Kalman Folding 3: Derivations (Review 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.

[1] B. Beckman, *Kalman Folding, Part 1*, `http://vixra.org/abs/1606.0328`.

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

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

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:

- z is a $\mathfrak{b}\times 1$ column vector containing one multidimensional observation
- $\mathbf{x}$ 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 $\mathbf{x}$
- $\mathbf{A}$ 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 $\mathbf{x}$ are $[[\mathbf{x}]]\stackrel{\text{def}}{=}(\mathcal{X},\mathfrak{n}\times 1)$ and of z are $[[z]]\stackrel{\text{def}}{=}(\mathcal{Z},\mathfrak{b}\times 1)$, then

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

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

3 Derivations

Below, we derive equations 1, 2 and 3. Again, these derivations are just sketches designed for clarity as opposed to rigorous proofs. These derivations only cover the static Kalman filter, where $\mathbf{x}$ are fixed, constant, static states of the model. See Bar-Shalom[2] for derivations of the Kalman filter with time-dependent states and part 2 of this series[3] for an example.

The plan is first to develop expressions for the prior estimate $\tilde{\mathbf{x}}$ and prior covariance $\tilde{\mathbf{P}}$, and then expressions for the posterior versions $\hat{\mathbf{x}}$ and $\hat{\mathbf{P}}$, defining the Kalman gain $\mathbf{K}$ matrix and the denominator matrix $\mathbf{D}$ along the way. Finally, we derive the particular, convenient expressions for $\mathbf{K}$ and $\mathbf{D}$ that appear in equations 1, 2, and 3. Bierman laid out this strategy for the derivation in his classic book *Factorization Methods for Discrete Sequential Estimation*.[4] We follow his plan, unpacking many of his elided steps for greater clarity.

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 for cases where 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. When we write a product like $\mathbf{A}\,\mathbf{B}$, we assume that the number of columns of $\mathbf{A}$ matches the number of rows of $\mathbf{B}$.

Matrix multiplication is non-commutative, meaning that $\mathbf{A}\,\mathbf{B}$ does not, in general, equal $\mathbf{B}\,\mathbf{A}$. However, if a matrix $\mathbf{D}$ is diagonal, meaning that it has non-zero entries only along its main diagonal from upper left to lower right, then $\mathbf{A}\,\mathbf{D}$ does always equal $\mathbf{D}\,\mathbf{A}$. We may freely use this fact without mentioning it explicitly.

Symmetric matrices do not always commute, even with other symmetric matrices. In particular, the product of two symmetric matrices is not always symmetric, as witnessed by the following counterexample:

$$\begin{pmatrix} 1 & 2 \\ 2 & 3 \end{pmatrix} \cdot \begin{pmatrix} 4 & 5 \\ 5 & 6 \end{pmatrix} = \begin{pmatrix} 14 & 17 \\ 23 & 28 \end{pmatrix}$$

Matrix multiplication is associative, meaning that the order in which pairwise multiplications is carried out does not matter. Thus

$$(\mathbf{A}\,\mathbf{B})\,\mathbf{C} = \mathbf{A}\,(\mathbf{B}\,\mathbf{C}) = \mathbf{A}\,\mathbf{B}\,\mathbf{C}$$

and we don't need to write parentheses. That means some expressions of long products can be hard to read. We occasionally use a center dot or $\times$ symbol to make multiplication easier to read,

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

[3] B. Beckman, *Kalman Folding 2: Tracking and System Dynamics*, `http://vixra.org/abs/1606.0348`.

[4] `http://tinyurl.com/h3jh4kt`

as in $\mathbf{A} \cdot \mathbf{x}$ or $\mathbf{A} \times \mathbf{x}$. We also use the $\times$ symbol when discussing the numbers of rows and columns of a matrix, as in "$\mathbf{A}$ is an $m \times n$ matrix," meaning that $\mathbf{A}$ has m rows and n columns.

We may freely exploit the following facts without mentioning them explicitly:

- 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 square of 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 $m \times m$.

3.1.1 Probability and Statistics

We use the terms *distribution* and *expectation value* without definition in this paper. If $\mathbf{x}$ is a random variable, then we denote the expectation value of some function f of $\mathbf{x}$ as $\mathsf{E}[f(\mathbf{x})]$.

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.

$\mathbf{x}$ 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.

