Example of a Kalman filter on a straight line

Robert K. Kutschke
March 13, 1998

Abstract

This document provides a worked example of how a Kalman filter actually works. It illustrates, how the information content changes as hits are added. The effect of multiple scattering is illustrated in section 4. The example is a straight line in 2D.

1 The Kalman Filter Equations

In this section the Kalman filter equations will be stated without proof. A derivation of these equations can be found in many places. A lengthy discussion can be found in, for example, reference [1], while a shorter one can be found in reference [2].

First, it is necessary to define some notation. Let η describe the track parameters, for example the helical parameters, $(\kappa, \phi_0, d_0, \cot\theta, z_0)$, the forward spectrometer parameters, $(\alpha, x', y', x_0, y_0)$, or the straight line parameters (m, b). The discussion of the coordinate system in which these track parameters are defined will be discussed later. Let V describe the covariance matrix of the track parameters. For this example we will only consider 1D measuring devices; let d_m denote the measurement made by the device and let σ denote the error on d_m. Also let $d(\eta)$ denote the the measurement as predicted by the track parameters. Finally, let D denote the derivatives of the measurement with respect to the track parameters,

$$D_i = \frac{\partial d_m}{\partial \eta_i}. \tag{1}$$

The Kalman filter equations take one estimator of the track (η, V) and add a new hit to obtain an improved estimator, (η', V'):

$$\begin{aligned} \eta' &= \eta + V'D\frac{d_m - d(\eta)}{\sigma^2} \\ V' &= V - \frac{VDD^TV}{\sigma^2 + D^TVD}. \end{aligned} \tag{2}$$

No matrix inversion is needed in this calculation.

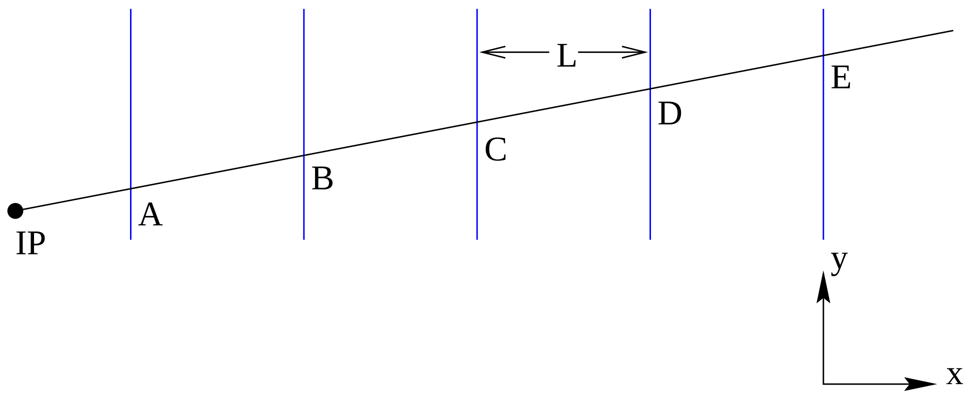

Figure 1: The diagonal line illustrates a track which is produced at the IP and which travels to the right, following a straight line. The vertical lines represent equally spaced measurement planes, which measure the y coordinate. The separation between measurement planes is L. The coordinate system used at the start of the fit is shown at the lower right.

2 Setting Up the Example

Consider a trajectory which is just a straight line in the (x, y) plane, as shown in figure 1. In this figure the track starts at the interaction point (IP) and it proceeds to the right. The trajectory intersects a series of planes which measure y at fixed values of x. The intersection points are labeled A, B, C, D and E. Each of the planes has a y-resolution of σ and we denote the values of the measurements at each point by $(y_A, y_B, y_C, y_D, y_E)$.

The goal of the Kalman filter is to obtain an estimate of the trajectory which is valid in the neighbourhood of the IP.

