Ramsey Faragher

Understanding the Basis of the Kalman Filter Via a Simple and Intuitive Derivation

This article provides a simple and intuitive derivation of the Kalman filter, with the aim of teaching this useful tool to students from disciplines that do not require a strong mathematical background. The most complicated level of mathematics required to understand this derivation is the ability to multiply two Gaussian functions together and reduce the result to a compact form.

The Kalman filter is over 50 years old but is still one of the most important and common data fusion algorithms in use today. Named after Rudolf E. Kálmán, the great success of the Kalman filter is due to its small computational requirement, elegant recursive properties, and its status as the optimal estimator for one-dimensional linear systems with Gaussian error statistics [1]. Typical uses of the Kalman filter include smoothing noisy data and providing estimates of parameters of interest. Applications include global positioning system receivers, phase-locked loops in radio equipment, smoothing the output from laptop trackpads, and many more.

From a theoretical standpoint, the Kalman filter is an algorithm permitting exact inference in a linear dynamical system, which is a Bayesian model similar to a hidden Markov model but where the state space of the latent variables is continuous and where all latent and observed variables have a Gaussian distribution (often a multivariate Gaussian distribution). The aim of this lecture note is to permit people who find this description confusing or terrifying to understand the basis of the Kalman filter via a simple and intuitive derivation.

Digital Object Identifier 10.1109/MSP.2012.2203621
Date of publication: 20 August 2012

RELEVANCE

The Kalman filter [2] (and its variants such as the extended Kalman filter [3] and unscented Kalman filter [4]) is one of the most celebrated and popular data fusion algorithms in the field of information processing. The most famous early use of the Kalman filter was in the Apollo navigation computer that took Neil Armstrong to the moon, and (most importantly) brought him back. Today, Kalman filters are at work in every satellite navigation device, every smart phone, and many computer games.

THE KALMAN FILTER IS OVER 50 YEARS OLD BUT IS STILL ONE OF THE MOST IMPORTANT AND COMMON DATA FUSION ALGORITHMS IN USE TODAY.

The Kalman filter is typically derived using vector algebra as a minimum mean squared estimator [5], an approach suitable for students confident in mathematics but not one that is easy to grasp for students in disciplines that do not require strong mathematics. The Kalman filter is derived here from first principles considering a simple physical example exploiting a key property of the Gaussian distribution—specifically the property that the product of two Gaussian distributions is another Gaussian distribution.

PREREQUISITES

This article is not designed to be a thorough tutorial for a brand-new student to the Kalman filter, in the interests of being concise, but instead aims to provide tutors with a simple method of teaching the concepts of the Kalman filter to students who are not strong mathematicians. The reader is expected to be familiar with vector notation and terminology associated with Kalman filtering such as the state vector and covariance matrix. This article is aimed at those who need to teach the Kalman filter to others in a simple and intuitive manner, or for those who already have some experience with the Kalman filter but may not fully understand its foundations. This article is not intended to be a thorough and standalone education tool for the complete novice, as that would require a chapter, rather than a few pages, to convey.

PROBLEM STATEMENT

The Kalman filter model assumes that the state of a system at a time t evolved from the prior state at time $t-1$ according to the equation

$$\mathbf{x}_t = \mathbf{F}_t\mathbf{x}_{t-1} + \mathbf{B}_t\mathbf{u}_t + \mathbf{w}_t, \qquad (1)$$

where

- $\mathbf{x}_t$ is the state vector containing the terms of interest for the system (e.g., position, velocity, heading) at time t
- $\mathbf{u}_t$ is the vector containing any control inputs (steering angle, throttle setting, braking force)
- $\mathbf{F}_t$ is the state transition matrix which applies the effect of each system state parameter at time $t-1$ on the system state at time t (e.g., the position and velocity at time $t-1$ both affect the position at time t)
- $\mathbf{B}_t$ is the control input matrix which applies the effect of each

1053-5888/12/$31.00©2012IEEE

control input parameter in the vector $\mathbf{u}_t$ on the state vector (e.g., applies the effect of the throttle setting on the system velocity and position)

- $\mathbf{w}_t$ is the vector containing the process noise terms for each parameter in the state vector. The process noise is assumed to be drawn from a zero mean multivariate normal distribution with covariance given by the covariance matrix $\mathbf{Q}_t$.

