}^\intercal + \mathbf{K}\,\mathbf{Z}\,\mathbf{K}^\intercal \tag{27}$$

We leave it to the reader to check, with reference to equations 4, that the physical dimensions work out. This completes the derivation of the recurrence equation 21.

To get the last form, we need a couple of small lemmas:

3.4.2 Lemma: $\mathbf{K}\,\mathbf{A}\,\tilde{\mathbf{P}}\,\mathbf{A}^\intercal = \tilde{\mathbf{P}}\,\mathbf{A}^\intercal\,\mathbf{A}\,\mathbf{K}$

You can call this "lemma kapat patak" if you like.

$$\begin{aligned}
\hat{\mathbf{P}}\,\mathbf{A}^\intercal\,\mathbf{A}\,\tilde{\mathbf{P}} &= \tilde{\mathbf{P}}\,\mathbf{A}^\intercal\,\mathbf{A}\,\hat{\mathbf{P}} && \text{Symmetric matrices commute} \\
\hat{\mathbf{P}}\,\mathbf{A}^\intercal\,\mathbf{A}\,\tilde{\mathbf{P}}\,\mathbf{A}^\intercal &= \tilde{\mathbf{P}}\,\mathbf{A}^\intercal\,\mathbf{A}\,\hat{\mathbf{P}}\,\mathbf{A}^\intercal && \text{Multiply on right by } \mathbf{A}^\intercal \\
\therefore \mathbf{K}\,\mathbf{A}\,\tilde{\mathbf{P}}\,\mathbf{A}^\intercal &= \tilde{\mathbf{P}}\,\mathbf{A}^\intercal\,\mathbf{A}\,\mathbf{K} && \text{Subst def of } \mathbf{K} = \hat{\mathbf{P}}\,\mathbf{A}^\intercal
\end{aligned} \tag{28}$$

3.4.3 Lemma: $\mathbf{K}\,\mathbf{D} = \tilde{\mathbf{P}}\,\mathbf{A}^\intercal$

You can call this "lemma kay-dee pat" if you like. It is equivalent to the main form for $\mathbf{K}$ used in 2. Assuming

$$\mathbf{D} \stackrel{\text{def}}{=} \mathbf{Z} + \mathbf{A}\,\tilde{\mathbf{P}}\,\mathbf{A}^\intercal \tag{29}$$

we get

$$\begin{aligned}
\mathbf{K}\,\mathbf{Z} &= \hat{\mathbf{P}}\,\mathbf{A}^\intercal && \text{Definition of } \mathbf{K}\text{, equation 16} \\
\mathbf{K}\,\mathbf{Z} &= \mathbf{Z}\,\left(\tilde{\mathbf{A}}^2 + \mathbf{A}^2\right)^{-1}\mathbf{A}^\intercal && \text{Definition of } \hat{\mathbf{P}}\text{, equation 15} \\
\left(\tilde{\mathbf{A}}^2 + \mathbf{A}^2\right)\mathbf{K} &= \mathbf{A}^\intercal && \text{Cancellation of } \mathbf{Z} \\
\tilde{\mathbf{A}}^2\,\mathbf{K} + \mathbf{A}^2\,\mathbf{K} &= \mathbf{A}^\intercal && \text{Distributive law} \\
\mathbf{K}\,\mathbf{Z} + \tilde{\mathbf{P}}\,\mathbf{A}^2\,\mathbf{K} &= \tilde{\mathbf{P}}\,\mathbf{A}^\intercal && \text{Left multiply by } \tilde{\mathbf{P}} \stackrel{\text{def}}{=} \mathbf{Z}\,\tilde{\mathbf{A}}^{-2} \\
\mathbf{K}\,\mathbf{Z} + \tilde{\mathbf{P}}\,\mathbf{A}^\intercal\,\mathbf{A}\,\mathbf{K} &= \tilde{\mathbf{P}}\,\mathbf{A}^\intercal && \text{expand } \mathbf{A}^2 \\
\mathbf{K}\,\mathbf{Z} + \mathbf{K}\,\mathbf{A}\,\tilde{\mathbf{P}}\,\mathbf{A}^\intercal &= \tilde{\mathbf{P}}\,\mathbf{A}^\intercal && \text{Equation 28} \\
\mathbf{K}\left(\mathbf{Z} + \mathbf{A}\,\tilde{\mathbf{P}}\,\mathbf{A}^\intercal\right) &= \tilde{\mathbf{P}}\,\mathbf{A}^\intercal \\
\therefore\ \mathbf{K}\,\mathbf{D} &= \tilde{\mathbf{P}}\,\mathbf{A}^\intercal && \text{Definition of } \mathbf{D}\text{, equation 29}
\end{aligned} \tag{30}$$