The equation of the trajectory is,

$$y = mx + b. \tag{3}$$

That is, the track parameters and covariance matrix have the form,

$$\eta = \begin{pmatrix} m \\ b \end{pmatrix} \qquad V = \begin{pmatrix} V_{mm} & V_{mb} \\ V_{mb} & V_{bb} \end{pmatrix}. \tag{4}$$

3 Fitting Without Multiple Scattering

3.1 Initialization

Many coordinate systems will be used in this example. The first of them is shown in figure 1; the origin of x is at the x of the measurement point farthest from the IP, point E. The coordinate axes (x, y) are aligned with the x and y axes of the measurement device. The origin of y is not important and we can choose it to be that shown in the figure.

In order to start the fit, the track parameters are initialized to some values which are obtained from the pattern recognition routines and the covariance matrix is initialized to a diagonal matrix with large numbers on the diagonal.

$$\eta = \begin{pmatrix} m_0 \\ b_0 \end{pmatrix} \qquad V = \begin{pmatrix} V_{0_{mm}} & 0 \\ 0 & V_{0_{bb}} \end{pmatrix}. \tag{5}$$

Appropriate values for the starting covariance matrix diagonal elements and constraints on the quality of the initial track parameters are discussed below.

3.2 Add the First Hit

In order to add the hit at point E to the track, we apply equation 2. Because the track parameters are expressed in a local coordinate system the derivatives, and therefore the rest of the algebra, have a particularly simple form,

$$\begin{aligned} D &= \begin{pmatrix} 0 \\ 1 \end{pmatrix} && (6) \\ D^T V D &= V_{0_{bb}} && (7) \\ V D D^T V &= \begin{pmatrix} 0 & 0 \\ 0 & V_{0_{bb}}^2 \end{pmatrix} && (8) \\ V' &= \begin{pmatrix} V_{0_{mm}} & 0 \\ 0 & V_{0_{bb}} - \frac{V_{0_{bb}}^2}{\sigma^2 + V_{0_{bb}}} \end{pmatrix} && (9) \end{aligned}$$

The above form is what the computer program actually evaluates. However its information content can be made more clear by the use of a Taylor expansion,

$$\begin{aligned} V' &\approx \begin{pmatrix} V_{0_{mm}} & 0 \\ 0 & \sigma^2 - \frac{\sigma^4}{V_{0_{bb}}} + \ldots \end{pmatrix} \\ &\approx \begin{pmatrix} V_{0_{mm}} & 0 \\ 0 & \sigma^2 \end{pmatrix}. && (10) \end{aligned}$$

Recall that the measurement at the first plane is denoted by y_E. And we identify $d(\eta) = b_0$. Therefore,

$$\begin{aligned} \eta' &= \begin{pmatrix} m_0 \\ b_0 \end{pmatrix} + \begin{pmatrix} V_{0_{mm}} & 0 \\ 0 & \sigma^2 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} \frac{(y_E - b_0)}{\sigma^2} && (11) \\ &= \begin{pmatrix} m_0 \\ y_E \end{pmatrix}. && (12) \end{aligned}$$

Not surprisingly, adding one hit tells us that we have a well defined impact parameter but no knowledge of the slope. An important property of this result is that it no longer depends on b_0 and $V_{0_{bb}}$.

At this point the interested reader can do various exercises to check appropriate values for $V_{0_{bb}}$. For example, a real life computer program will evaluate

$$V'_{bb} \quad = \quad V_{0_{bb}} - \frac{V^2_{0_{bb}}}{\sigma^2 + V_{0_{bb}}} \tag{13}$$

exactly as written here, which will have precision problems if $V_{0_{bb}}$ is too large. If, on the other hand, $V_{0_{bb}}$ is too small, then the Taylor series in equation 10 may not be truncated. If one retains the next order term and computes η', one will see that η' is biased towards b_0. This is an undesirable property because we want the final result to depend only on the hits, not on the initial guess at the trajectory.

3.3 Transport to Next Hit

In this step we transport the track from point E to point D. To do this we set up a new coordinate system on the second measurement plane and make a basis transformation into this new coordinate system. To be specific the new coordinate system has axes which are parallel to those of the old system and has an origin which is offset by $(-L, 0)$. It should be emphasized that this procedure is simply expressing the same track in a new basis.