$\mathbf{A}\,\mathbf{x}$ is 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 may depend on t. The model is a column vector of dimensionality $b \times 1$, the same as 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 $\mathbf{A}(t)$; it's an aesthetic judgment to omit explicit t dependence because it would make the derivations 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 may depend 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 $\mathfrak{m}\mathfrak{b} \times \mathfrak{n}$, where $\mathfrak{m}$ is the number of prior observations, $\mathfrak{b}$ is the dimensionality of a single observation $\boldsymbol{z}$, and $\mathfrak{n}$ is the dimensionality of the states $\boldsymbol{x}$. Thus $\tilde{\mathbf{A}}$ is tall in the typical *overdetermined* case where $\mathfrak{m} > \mathfrak{n}$, 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 for the derivations below.

$\boldsymbol{z}$ is the *current observation*. It is known and non-random. Its dimensionality is $\mathfrak{b} \times 1$.

$\tilde{\boldsymbol{z}}$ is a stack of all prior observations. It is known, non-random, $\mathfrak{m}\mathfrak{b} \times 1$. 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.

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

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

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

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

$\tilde{\mathbf{P}}$ *covariance of the priors*, equals $\tilde{\mathbf{A}}^{-2}$ (de-dimensionalized; proof sketch below). This is called just $\mathbf{P}$ in part one of this series. It is known, non-random, $\mathfrak{n} \times \mathfrak{n}$.

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

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

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

ζ *observation noise*: random column-vector with zero mean and covariance $\mathbf{Z}$ (unity, $\mathbf{1}$, after de-dimensionalization). It has $\mathfrak{b}$ rows and 1 column, like $\boldsymbol{z}$.

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

$\tilde{\boldsymbol{z}} = \tilde{\mathbf{A}}\boldsymbol{x} + \zeta$ the *observation equation*, which equates $\tilde{\boldsymbol{z}}$, the stack of all prior observations, to the product of $\tilde{\mathbf{A}}$, the stack of all prior partials matrices, and an unknown random vector of states, $\boldsymbol{x}$, plus some unknown random observation noise ζ. The stack of prior observations $\tilde{\boldsymbol{z}}$ is known, non-random, $\mathfrak{m}\mathfrak{b} \times 1$; the stack of prior partials matrices $\tilde{\mathbf{A}}$ is known, non-random, $\mathfrak{m}\mathfrak{b} \times \mathfrak{n}$; the state vector $\boldsymbol{x}$ is unknown, random, $\mathfrak{n} \times 1$; The noise vector ζ is unknown, random, $\mathfrak{m}\mathfrak{b} \times 1$. The observation equation looks similar to the expression for the residual above. It's worthwhile to take a little time to examine the notations carefully and make sure that you have a good mental picture of the meanings of these notations. The observation equation

looks tall in the typical, overdetermined case, where as the residual is usually equivalent to a scalar expression.

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

D *Kalman denominator* $= \mathbf{Z} + \mathbf{A}\,\tilde{\mathbf{P}}\,\mathbf{A}^{\intercal}$, or $\mathbf{1} + \mathbf{A}\,\tilde{\mathbf{P}}\,\mathbf{A}^{\intercal}$ de-dimensionalized. (proof sketch below). Non-random, $b \times b$.

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

The fact that the prior covariance, $\tilde{\mathbf{P}}$, equals the the inverse square of the stack of prior partials matrices (de-dimensionalized), $\tilde{\mathbf{A}}^{-2}$, is the secret to Kalman's efficient, in fact constant, use of computer memory. The stack of prior partials matrices $\tilde{\mathbf{A}}$ can be very tall and impractical to store. But its square, $\tilde{\mathbf{A}}^2$ is only $n \times n$, and its inverse square is also just $n \times n$. Kalman packs all statistical information about the model into this small matrix of constant size, and incrementally improves the statistics as observations accumulate, without increasing the size of the matrix, and thus without increasing the amount of computer memory needed to keep all important information. The Kalman filter is *optimal*, meaning that the small covariance matrices keep all available information. No other method would be able to squeeze more information out of the observations and the model — at least when the noise is Gaussian. A rigorous optimality proof is out of scope for this paper, but the least-squares derivation below contains the central idea: Kalman tracks the estimate and covariance that minimize the sum of squared residuals. Kalman is optimal in the sense that no other method would find a smaller sum of squared residuals.