where we have freely used the fact that the diagonal matrix $\mathbf{Z}$ commutes with all other matrix products. This also demonstrates our original definition of the Kalman gain, $\mathbf{K} = \hat{\mathbf{P}}\,\mathbf{A}^\intercal\,\mathbf{D}^{-1}$ from equation 2.

We now show that $\mathbf{K}\,\mathbf{D} = \tilde{\mathbf{P}}\,\mathbf{A}^\intercal$ implies $\hat{\mathbf{P}} = \tilde{\mathbf{P}} - \mathbf{K}\,\mathbf{D}\,\mathbf{K}^\intercal$.

$$\begin{aligned}
\mathbf{K}\,\mathbf{D} &= \tilde{\mathbf{P}}\,\mathbf{A}^\intercal \\
\mathbf{K}\,\left(\mathbf{Z} + \mathbf{A}\,\tilde{\mathbf{P}}\,\mathbf{A}^\intercal\right) &= \tilde{\mathbf{P}}\,\mathbf{A}^\intercal && \text{Definition of } \mathbf{D}\text{, equation 29} \\
\mathbf{K}\,\mathbf{Z} + \mathbf{K}\,\mathbf{A}\,\tilde{\mathbf{P}}\,\mathbf{A}^\intercal &= \tilde{\mathbf{P}}\,\mathbf{A}^\intercal && \text{Expand} \\
\hat{\mathbf{P}}\,\mathbf{A}^\intercal + \mathbf{K}\,\mathbf{A}\,\tilde{\mathbf{P}}\,\mathbf{A}^\intercal &= \tilde{\mathbf{P}}\,\mathbf{A}^\intercal && \text{Definition of } \mathbf{K}\text{, equation 16} \\
\hat{\mathbf{P}}\,\mathbf{A}^\intercal - \tilde{\mathbf{P}}\,\mathbf{A}^\intercal &= -\mathbf{K}\,\mathbf{A}\,\tilde{\mathbf{P}}\,\mathbf{A}^\intercal && \text{Rearrange} \\
(\hat{\mathbf{P}} - \tilde{\mathbf{P}})\,\mathbf{A}^\intercal &= -\mathbf{K}\,\mathbf{A}\,\tilde{\mathbf{P}}\,\mathbf{A}^\intercal && \text{Collect} \\
(\hat{\mathbf{P}} - \tilde{\mathbf{P}})\,\mathbf{A}^\intercal &= -\mathbf{K}\,(\mathbf{K}\,\mathbf{D})^\intercal\,\mathbf{A}^\intercal && \text{Hypothesis and symmetry of } \tilde{\mathbf{P}} \\
\therefore\ (\hat{\mathbf{P}} - \tilde{\mathbf{P}})\,\mathbf{A}^\intercal &= -(\mathbf{K}\,\mathbf{D}\,\mathbf{K}^\intercal)\,\mathbf{A}^\intercal && \text{Symmetry of } \mathbf{D}
\end{aligned} \tag{31}$$

For arbitrary $\mathbf{A}^\intercal$, this can only be true if $\hat{\mathbf{P}} = \tilde{\mathbf{P}} - \mathbf{K}\,\mathbf{D}\,\mathbf{K}^\intercal$.

4 Concluding Remarks

These derivations are helpful for gaining intuition into the underlying statistics and dimensional structures of the Kalman filter and its many variants. They are a bit involved, but it is worthwhile

to ingest these fundamentals, especially for those who need to research new filters and applications. For more rigorous proofs built on a Bayesian perspective, see Bar-Shalom.[3] Emacs 24.5.1 (Org mode 8.3.4)