By definition the equation of the same trajectory in the new coordinate system is given by,

$$y' \quad = \quad m'x' + b', \tag{14}$$

and we identify, $y' = y$, $x' = x + L$, $m' = m$ and $b' = b - mL$. Here (x', y') and (x, y) denote the coordinates of a fixed point in the new (old) coordinate system; they are not the coordinates of the axes of one system in the other system. In this notation, the new track parameters are,

$$\eta'' \quad = \quad \begin{pmatrix} m_0 \\ y_E - m_0 L \end{pmatrix}. \tag{15}$$

(The quantity L was defined above to be positive.)

In order to transform the covariance matrix we need to evaluate,

$$\begin{aligned} A_{ij} \quad &= \quad \frac{\partial \eta'_i}{\partial \eta_j} \\ &= \quad \begin{pmatrix} 1 & 0 \\ -L & 1 \end{pmatrix}. \end{aligned} \tag{16}$$

Therefore,

$$
\begin{aligned}
V'' &= AV'A^T \\
&= \begin{pmatrix} V_{0_{mm}} & -LV_{0_{mm}} \\ -LV_{0_{mm}} & \sigma^2 + L^2 V_{0_{mm}} \end{pmatrix}.
\end{aligned} \tag{17}
$$

Several properties of this result are worth comment. The correlation coefficients are very, very close to -1.0. The error on the impact parameter has again become large, because the extrapolation uses a slope with a large error. The error on the slope is unchanged by this operation.

To reiterate, this is the same track which we had at the start of this section. The only difference is that it is described in a new coordinate system.

3.4 Add on the Next Hit

The Kalman filter equations can now be applied again to add the hit at point D onto the track. Since the track is described in a local coordinate system, the derivatives are again simple,

$$
D = \begin{pmatrix} 0 \\ 1 \end{pmatrix}. \tag{18}
$$

This gives the new covariance matrix,

$$
\begin{aligned}
V''' &= \begin{pmatrix} V_{0_{mm}} & -LV_{0_{mm}} \\ -LV_{0_{mm}} & \sigma^2 + L^2 V_{0_{mm}} \end{pmatrix} \\
&\quad - \frac{1}{2\sigma^2 + L^2 V_{0_{mm}}} \begin{pmatrix} L^2 V_{0_{mm}}^2 & -LV_{0_{mm}}(\sigma^2 + L^2 V_{0_{mm}}) \\ -LV_{0_{mm}}(\sigma^2 + L^2 V_{0_{mm}}) & (\sigma^2 + L^2 V_{0_{mm}})^2 \end{pmatrix} \\
&\approx \begin{pmatrix} \frac{2\sigma^2}{L^2} & \frac{-\sigma^2}{L} \\ \frac{-\sigma^2}{L} & \sigma^2 \end{pmatrix}.
\end{aligned} \tag{19}
$$

This again has the expected form — V_{bb} depends only on the local information while V_{mm} depends on both σ and L. Again the Taylor expansion is done only to illustrate the information content; the actual program computes the full expression.

This time one identifies $d(\eta) = y_E - m_0 L$ and the new track parameters are,

$$
\begin{aligned}
\eta''' &= \begin{pmatrix} m_0 \\ y_E - m_0 L \end{pmatrix} + V''' \begin{pmatrix} 0 \\ 1 \end{pmatrix} \frac{(y_D - (y_E - m_0 L))}{\sigma^2} \\
&= \begin{pmatrix} \frac{(y_E - y_D)}{L} \\ y_D \end{pmatrix}
\end{aligned} \tag{20}
$$

Again, the form is as expected. In particular, all of m_0, b_0, $V_{0_{mm}}$ and $V_{0_{bb}}$ have dropped out of the answer.

At this point we have an estimator of the trajectory, which is valid in the neighbourhood of point D, but which does not use all of the available information.

3.5 Transport to and Adding the Third Hit

Now consider adding the hit at point C to the track. As before, we transport to the next hit by setting up a new coordinate system on the measurement plane; again the translation between the coordinate systems is $(-L, 0)$.

