Data Fusion with Linear Kalman Filter
Least Squares Estimation

Steven Dumble, PhD

In this chapter we will investigate least squares estimation, in which the basic idea is to estimate a quantity by minimizing a cost function of the squared error. The main principle is to use a quadratic cost function so that a minimum point (i.e the stationary) point of the cost surface can be found, which is then by design, the solution that minimizes the error and hence is the best estimate.

1 Estimation of a Constant Scalar

In this section we will look at how to estimate a constant from a series of nosy measurements of that constant. For example, we might have a temperature sensor attached to an engine and we would want to estimate the temperature of the engine. Each temperature measurement is very noisy since we have only a cheap sensor. So we will take multiple measurements with the sensor to try to get a better estimate of the true temperature.
So lets put this in mathematical terms, Suppose we have a number of measurements y_i where $i \in \{1, \ldots, k\}$ of the unknown quantity x (in this case the engine temperature). Each of the measurements are corrupted by some amount of random, white, uncorrelated noise of zero mean v_i, such that $E(V) = 0$ and $E(v_i v_j^T) = 0$ where $v_i \in V$. Each measurement of the quantity can be expressed as:

$$y_i = x + v_i$$

Now to estimate the true temperature from all the measurements, a reasonable idea would just to average together all the measurements that have been made:

$$\bar{y} = \frac{1}{k} \sum_{i=1}^{k} y_i$$

$$\begin{aligned} \bar{y} &= \frac{1}{k} \sum_{i=1}^{k} (x + v_i) \\ &= \frac{1}{k}(kx) + \frac{1}{k} \sum_{i=1}^{k} v_i \\ &= x + \bar{v} \end{aligned}$$

We know that the mean of all the noise (or the expected value) is equal to zero from the probability distribution:

$$\bar{v} = E(V) = 0$$

So the best estimate of x which we will denote as $\hat{x}$ is calculated from the mean of all the measurements:

$$\hat{x} = \bar{y}$$

This relies on the fact that if we take enough measurements, the average value of all the noise should be zero mean, so it should all cancel out. The number of measurements that need to be made will depend on the required accuracy of the estimate required and the size of the noise distribution.

2 Linear Least Squares

Lets extend the estimation of a constant scalar in the previous section to a constant vector. Suppose we that we have a measurement noisy measurement y_i and this measurement is now a linear combination of the elements of the vector x to be estimated, such that:

$$\begin{aligned} y_1 &= H_{11}x_1 + \cdots + H_{1n}x_n + v_1 \\ &\vdots \\ y_k &= H_{k1}x_1 + \cdots + H_{kn}x_n + v_k \end{aligned}$$

The quantity that we want to estimate is now a n-dimensional vector and we can write the above system of equations in a matrix form:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_k \end{bmatrix} = \begin{bmatrix} H_{11} & H_{12} & \dots & H_{1n} \\ H_{21} & H_{22} & \dots & H_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ H_{k1} & H_{k2} & \dots & H_{kn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_k \end{bmatrix}$$

This equation can be be written as:

$$y = Hx + v$$

We want to calculate the value of x, but we don't know the random values of v, so we want to come up with the best estimate of x that we can with all the information that we have available. Lets denote this best estimate again as $\hat{x}$. The error residual ϵ or difference between the measurements y and the best estimate vector $H\hat{x}$ is:

$$\epsilon = y - H\hat{x}$$

If we can make the vector ϵ as small as possible (in magnitude), then the estimate $\hat{x}$ should be as close to the true value of x that we can get. Lets define a cost function J,

this function is a single scalar value that is related to the error residual, if we make this value of J small, then the whole error residual vector ϵ should also be small. Let the cost function J be:

$$\begin{aligned} J &= \epsilon_1^2 + \epsilon_2^2 + \cdots + \epsilon_k^2 \\ &= \epsilon^T \epsilon \end{aligned}$$

The cost function J is simply the sum of the squared errors. We can expand the cost function to be:

$$\begin{aligned} J &= \epsilon^T \epsilon \\ &= (y - H\hat{x})^T (y - H\hat{x}) \\ &= y^T y - \hat{x}^T H^T y - y^T H\hat{x} + x^T H^T H\hat{x} \end{aligned}$$

To find the minimum of this equation, we can differentiate it with respect to $\hat{x}$ and set the derivate to zero. This calculates the stationary points of the equation, and we know that the stationary points can be a maximum or minimum points on a curve. Solving the derivative for $\hat{x}$ should then calculate the location or value of $\hat{x}$ that minimizes the function J.

$$\begin{aligned} \frac{\partial J}{\partial \hat{x}} &= -y^T H - y^T H + 2\hat{x}^T H^T H \\ &= 0 \end{aligned}$$

Solving for $\hat{x}$ gives:

$$\begin{aligned} H^T y &= H^T H\hat{x} \\ \hat{x} &= \left(H^T H\right)^{-1} H^T y \end{aligned}$$

This is then the equation for the least squares solution (the equation that minimizes the sum of the squared error). For this equation to be tractable, the matrix H must be full rank and the inverse of $(H^T H)$ must exist. The number of measurements k must be greater than the number of elements n in the vector $\hat{x}$ to estimate.