3.3.1 Covariance of Any Random Vector Variable

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

3.3.2 Prior Estimate $\tilde{\mathbf{x}}$

One of our random variables is $\mathbf{x}$, the column n-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 $\tilde{\boldsymbol{z}}$ (tall, $mb \times 1$).

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

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

3.3.3 Sum of Squared Residuals

Consider the following *performance functional*, computed over the population of $\mathbf{x}$.

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

$J(\mathbf{x})$ is a scalar: the sum of squared residuals. A *residual* is a difference between an actual observation $\boldsymbol{z}$ and a predicted observation $\boldsymbol{A}\,\boldsymbol{x}$. An *actual observation* $\boldsymbol{z}$ is a known, concrete b-vector of

numbers, and the partials matrix $\mathbf{A}$ is a known, concrete $(\mathfrak{b} \times \mathfrak{n})$-matrix of numbers corresponding to that observation. The observation equation

- stacks all prior observations (known, concrete numbers) into $\tilde{\mathbf{z}}$
- stacks all prior values of the partials matrix $\mathbf{A}$ into $\tilde{\mathbf{A}}$ (known, concrete numbers)
- multiplies by the unknown random state estimate $\mathbf{x}$ to get the (unknown, random) predicted observations $\tilde{\mathbf{A}}\,\mathbf{x}$
- finally adds some unknown random noise ζ (column vector of height $\mathfrak{m}\mathfrak{b}$)

The performance functional collapses all that information into a scalar random variable $J(\mathbf{x})$ with the same (Gaussian) distribution as the noise ζ. Recall that any *random variable* is, in fact, always a function, even if only the identity function, as when we say that $\mathbf{x}$ is a random variable. This is the standard nomenclature of probability and statistics established by Kolmogorov, and it admittedly can be confusing.

The job of finding the optimal estimate of the state vector $\mathbf{x}$ is the job of finding the concrete, numerical value of $\mathbf{x}$ that minimizes the performance functional $J(\mathbf{x})$, which depends on all the known, non-random, concrete numbers in $\tilde{\mathbf{z}}$ and $\tilde{\mathbf{A}}$.

To find the $\mathbf{x}$ that minimizes $J(\mathbf{x})$, we could take the classic, school approach of setting to zero the partial derivatives of $J(\mathbf{x})$ with respect to $\mathbf{x}$ and solving the resulting equations for $\mathbf{x}$. The following is an easier way. Multiply the residuals across by the wide matrix $\tilde{\mathbf{A}}^{\intercal}$:

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

producing an $\mathfrak{n}$-vector, and then construct a modified performance functional:

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

$J(\mathbf{x})$ is minimum with respect to $\mathbf{x}$ if and only if (iff) $J'(\mathbf{x})$ is minimum (this assertion needs a rigorous proof; as warned, we present only sketches in this paper). 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 $\mathfrak{n} \times \mathfrak{n}$ square matrix, is invertible (non-singular), in which case

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

produces that minimum value of $J'(\mathbf{x})$, because then

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

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 $\tilde{\mathbf{x}}$

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

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

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

Equation 8 is another assertion that requires a rigorous proof, out of scope for this paper of sketches. But, assuming it is true, we have