Measurements of the system can also be performed, according to the model

$$\mathbf{z}_t = \mathbf{H}_t\mathbf{x}_t + \mathbf{v}_t, \tag{2}$$

where

- $\mathbf{z}_t$ is the vector of measurements
- $\mathbf{H}_t$ is the transformation matrix that maps the state vector parameters into the measurement domain
- $\mathbf{v}_t$ is the vector containing the measurement noise terms for each observation in the measurement vector. Like the process noise, the measurement noise is assumed to be zero mean Gaussian white noise with covariance $\mathbf{R}_t$.

In the derivation that follows, we will consider a simple one-dimensional tracking problem, particularly that of a train moving along a railway line (see Figure 1). We can therefore consider some example vectors and matrices in this problem. The state vector $\mathbf{x}_t$ contains the position and velocity of the train

$$\mathbf{x}_t = \begin{bmatrix} x_t \\ \dot{x}_t \end{bmatrix}.$$

The train driver may apply a braking or accelerating input to the system, which we will consider here as a function of an applied force f_t and the mass of the train m. Such control information is stored within the control vector $\mathbf{u}_t$

$$\mathbf{u}_t = \frac{f_t}{m}.$$

The relationship between the force applied via the brake or throttle during the time period Δt (the time elapsed between time epochs $t-1$ and t) and the position and velocity of the train is given by the following equations:

$$x_t = x_{t-1} + (\dot{x}_{t-1} \times \Delta t) + \frac{f_t(\Delta t)^2}{2m}$$

$$\dot{x}_t = \dot{x}_{t-1} + \frac{f_t \Delta t}{m}.$$

These linear equations can be written in matrix form as

$$\begin{bmatrix} x_t \\ \dot{x}_t \end{bmatrix} = \begin{bmatrix} 1 & \Delta t \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ \dot{x}_{t-1} \end{bmatrix} + \begin{bmatrix} \frac{(\Delta t)^2}{2} \\ \Delta t \end{bmatrix} \frac{f_t}{m}.$$

And so by comparison with (1), we can see for this example that

$$\mathbf{F}_t = \begin{bmatrix} 1 & \Delta t \\ 0 & 1 \end{bmatrix} \text{and } \mathbf{B}_t = \begin{bmatrix} \frac{(\Delta t)^2}{2} \\ \Delta t \end{bmatrix}.$$

The true state of the system $\mathbf{x}_t$ cannot be directly observed, and the Kalman filter provides an algorithm to determine an estimate $\hat{\mathbf{x}}_t$ by combining models of the system and noisy measurements of certain parameters or linear functions of parameters. The estimates of the parameters of interest in the state vector are therefore now provided by probability density functions (pdfs), rather than discrete values. The Kalman filter is based on Gaussian pdfs, as will become clear following the derivation outlined below in the "Solutions" section. To fully describe the Gaussian functions, we need to know their variances and covariances, and these are stored in the covariance matrix $\mathbf{P}_t$. The terms along the main diagonal of $\mathbf{P}_t$ are the variances associated with the corresponding terms in the state vector. The off-diagonal terms of $\mathbf{P}_t$ provide the covariances between terms in the state vector. In the case of a well-modeled, one-dimensional linear system with measurement errors drawn from a zero-mean Gaussian distribution, the Kalman filter has been shown to be the optimal estimator [1]. In the remainder of this article, we will derive the Kalman filter equations that allow us to recursively calculate $\hat{\mathbf{x}}_t$ by combining prior knowledge, predictions from systems models, and noisy measurements.