Lets take another look at our engine temperature problem. This time we would like to calculate how the engine temperature changes with the speed of the engine, i.e. RPMs (revolutions per minute). We take measurements of the engine temperature y_i at different RPMs speeds r_i and we want to find a linear line (of line form $y = ax + b$) of best fit between the temperature and RPM, such that:

$$y_i = x_1 r_i + x_2 + v_i$$

We can do this via the least squares solution of writing the measurements as a linear combination of the vector we want to estimate:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_k \end{bmatrix} = \begin{bmatrix} r_1 & 1 \\ r_2 & 1 \\ \vdots & \vdots \\ r_k & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_k \end{bmatrix}$$

so that the estimate vector $\hat{x}$ is just the parameters of the line of best fit that we want to estimate.

3 Weighted Least Squares

In the previous solutions, we did not know or include any information about the noise values v_i in the solutions. What happens if we do know some information? What happens if we know some information about how accurate each measurement is or that some measurements may be more accurate than others? How we include that information into the solution, and can we use that information to calculate how accurate our solution is?

Lets extend the example, suppose the vector x to be estimated is a constant n-dimensional vector, and y is a k-dimensional nosy measurement vector that is a linear combination of x via the model matrix H. Each element in the measurement matrix has some additive measurement noise components v_i and that noise has a variance of σ_i^2. The problem can be expressed mathematically as:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_k \end{bmatrix} = \begin{bmatrix} H_{11} & H_{12} & \dots & H_{1n} \\ H_{21} & H_{22} & \dots & H_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ H_{k1} & H_{k2} & \dots & H_{kn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_k \end{bmatrix}$$

$$E(v_i^2) = \sigma_i^2 \quad (i = 1, \dots, k)$$

$$E(vv^T) = R = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_k^2 \end{bmatrix}$$

Where we want to minimize the cost function J with respect to $\hat{x}$ such that:

$$\begin{aligned} J &= \frac{\epsilon_1^2}{\sigma_1^2} + \frac{\epsilon_2^2}{\sigma_2^2} + \dots + \frac{\epsilon_1^k}{\sigma_k^2} \\ &= \epsilon^T R^{-1} \epsilon \end{aligned}$$

So this new cost function minimizes the sum of the squares of the weighted error residuals (hence the name weighted least squares). Each weighting is based on the expected noise variance, the larger the variance, the small the weight is given to that measurement. It is only the relative weightings between the measurements that matter, if all the weights were the same then the solution would be the same as the least squares. The overall magnitude of the wights do not matter, they don't change the minimum solution $\hat{x}$, only the relative size of J at the minimum point.

We can again calculate the least squares solution by differentiating the cost function and finding the minimum point by calculating the solution that sets the derivative to zero:

$$
\begin{aligned}
J &= \epsilon^T R^{-1} \epsilon \\
&= (y - H\hat{x})^T R^{-1} (y - H\hat{x}) \\
&= y^T R^{-1} y - \hat{x}^T H^T R^{-1} y - y^T R^{-1} H\hat{x} + x^T H^T R^{-1} H\hat{x}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial J}{\partial \hat{x}} &= -y^T R^{-1} H + \hat{x}^T H^T R^{-1} H \\
&= 0 \\
H^T R^{-1} y &= H^T R^{-1} H\hat{x} \\
\hat{x} &= (H^T R^{-1} H)^{-1} H^T R^{-1} y
\end{aligned}
$$

We can also estimate the uncertainty on the estimates using the transformation of uncertainty error propagation. We know that if we have a linear relationship $f = Ax$, then we can transform the x uncertainty covariance matrix Σ_x into the f uncertainty covariance matrix Σ_f via the relationship $\Sigma_f = A\Sigma_x A^T$, so extending that to the weighted least squares solution:

$$\Sigma_{\hat{x}} = \left[(H^T R^{-1} H)^{-1} H^T R^{-1}\right] \Sigma_y \left[(H^T R^{-1} H)^{-1} H^T R^{-1}\right]^T$$

If assume that we have modelled the uncertainty correctly such that $\Sigma_y = R$ then the equation simplifies to:

$$\Sigma_{\hat{x}} = (H^T R^{-1} H)^{-1}$$

4 Recursive Least Squares

We have seen how to calculate an estimate of a constant with the least squares solution. We build up a model matrix H from the different measurements and solve a matrix equation. This requires all the measurements to be available at the time we calculate the solution, what happens if we want to update the estimate as the measurements become available? Do we keep storing all the past measurements and add new rows to the H matrix? A better way is to recursive update the estimate every time a new measurement is available, doing it this way would mean that we don't have to keep a history of all the previous measurements, we would only have to keep the best estimate from last time.

Suppose $\hat{x}_k$ is the estimated constant n-dimensional vector that includes all the measurement information for all the measurements up to and including the k-th measurement. Let y_k be the k-th noisy measurement vector that is a linear combination of x via the model matrix H_k and is corrupted with some random additive measurement noise v_k.