$$
\begin{aligned}
J(\mathbf{x}) &= \left(\tilde{\mathbf{z}} - \tilde{\mathbf{A}}\,\mathbf{x}\right)^\intercal \cdot \left(\tilde{\mathbf{z}} - \tilde{\mathbf{A}}\,\mathbf{x}\right) \\
&= \left(\tilde{\mathbf{z}} - \tilde{\mathbf{A}}\,\mathbf{x}\right)^\intercal \tilde{\mathbf{A}}\,\tilde{\mathbf{A}}^{-2}\,\tilde{\mathbf{A}}^\intercal \left(\tilde{\mathbf{z}} - \tilde{\mathbf{A}}\,\mathbf{x}\right) && \text{insert } \mathbf{1} \text{ from equation 8} \\
&= \left(\tilde{\mathbf{z}} - \tilde{\mathbf{A}}\,\mathbf{x}\right)^\intercal \tilde{\mathbf{A}}\,(\tilde{\mathbf{A}}^{-2}\,\tilde{\mathbf{A}}^2)\,\tilde{\mathbf{A}}^{-2}\,\tilde{\mathbf{A}}^\intercal \left(\tilde{\mathbf{z}} - \tilde{\mathbf{A}}\,\mathbf{x}\right) && \text{insert } \mathbf{1} = \tilde{\mathbf{A}}^{-2}\,\tilde{\mathbf{A}}^2 \\
&= \left[\left(\tilde{\mathbf{z}} - \tilde{\mathbf{A}}\,\mathbf{x}\right)^\intercal \tilde{\mathbf{A}}\,\tilde{\mathbf{A}}^{-2}\right] \tilde{\mathbf{A}}^2 \left[\tilde{\mathbf{A}}^{-2}\,\tilde{\mathbf{A}}^\intercal \left(\tilde{\mathbf{z}} - \tilde{\mathbf{A}}\,\mathbf{x}\right)\right] && \text{Regroup} \\
&= \left[\tilde{\mathbf{A}}^{-2}\,\tilde{\mathbf{A}}^\intercal \left(\tilde{\mathbf{z}} - \tilde{\mathbf{A}}\,\mathbf{x}\right)\right]^\intercal \tilde{\mathbf{A}}^2 \left[\tilde{\mathbf{A}}^{-2}\,\tilde{\mathbf{A}}^\intercal \left(\tilde{\mathbf{z}} - \tilde{\mathbf{A}}\,\mathbf{x}\right)\right] && \text{Symmetry of } \tilde{\mathbf{A}} \text{ and } \tilde{\mathbf{A}}^{-2} \\
&= (\tilde{\mathbf{x}} - \mathbf{x})^\intercal\,\tilde{\mathbf{A}}^2\,(\tilde{\mathbf{x}} - \mathbf{x}) && \text{Definition of } \tilde{\mathbf{x}} \text{ from equation 7}
\end{aligned} \tag{9}
$$

This has physical dimensions $\mathcal{Z}^2$ where $\mathcal{Z}$ are the physical dimensions of the observations $\mathbf{z}$.

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

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

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

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

$$
\begin{aligned}
\tilde{\mathbf{z}} &= \tilde{\mathbf{A}}\,\mathbf{x} + \zeta && \text{the observation equation, Equation 6} \\
\tilde{\mathbf{A}}^{-2}\,\tilde{\mathbf{A}}^\intercal\,\tilde{\mathbf{z}} &= \mathbf{x} + \tilde{\mathbf{A}}^{-2}\,\tilde{\mathbf{A}}^\intercal\,\zeta && \text{Multiply on left by } \tilde{\mathbf{A}}^{-2}\,\tilde{\mathbf{A}}^\intercal \\
\tilde{\mathbf{x}} &= \mathbf{x} + \tilde{\mathbf{A}}^{-2}\,\tilde{\mathbf{A}}^\intercal\,\zeta && \text{Definition of } \tilde{\mathbf{x}} \text{ from equation 7} \\
\text{therefore}\quad \tilde{\mathbf{x}} - \mathbf{x} &= \tilde{\mathbf{A}}^{-2}\tilde{\mathbf{A}}^\intercal\zeta
\end{aligned}
$$

Now rewrite equation 10, the definition of the prior covariance $\tilde{\mathbf{P}}$:

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

We can collapse the expectation value inwards because the stack of observation partials $\tilde{\mathbf{A}}$ is a matrix of concrete, non-random numbers.

Noise ζ is Gaussian, normal, with diagonal covariance matrix $\mathbf{Z}$, by hypothesis. Equation 11 becomes

$$\begin{aligned}\tilde{\mathbf{P}} = \tilde{\mathbf{A}}^{-2}\tilde{\mathbf{A}}^{\intercal}\,\mathrm{E}\left[\zeta\,\zeta^{\intercal}\right](\tilde{\mathbf{A}}^{-2}\tilde{\mathbf{A}}^{\intercal})^{\intercal} &= \tilde{\mathbf{A}}^{-2}\tilde{\mathbf{A}}^{\intercal}\,\mathbf{Z}\,(\tilde{\mathbf{A}}^{-2}\tilde{\mathbf{A}}^{\intercal})^{\intercal} \\ &= \tilde{\mathbf{A}}^{-2}\tilde{\mathbf{A}}^{\intercal}\,\mathbf{Z}\,\tilde{\mathbf{A}}(\tilde{\mathbf{A}}^{-2})^{\intercal} \\ &= \tilde{\mathbf{A}}^{-2}\tilde{\mathbf{A}}^{\intercal}\,\mathbf{Z}\,\tilde{\mathbf{A}}(\tilde{\mathbf{A}}^{-2}) \end{aligned} \tag{12}$$