In this new basis, the track parameters of the trajectory are,

$$\begin{aligned} \eta^{(iv)} &= \begin{pmatrix} m''' \\ b''' - m'''L \end{pmatrix} \\ &= \begin{pmatrix} \frac{(y_E - y_D)}{L} \\ y_D - (y_E - y_D) \end{pmatrix} \end{aligned} \tag{21}$$

It should be emphasized that this step is simply re-expressing the same trajectory in a new basis. Again the transformation matrix for V is,

$$A = \begin{pmatrix} 1 & 0 \\ -L & 1 \end{pmatrix}. \tag{22}$$

Therefore,

$$\begin{aligned} V^{(iv)} &= AV'''A^T \\ &= \begin{pmatrix} \frac{2\sigma^2}{L^2} & \frac{-3\sigma^2}{L} \\ \frac{-3\sigma^2}{L} & 5\sigma^2 \end{pmatrix}. \end{aligned} \tag{23}$$

This has the sensible behavior that, relative to equation 19, the error on the impact parameter has grown (since we are extrapolating).

Now we can add the hit at point C to the covariance matrix,

$$\begin{aligned} \sigma^2 + D^T V^{(iv)} D &= 6\sigma^2 \\ V^{(v)} &= \begin{pmatrix} \frac{\sigma^2}{2L^2} & \frac{-\sigma^2}{2L} \\ \frac{-\sigma^2}{2L} & \frac{5\sigma^2}{6} \end{pmatrix}. \end{aligned} \tag{24}$$

Notice that the two diagonal elements are now smaller than they were when only two hits were on the track (equation 19).

Finally, the new estimator of the track parameters is,

$$\eta^{(v)} = \begin{pmatrix} \frac{y_E - y_C}{2L} \\ \frac{2y_D - y_E + 5y_C}{6} \end{pmatrix}. \tag{25}$$

The measurement y_D does not enter into the expression for the slope because of the symmetries of this particular example. When more hits are added to the track, y_D will reappear in the expression for m via $d(\eta)$ in equation 2.

An interesting exercise is to consider the case that the distance to the third hit is some distance other than L, say L'. One property which must be true is that the longer L' is, the smaller is $V^{(v)}_{mm}$ and the closer $V^{(v)}_{bb}$ approaches σ^2 from below. This result is achieved as follows. The longer is L', the closer to -1.0 is the correlation coefficient in equation 23. It is this large correlation coefficient which forces expected result when the next hit is added.

4 Including Multiple Scattering in the Fit

For this discussion the example is modified to include an infinitesimally thin scattering surface coincident with each measurement plane. This discussion proceeds as above for the stages of initialization and adding the first hit. The multiple scattering at the first plane is lost in the large initial value of $V_{0_{mm}}$. The discussion continues as above for the propagation to the second plane and the addition of the second hit. That is, we pick up the discussion at equations 19 and 20.

In this model the the scattering surfaces are infinitesimally thin. Therefore, at the scattering surface they contribute an error only to the slope, not to the intercept and not to the off-diagonal term. As the track is transported away from the scattering surface, the effects of the scattering are propagated into the error on the intercept and into the off-diagonal term.

Let the scattering surface introduce an error δ, in the slope. When this is added to the covariance matrix, still at the plane of measurement D, equation 19 becomes,

$$V^{(vi)} = \begin{pmatrix} \frac{2\sigma^2}{L^2} + \delta^2 & \frac{-\sigma^2}{L} \\ \frac{-\sigma^2}{L} & \sigma^2 \end{pmatrix}. \tag{26}$$

The track parameters, equation 20, remain unchanged.

After transport to point C the covariance matrix becomes,

$$V^{(vii)} = \begin{pmatrix} \frac{2\sigma^2}{L^2} + \delta^2 & \frac{-3\sigma^2}{L} - \delta^2 L \\ \frac{-3\sigma^2}{L} - \delta^2 L & 5\sigma^2 + \delta^2 L^2 \end{pmatrix}. \tag{27}$$

Compared with equation 23, this has a larger error in the impact parameter and correlation coefficients of larger magnitude. Any more detailed comments would require specific assumptions about the relative magnitudes of σ, δ and L.