We can form a linear recursive estimator with the form:

$$
\begin{aligned}
y_k &= H_k x + v_k \\
\hat{x}_k &= \hat{x}_{k-1} + K_k(y_k - H_k \hat{x}_{k-1})
\end{aligned}
$$

We compute the latest estimate for $\hat{x}_k$ based on the previous estimate $\hat{x}_{k-1}$ and the new measurement y_k. The amount that the estimate $\hat{x}$ changes from the previous estimate is based on the error between the current measurement y_k and the estimate measurement calculated from $H_k\hat{x}_{k-1}$. The error is multiplied by an gain matrix K_k to calculate the amount to update the estimate. This gain matrix K_k will be selected to be a matrix which is optimum, to make the solution minimize a least squares cost function.
Lets have a look at the estimation error $\epsilon_k = x - \hat{x}_k$ and write it out in a recursive form:

$$\begin{aligned}
\epsilon_k &= x - \hat{x}_k \\
&= x - \hat{x}_{k-1} - K_k(y_k - H_k\hat{x}_{k-1}) \\
&= \epsilon_{k-1} - K_k(H_k x + v_k - H_k\hat{x}_{k-1}) \\
&= \epsilon_{k-1} - K_k H_k(x - \hat{x}_{k-1}) - K_k v_k \\
&= (I - K_k H_k)\epsilon_{k-1} - K_k v_k
\end{aligned}$$

So the latest estimation error is a function of the previous estimation error and the current measurement noise. Lets define a estimation error covariance matrix P_k which is a measure of the error in the estimation:

$$P_k = \epsilon_k \epsilon_k^T$$

Now lets expand the covariance matrix P_k into a recursive form:

$$\begin{aligned}
P_k &= \epsilon_k \epsilon_k^T \\
&= [(I - K_k H_k)\epsilon_{k-1} - K_k v_k][\dots]^T \\
&= (I - K_k H_k)\epsilon_{k-1}\epsilon_{k-1}^T(I - K_k H_k)^T + K_k v_k v_k^T K_k^T \\
&\quad - (I - K_k H_k)\epsilon_{k-1} v_k^T K_k^T - K_k v_k \epsilon_{k-1}^T (I - K_k H_k)^T
\end{aligned}$$

Now we know $P_{k-1} = \epsilon_{k-1}\epsilon_{k-1}^T$ is the estimation error covariance matrix for the previous estimate and the measurement noise covariance is $R_k = v_k v_k^T$ from the noise properties of the measurement. We also know that the estimation error for $k-1$ is independent of the noise on measurement k, so that $E(\epsilon_{k-1} v_k^T) = E(v_k \epsilon_{k-1}^T) = 0$, therefore we can simplify the recursive covariance matrix equation to:

$$P_k = (I - K_k H_k)P_{k-1}(I - K_k H_k)^T + K_k R_k K_k^T$$

So now lets define a cost function J_k to minimize the sum of all the estimation errors:

$$\begin{aligned}
J_k &= (x_1 - \hat{x}_1)^2 + (x_2 - \hat{x}_2)^2 + \cdots + (x_k - \hat{x}_k)^2 \\
&= \epsilon_1^2 + \epsilon_2^2 + \cdots + \epsilon_k^2 \\
&= \epsilon_k^T \epsilon_k
\end{aligned}$$

So we can use the relationship $Tr(P_k) = \epsilon_k^T \epsilon_k$, so that minimizing the cost function J_k is the same as minimizing the trace of the estimation error covariance matrix P_k.

We have the cost function J that we would like to minimize, so we can simply follow the method we have used in the past, but this time instead of calculating $\hat{x}$ to minimize the solution, we will calculate the optimum gain matrix K_k which will minimize the estimation error and hence drive $\hat{x}_k$ towards x.
The derivative of the cost function is:

$$\frac{\partial J_k}{\partial K_k} = -2(I - K_k H_k)P_{k-1}H_k^T + 2K_k R_k$$

Setting the derivative of the cost function to zero and solving for K_k gives:

$$K_k R_k = (I - K_k H_k)P_{k-1}H_k^T \tag{1}$$
$$K_k(R_k + H_k P_{k-1}H_k^T) = P_{k-1}H_k^T \tag{2}$$
$$K_k = P_{k-1}H_k^T(H_k P_{k-1}H_k^T + R_k)^{-1} \tag{3}$$

So to use the recursive least squares you first initialize the estimator:

$$\hat{x}_0 = E(x)$$

$$P_0 = E[(x - \hat{x}_0)(x - \hat{x}_0)^T]$$

So that $\hat{x}_0$ is the best estimate or initial guess and P_0 is the current uncertainty of the estimate.
Next every time a new measurement y_k is made then you update the solution using:

$$K_k = P_{k-1}H_k^T(H_k P_{k-1}H_k^T + R_k)^{-1}$$
$$\hat{x}_k = \hat{x}_{k-1} + K_k(y_k - H_k\hat{x}_{k-1})$$
$$P_k = (I - K_k H_k)P_{k-1}(I - K_k H_k)^T + K_k R_k K_k^T$$