because $\tilde{\mathbf{A}}^{-2}$ is symmetric. At this point, no further simplification is possible, in general, because $\mathbf{Z}$ is $b \times b$ and can only be sandwiched between $\tilde{\mathbf{A}}^{\intercal}$, $n \times b$, and $\tilde{\mathbf{A}}$, $b \times n$. However, we can greatly simplify this and all subsequent computations by de-dimensionalizing. There are numerical benefits, as well, to be discussed in the next section.

3.3.5 De-Dimensionalizing the Observation Equation

Fully spelled out, and in the general case of b-vector observations z, one block of height b of the observation equation is

$$\begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_b \end{pmatrix} = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{b1} & A_{b2} & \cdots & A_{bn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} \zeta_1 \\ \zeta_2 \\ \vdots \\ \zeta_b \end{pmatrix}$$

If we divide each row i by the standard deviation σ_{z_i} of the i-th component of the observation z, we get

$$\begin{pmatrix} \frac{z_1}{\sigma_{z_1}} \\ \frac{z_2}{\sigma_{z_2}} \\ \vdots \\ \frac{z_b}{\sigma_{z_b}} \end{pmatrix} = \begin{pmatrix} \frac{A_{11}}{\sigma_{z_1}} & \frac{A_{12}}{\sigma_{z_1}} & \cdots & \frac{A_{1n}}{\sigma_{z_1}} \\ \frac{A_{21}}{\sigma_{z_2}} & \frac{A_{22}}{\sigma_{z_2}} & \cdots & \frac{A_{2n}}{\sigma_{z_2}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{A_{b1}}{\sigma_{z_b}} & \frac{A_{b2}}{\sigma_{z_b}} & \cdots & \frac{A_{bn}}{\sigma_{z_b}} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} \frac{\zeta_1}{\sigma_{z_1}} \\ \frac{\zeta_2}{\sigma_{z_2}} \\ \vdots \\ \frac{\zeta_b}{\sigma_{z_b}} \end{pmatrix} \tag{13}$$

The covariance of the noise ζ, so normalized, is non-dimensional unity and equation 12 collapses completely to just

$$\tilde{\mathbf{P}} = \tilde{\mathbf{A}}^{-2} \tag{14}$$

and the estimate of the priors, equation 7 now becomes

$$\tilde{\mathbf{x}} \stackrel{\text{def}}{=} \tilde{\mathbf{P}}\,\tilde{\mathbf{A}}^{\intercal}\tilde{z} \tag{15}$$

This is remarkable. All information about the covariance of the noise is pulled into the (new, normalized) observation partials.

I remember, when working in the early 1980's at the Deep Space Network at JPL on direct measurement of tectonic drift,[5] one difficulty was the wide disparity in uncertainties of horizontal measurments (right ascension and declination) and range. For instance, we knew the RA-dec

[5]JPL Geodynamics Program `http://www.jpl.nasa.gov/report/1981.pdf`

position of the centroid of Saturn within 75 meters but its distance to no better than a million kilometers. That's a disparity of seven orders of magnitude (the situation is greatly improved, now, due to the accumulation of range data for multiple spacecraft coupled with decades of orbital mechanics[6]). At the time, this meant that we had to deal with error ellipsoids that were long, thin needles. That means covariance matrices with components differing by up to fourteen orders of magnitude, and that's not practical with floating-point computer arithmetic. One mitigation was de-dimensionalizing or normalizing, as described here, which brings the uncertainties of all components of an observation into the same numerical range, near unity. Another mitigation was Square Root Information Filtering (SRIF), the subject of another paper in this series.

In any event, for all subsequent calculations in this paper, we assume that the observation equation has been normalized and that $\mathbf{Z} = \mathbf{1}$.

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 $\mathbf{z}$, both of which are matrices of known, concrete, non-random numbers. This is exactly what our *kalmanStatic* function from equation 1 does, of course, in functional form. We derive the posteriors from scratch to seek opportunities to define $\mathbf{K}$ and $\mathbf{D}$ and to radically shorten the expressions.