After transport to point C, the track parameters are unchanged from before,

$$\eta^{(vii)} = \eta^{(iv)} = \begin{pmatrix} \frac{(y_E - y_D)}{L} \\ y_D - (y_E - y_D) \end{pmatrix}.$$

Adding the hit at point C to the covariance matrix gives,

$$
\begin{aligned}
V^{(viii)} &= \begin{pmatrix} \frac{2\sigma^2}{L^2} + \delta^2 & \frac{-3\sigma^2}{L} - \delta^2 L \\ \frac{-3\sigma^2}{L} - \delta^2 L & 5\sigma^2 + \delta^2 L^2 \end{pmatrix} \\
&\quad - \frac{1}{6\sigma^2 + \delta^2 L^2} \begin{pmatrix} (3\sigma^2/L + \delta^2 L)^2 & -(3\sigma^2/L + \delta^2 L)(5\sigma^2 + \delta^2 L^2) \\ -(3\sigma^2/L + \delta^2 L)(5\sigma^2 + \delta^2 L^2) & (5\sigma^2 + \delta^2 L^2)^2 \end{pmatrix}.
\end{aligned} \tag{28}
$$

This is sufficiently complicated that it seems wisest to choose a particular example, say, $\delta^2 L^2 = \sigma^2$. Then,

$$
V^{(viii)} = \begin{pmatrix} \frac{5\sigma^2}{7L^2} & \frac{-4\sigma^2}{7L} \\ \frac{-4\sigma^2}{7L} & \frac{6\sigma^2}{7} \end{pmatrix}. \tag{29}
$$

Compare this with equation 24 and see that, as expected, the slope and intercept are more poorly measured with multiple scattering than without. Also the covariance matrix elements have a slightly smaller magnitude. Among other things, this means that the next hit will add less information to the slope. Keeping with the example of $\delta^2 L^2 = \sigma^2$, the track parameters in the neighbourhood of point C are,

$$
\eta^{(viii)} = \begin{pmatrix} \frac{3y_E + y_D - 4y_C}{7L} \\ \frac{2y_D - y_E + 6y_C}{7} \end{pmatrix}. \tag{30}
$$

This can be compared with equation 25. Now, the symmetry has been broken by the multiple scattering and m again depends on y_D. Also b is now pulled more strongly by y_C than it was without multiple scattering.

5 Finishing the Fit

In this example, with or without multiple scattering, the next step is to propagate the track to the plane at point B and to add the hit from that plane. Then repeat again for point A. There is nothing new to learn by wading through the algebra of these steps.

When the hit at point A has been added, one has an estimator of the trajectory which uses all of the available information and which is valid in the neighbourhood of point A. To obtain an estimator of the trajectory in the neighbourhood of the IP, one would extrapolate the track from point A to the IP. In practice one would make one last basis transformation, using the nominal IP as the origin.

6 Computing Hit Residuals

After finishing the fit as described in the previous section, the result is a set of track parameters, and their covariance matrix, which are valid in the neighbourhood of point A. These track parameters can be used to compute the residual of the measurement at point A. It is clear how to compute the residual either including the hit at point A in the fit, or excluding it from the fit.

The entire procedure could also have been started at point A and run to point E to give an estimator of the trajectory which is valid in the neighbourhood of point E. This set of track parameters can be used to compute the hit residual at point E.

But what about the residual at point C? One must not start with the track which is valid in the neighbourhood of A and extrapolate it to C. This would give an incorrect treatment of the multiple scattering between C and A. Similarly, one must not start with the track from the reverse fit, which is valid in the neighbourhood of E. When doing the fit from E to A, there was a track that was valid in the neighbourhood of C. However that track did not include the information from two hits and so it must not be used to compute the residual at C. Similarly for the track valid in the neighbourhood of C which existed during the fit from A to E.