The Kalman filter algorithm involves two stages: prediction and measurement update. The standard Kalman filter equations for the prediction stage are

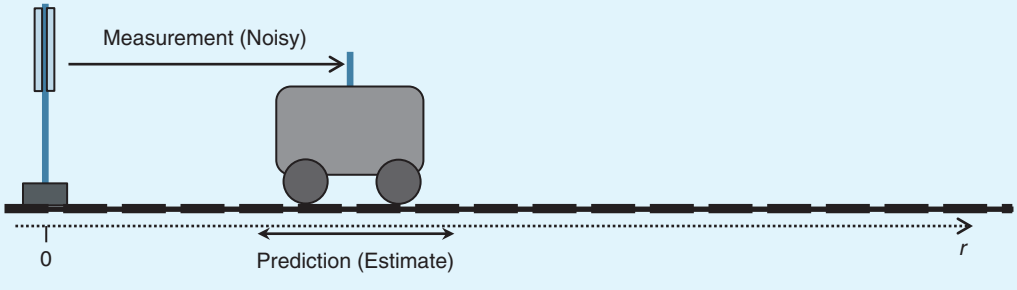
$$\hat{\mathbf{x}}_{t|t-1} = \mathbf{F}_t\hat{\mathbf{x}}_{t-1|t-1} + \mathbf{B}_t\mathbf{u}_t \tag{3}$$

$$\mathbf{P}_{t|t-1} = \mathbf{F}_t\mathbf{P}_{t-1|t-1}\mathbf{F}_t^{\mathrm{T}} + \mathbf{Q}_t, \tag{4}$$

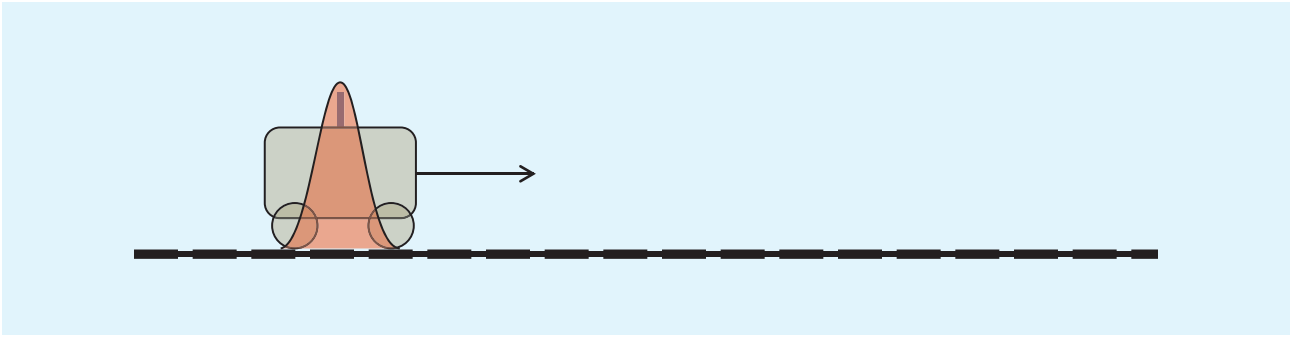
where $\mathbf{Q}_t$ is the process noise covariance matrix associated with noisy control inputs. Equation (3) was derived explicitly in the discussion above. We can derive (4) as follows. The variance associated with the prediction $\hat{x}_{t|t-1}$ of an unknown true value x_t is given by

$$P_{t|t-1} = \mathrm{E}[(x_t - \hat{x}_{t|t-1})(x_t - \hat{x}_{t|t-1})^T],$$

and taking the difference between (3) and (1) gives



[FIG1] This figure shows the one-dimensional system under consideration.