First, define a new performance functional $J_1(\mathbf{x})$ as the sum of the performance of the priors $\tilde{J}(\mathbf{x})$ from equation 9, now written with tildes overhead, and a new term $J_2(\mathbf{x})$ for the performance of the new data:

$$J_1(\mathbf{x}) \stackrel{\text{def}}{=} \tilde{J}(\mathbf{x}) + J_2(\mathbf{x}) \tag{16}$$

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

$$\begin{aligned} J_2(\mathbf{x}) &\stackrel{\text{def}}{=} (\mathbf{z} - \mathbf{A}\mathbf{x})^2 \\ &= (\mathbf{z} - \mathbf{A}\mathbf{x})^\intercal \cdot (\mathbf{z} - \mathbf{A}\mathbf{x}) \\ &= \mathbf{z}^2 - \mathbf{z}^\intercal \mathbf{A}\mathbf{x} - \mathbf{z}\mathbf{x}^\intercal \mathbf{A}^\intercal + (\mathbf{A}\mathbf{x})^2 \\ &= \mathbf{z}^2 - 2\,\mathbf{z}^\intercal \mathbf{A}\mathbf{x} + (\mathbf{A}\mathbf{x})^2 \end{aligned} \tag{17}$$

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 $d\,J_1(\mathbf{x})/d\,\mathbf{x} = 0$. The usual way to write a vector derivative is with the *nabla* operator ∇, which produces *gradient* vectors from scalar functions.

$$\nabla f(\mathbf{x}) \stackrel{\text{def}}{=} \begin{bmatrix} df(\mathbf{x})/dx_0 \\ df(\mathbf{x})/dx_1 \\ \vdots \\ df(\mathbf{x})/dx_{n-1} \end{bmatrix}$$

The particular scalar function we're differentiating is, of course, the new performance functional $J_1(\mathbf{x}) = \tilde{J}(\mathbf{x}) + J_2(\mathbf{x})$. Because $\tilde{\mathbf{A}}^2$ is symmetric,

[6] http://ipnpr.jpl.nasa.gov/progress_report/42-178/178C.pdf

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

an n-vector, and we similarly compute the gradient of $J_2(x)$, which contains the new observation and partials:

$$\begin{aligned}\nabla J_2(x) &= \nabla \left(z^2 - 2 z^T A x + (A x)^2 \right) \\ &= 2 A^T (A x - z) \\ &= 2 \left(A^2 x - A^T z \right)\end{aligned}$$

another n-vector. We can solve the resulting equation for x on sight, writing the new solution — the new estimate — with an overhat. Be aware that that A is a wide matrix, in fact an n-row when $b = 1$, a common case, and A^2 is thus an outer product and an $n \times n$ matrix.

$$\begin{aligned}\nabla J_1(x) &= \nabla \tilde{J}(x) + \nabla J_2(x) = 0 \\ &= \tilde{A}^2 x - \tilde{A}^2 \tilde{x} + A^2 x - A^T z \\ &\text{if and only if} \quad x = \hat{x} \stackrel{\text{def}}{=} \left(\tilde{A}^2 + A^2 \right)^{-1} \cdot \left(A^T z + \tilde{A}^2 \tilde{x} \right)\end{aligned} \tag{18}$$

Look how pretty this is. Equation 15 for the priors gave us the form $\tilde{x} = \tilde{P} \tilde{A}^T \tilde{z}$, a covariance $\tilde{P}$ times the prior observations $\tilde{z}$ scaled by the prior partials, transposed, $\tilde{A}^T$. The new estimate $\hat{x}$ has exactly the same form if we regard the first matrix factor $\left(\tilde{A}^2 + A^2 \right)^{-1}$ as a covariance $\hat{P}$ and if we regard *all* the priors $\tilde{A}^2 \tilde{x}$ as a *single* scaled observation to add to the current scaled observation $A^T z$. We may regard $\tilde{A}^2 \tilde{x}$ as a scaled observation because equations 12 and 15 imply that $\tilde{A}^T \tilde{z} = \tilde{A}^2 \tilde{x}$. We may view the second term above, $A^T z + \tilde{A}^2 \tilde{x}$, as $A^T z + \tilde{A}^T \tilde{z}$.