The answer is the following. Fit the track from E to D and extrapolate the fit result to C. Fit the track from A to B and extrapolate the result to C. Call these two extrapolated results (η_1, V_1) and (η_2, V_2). The optimal estimator of the trajectory in the neighbourhood of C, (η, V), is,

$$\begin{aligned} V &= (V_1^{-1} + V_2^{-1})^{-1} \\ \eta &= V(V_1^{-1}\eta_1 + V_2^{-1}\eta_2). \end{aligned} \tag{31}$$

The track parameters (η, V) can be used to compute the residual at point C, without including the hit at point C itself in the fit. If one wants the hit included in the fit, one can include it in either one of the contributing fits or one can add it in at the end.

And a side note. Equation 31 is reminiscent of the formula for the weighted mean of two numbers,

$$\bar{x} = \frac{1}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}} \left(\frac{x_1}{\sigma_1^2} + \frac{x_2}{\sigma_2^2} \right). \tag{32}$$

Now, what about the residual at point B? As it turns out, one simply follows the same procedure as for point C. One might worry that the track fit which started at A contains only the information from one hit, which is not enough to define a straight line. While this is indeed true, it has no effect on the result; the information about the track is contained in its covariance matrix and it will be give the correct weight when applying equation 31.

The above procedure can be implemented for every hit on the track using a single E to A pass of the filter and a single A to E pass of the filter. It is necessary, of course, to store the intermediate results at each hit during the first pass. With this technique it is as quick and easy to compute residuals excluding the local hit from the fit as it is to compute residuals which include the local hit in the fit. This is in contrast to global fitter techniques which, for a track with N_{hit} hits, require N_{hit} separate fits to compute residuals that exclude each local hit in turn.

7 More Complex Treatments of Multiple Scattering

In the above example the treatment of multiple scattering was very simplified — it was assumed that the detector planes were infinitesimally thin and that there was no scattering in the gas between planes. In order to treat the planes as a thick scattering surface, one need only modify equation 26 and add terms to the remaining elements of V. One can also add the scattering accumulated in the gas just before adding each hit. A more complete version might add half of the scattering in each plane before adding the hit and then add the remaining half.

8 Constraints on the Initial Track Parameters

Inspection of the above example will show that, for a linear trajectory, the initial values of m_0 and b_0 are completely irrelevant. So long as $V_{0_{mm}}$ and $V_{0_{bb}}$ are sufficiently large, m_0 and b_0 have no effect on the output. This arises from two things: the derivatives of the measurements with respect to the track parameters are independent of the track parameters and the transport derivatives (equation 16) are independent of the track parameters.

When the track follows a non-linear trajectory, for example a helix, these two conditions no longer apply and the result of the fit will depend on the starting parameters. In practice one finds that, for a very wide range of starting parameters, the variation in the final filtered values is on the order of 1% of the error on the parameters. Variations outside of this range only occur when there has been a gross failure of pattern recognition and the Kalman filter is asked to operate on hits which do not, in fact, form a track. The bottom line is that one can safely seed a Kalman filter using the same values which are used to seed global final fitters.

With non-linear trajectories there is another problem which can occur early in the fit — there are rare cases in which an unusual pattern of residuals among the first few hits can throw the fit outside of its radius of convergence. The solution to this problem is well understood and is described in reference [1]. In brief, the solution is to expand around the seed track and not to update the seed until the error on the expansion parameters is small.

9 Final Comments

The method of using a new coordinate system for each hit is not intrinsic to the Kalman filter algorithm. The example could equally have been implemented using a fixed coordinate system. In the method of moving coordinate systems, however, the information content of the procedure is much more clear. Other important features of the moving coordinate system are that it greatly simplifies the algebra and that it makes the code more robust against numerical instabilities.

References

[1] R. Kutschke and A. Ryd, "Billoir Fitter for CLEO II", `http://cithe502.cithep.caltech.edu/ ryd/klmn.ps`.

[2] P. Avery, "Fitting Theory V", `http://www.phys.ufl.edu/ avery/fitting.html`. You should ignore the section which discusses the initial instabilities in the fit.