[FIG2] The initial knowledge of the system at time $t = 0$. The red Gaussian distribution represents the pdf providing the initial confidence in the estimate of the position of the train. The arrow pointing to the right represents the known initial velocity of the train.

$$
\begin{aligned}
x_t - \hat{x}_{t|t-1} &= F(x_{t-1} - \hat{x}_{t|t-1}) + w_t \\
\Rightarrow P_{t|t-1} &= \mathrm{E}[(F(x_{t-1} - \hat{x}_{t-1|t-1}) \\
&\quad + w_t) \times (F(x_{t-1} - \hat{x}_{t-1|t-1}) \\
&\quad + w_t)^T] \\
&= F\mathrm{E}[(x_{t-1} - \hat{x}_{t-1|t-1}) \\
&\quad \times (x_{t-1} - \hat{x}_{t-1|t-1})^T] \\
&\quad \times F^{\mathrm{T}} + F\mathrm{E}[(x_{t-1} \\
&\quad - \hat{x}_{t-1|t-1}) w_t^T] \\
&\quad + \mathrm{E}[w_t x_{t-1} \\
&\quad - \hat{x}_{t-1|t-1}{}^T] F^T \\
&\quad + \mathrm{E}[w_t w_t^T].
\end{aligned}
$$

Noting that the state estimation errors and process noise are uncorrelated

$$
\begin{aligned}
&E[(x_{t-1} - \hat{x}_{t-1|t-1}) w_t^T] \\
&\quad = E[w_t (x_{t-1} - \hat{x}_{t-1|t-1})^T] = 0 \\
\Rightarrow P_{t|t-1} &= F\mathrm{E}[(x_{t-1} - \hat{x}_{t-1|t-1})(x_{t-1} \\
&\quad - \hat{x}_{t-1|t-1})^T] F^T + \mathrm{E}[w_t w_t^T] \\
\Rightarrow P_{t|t-1} &= F P_{t-1|t-1} F^T + Q_t.
\end{aligned}
$$

The measurement update equations are given by

$$\hat{\mathbf{x}}_{t|t} = \hat{\mathbf{x}}_{t|t-1} + \mathbf{K}_t(\mathbf{z}_t - \mathbf{H}_t \hat{\mathbf{x}}_{t|t-1}) \quad (5)$$

$$\mathbf{P}_{t|t} = \mathbf{P}_{t|t-1} - \mathbf{K}_t \mathbf{H}_t \mathbf{P}_{t|t-1}, \quad (6)$$

where

$$\mathbf{K}_t = \mathbf{P}_{t|t-1} \mathbf{H}_t^{\mathbf{T}} (\mathbf{H}_t \mathbf{P}_{t|t-1} \mathbf{H}_t^{\mathbf{T}} + \mathbf{R}_t)^{-1}. \quad (7)$$

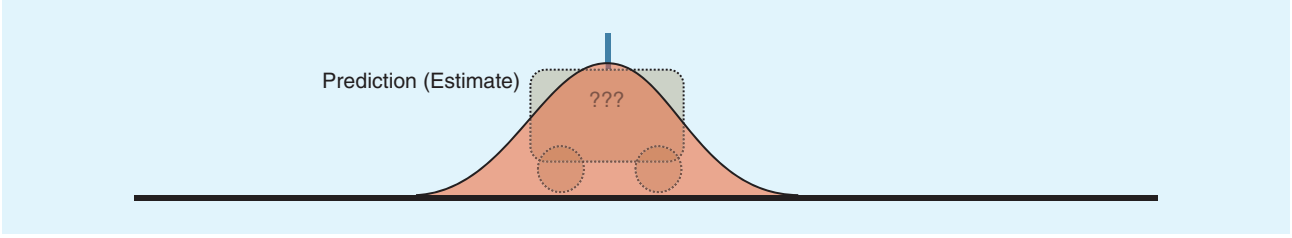
In the remainder of this article, we will derive the measurement update equations [(5)–(7)] from first principles.