3.4.1 Posterior estimate, $\hat{x}$

We must wrangle equation 5 from equation 18. Equation 5 is the recurrence we want, namely $\hat{x} = \tilde{x} + K(z - A\tilde{x})$, and equation 18 is the recurrence we have, namely
$\hat{x} = \left(\tilde{A}^2 + A^2 \right)^{-1} \left(A^T z + \tilde{A}^2 \tilde{x} \right)$.

First, formally define the new, posterior covariance.

$$\hat{P} \stackrel{\text{def}}{=} \left(\tilde{A}^2 + A^2 \right)^{-1} \tag{19}$$

Now write equation 18 as

$$\hat{x} = \hat{P} \left(A^T z + \tilde{A}^2 \tilde{x} \right)$$

$$= \hat{P} A^T z + \hat{P} \tilde{A}^2 \tilde{x}$$

The form above strongly suggests that we define

$$\mathbf{K} \stackrel{\text{def}}{=} \hat{\mathbf{P}}\,\mathbf{A}^{\intercal} \tag{20}$$

yielding

$$\hat{\mathbf{x}} = \mathbf{K}\,\boldsymbol{z} + \hat{\mathbf{P}}\,\tilde{\mathbf{A}}^{2}\,\tilde{\mathbf{x}} \tag{21}$$

Now, to get the recurrence we want

$$\begin{aligned} \hat{\mathbf{x}} &= \tilde{\mathbf{x}} + \mathbf{K}\,(\boldsymbol{z} - \mathbf{A}\,\tilde{\mathbf{x}}) \\ &= \tilde{\mathbf{x}} + \mathbf{K}\,\boldsymbol{z} - \mathbf{K}\,\mathbf{A}\,\tilde{\mathbf{x}} \end{aligned} \tag{22}$$

we need only set equation 21 equal to equation 22. Cancelling terms and rearranging, we get

$$(\mathbf{1} - \mathbf{K}\,\mathbf{A})\,\tilde{\mathbf{x}} = \hat{\mathbf{P}}\,\tilde{\mathbf{A}}^{2}\,\tilde{\mathbf{x}} = \hat{\mathbf{P}}\,\tilde{\mathbf{P}}^{-1}\,\tilde{\mathbf{x}} \tag{23}$$

by definition of the prior covariance, equation 14. For arbitrary $\tilde{\mathbf{x}}$, this will be true if

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

Rearrange and right-multiply by $\tilde{\mathbf{P}}$ to get

$$\hat{\mathbf{P}} = (\mathbf{1} - \mathbf{K}\,\mathbf{A})\;\tilde{\mathbf{P}} = \hat{\mathbf{P}}\,\tilde{\mathbf{A}}^{2}\,\tilde{\mathbf{P}} \tag{24}$$

showing that equations 23 and 5 are just alternative expressions for the same thing.

Let's write this more compactly

$$\hat{\mathbf{P}} = \mathbf{L}\,\tilde{\mathbf{P}} \tag{25}$$

where

$$\mathbf{L} \stackrel{\text{def}}{=} (\mathbf{1} - \mathbf{K}\,\mathbf{A}) = \hat{\mathbf{P}}\,\tilde{\mathbf{A}}^{2} \tag{26}$$

and we have one of the three equivalent recurrences for the posterior covariance from the first paper in this series

$$\mathbf{P} \leftarrow \mathbf{L}\,\mathbf{P} \tag{27}$$

3.4.2 A Gain Matrix K We Can Actually Compute

Of course, the gain matrix $\mathbf{K}$ is formally defined in terms of the posterior covariance, that is, as $\hat{\mathbf{P}}\mathbf{A}^\intercal$, but we don't have the posterior covariance $\hat{\mathbf{P}}$ by equation 24 until we have the gain matrix $\mathbf{K}$. To get out of this fix, we note that

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

and solve for $\mathbf{K}$:

$$\begin{aligned} \mathbf{K} &= \tilde{\mathbf{P}}\mathbf{A}^\intercal - \mathbf{K}\mathbf{A}\tilde{\mathbf{P}}\mathbf{A}^\intercal \\ \mathbf{K}(\mathbf{1} + \mathbf{A}\tilde{\mathbf{P}}\mathbf{A}^\intercal) &= \tilde{\mathbf{P}}\mathbf{A}^\intercal \\ \mathbf{K} &= \tilde{\mathbf{P}}\mathbf{A}^\intercal(\mathbf{1} + \mathbf{A}\tilde{\mathbf{P}}\mathbf{A}^\intercal)^{-1} \end{aligned} \tag{28}$$

Defining the Kalman denominator matrix $\mathbf{D}$ as follows:

$$\mathbf{D} \stackrel{\text{def}}{=} \mathbf{1} + \mathbf{A}\tilde{\mathbf{P}}\mathbf{A}^\intercal \tag{29}$$

we finally get a form for the Kalman gain matrix $\mathbf{K}$ entirely in terms of priors and the new observation partials (sometimes called the *innovation*):

$$\mathbf{K} = \tilde{\mathbf{P}}\mathbf{A}^\intercal\mathbf{D}^{-1} \tag{30}$$

$$\text{where } \mathbf{D} = \mathbf{1} + \mathbf{A}\mathbf{P}\mathbf{A}^\intercal \tag{31}$$

These are almost the same as the original definitions, equations 2 and 3, which were written in dimensional form. We leave it to the reader to show that the dimensional form for $\mathbf{D}$ is $\mathbf{Z} + \mathbf{A}\mathbf{P}\mathbf{A}^\intercal$.

3.4.3 Two More Recurrences

There remain two more recurrences to derive, namely

$$\mathbf{P} \leftarrow \mathbf{L}\mathbf{P}\mathbf{L}^\intercal + \mathbf{K}\mathbf{Z}\mathbf{K}^\intercal \tag{32}$$

and the canonical form,

$$\mathbf{P} \leftarrow \mathbf{P} - \mathbf{K}\mathbf{D}\mathbf{K}^\intercal \tag{33}$$

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

The posterior covariance is, from the statistical viewpoint,

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

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

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

where, again

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

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

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

implying that $(\hat{\mathbf{x}} - \mathbf{x}) = \mathbf{L}\;(\tilde{\mathbf{x}} - \mathbf{x}) + \mathbf{K}\,\zeta$.

Remembering that $\mathsf{E}\,[\zeta] = \mathbf{0}$, $\mathsf{E}\,[\zeta\,\zeta^\intercal] = \mathbf{Z}$, glibly re-dimensionalizing and skipping intermediate steps, we find that

$$\hat{\mathbf{P}} = \mathbf{L}\,\tilde{\mathbf{P}}\,\mathbf{L}^\intercal + \mathbf{K}\,\mathbf{Z}\,\mathbf{K}^\intercal \tag{34}$$

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 32.

The last form, $\hat{\mathbf{P}} = \tilde{\mathbf{P}} - \mathbf{K}\,\mathbf{D}\,\mathbf{K}^\intercal$, is easy to show from what we already know, that $\hat{\mathbf{P}} = \mathbf{L}\,\tilde{\mathbf{P}} = (\mathbf{1} - \mathbf{K}\,\mathbf{A})\,\tilde{\mathbf{P}}$. We just need to show that $\mathbf{K}\,\mathbf{A}\,\tilde{\mathbf{P}} = \mathbf{K}\,\mathbf{D}\,\mathbf{K}^\intercal$. Substitute $\mathbf{D}^{-\intercal}\,\mathbf{A}\,\tilde{\mathbf{P}}^\intercal$ for $\mathbf{K}^\intercal$ by transposing equation 30. Note that for square matrices, the inverse of the transpose is the transpose of the inverse. Therefore $\mathbf{D}^{-\intercal} = \mathbf{D}^{-1}$ because $\mathbf{D}$ is symmetric. Likewise $\tilde{\mathbf{P}}^\intercal = \tilde{\mathbf{P}}$. The result follows:

$$\mathbf{K}\,\mathbf{D}\,\mathbf{K}^\intercal = \mathbf{K}\,\mathbf{D}\,\mathbf{D}^{-\intercal}\,\mathbf{A}\tilde{\mathbf{P}} = \mathbf{K}\,\mathbf{A}\tilde{\mathbf{P}}$$

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.[2] For more careful dimensional analysis of the present derivations, see part 6 of this series.[7]

[7] B. Beckman, *Kalman Folding 6: Dimensional Analysis*, to appear.