SOLUTIONS

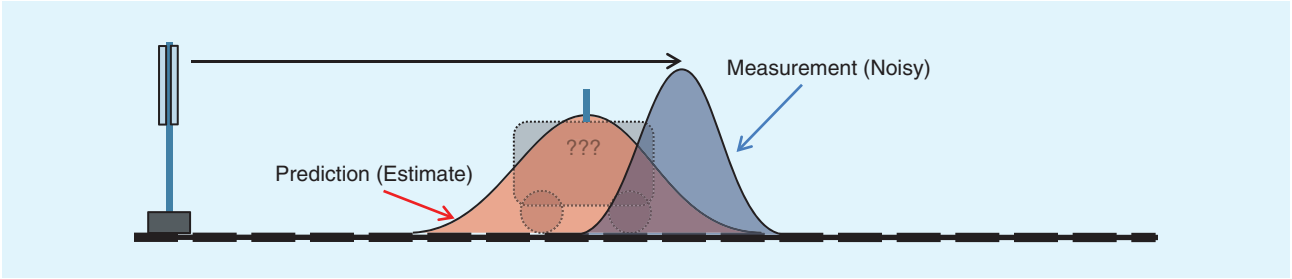
The Kalman filter will be derived here by considering a simple one-dimensional tracking problem, specifically that of a train is moving along a railway line. At every measurement epoch we wish to know the best possible estimate of the location of the train (or more precisely, the location of the radio antenna mounted on the train roof). Information is available from two sources: 1) predictions based on the last known position and velocity of the train and 2) measurements from a radio ranging system deployed at the track side. The information from the predictions and measurements are combined to provide the best possible estimate of the location of the train. The system is shown graphically in Figure 1.

THE BEST ESTIMATE WE CAN MAKE OF THE LOCATION OF THE TRAIN IS PROVIDED BY COMBINING OUR KNOWLEDGE FROM THE PREDICTION AND THE MEASUREMENT.

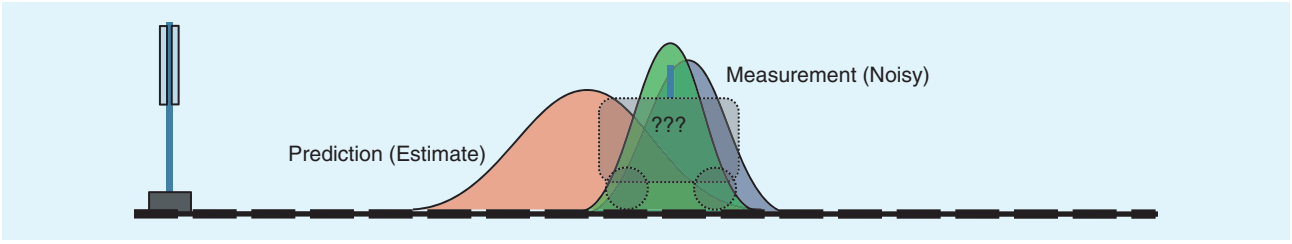
The initial state of the system (at time $t = 0$ s) is known to a reasonable accuracy, as shown in Figure 2. The location of the train is given by a Gaussian pdf. At the next time epoch ($t = 1$ s), we can estimate the new position of the train, based on known limitations such as its position and velocity at $t = 0$, its maximum possible acceleration and deceleration, etc. In practice, we may have some knowledge of the control inputs on the brake or accelerator by the driver. In any case, we have a prediction of the new position of the train, represented in Figure 3 by a new Gaussian pdf with a new mean and variance. Mathematically this step is represented by (1). The variance has increased [see (2)], representing our reduced certainty in the accuracy of our position estimate compared to $t = 0$, due to the uncertainty associated with any process noise from accelerations or decelerations undertaken from time $t = 0$ to time $t = 1$.



[FIG3] Here, the prediction of the location of the train at time $t = 1$ and the level of uncertainty in that prediction is shown. The confidence in the knowledge of the position of the train has decreased, as we are not certain if the train has undergone any accelerations or decelerations in the intervening period from $t = 0$ to $t = 1$.



[FIG4] Shows the measurement of the location of the train at time $t = 1$ and the level of uncertainty in that noisy measurement, represented by the blue Gaussian pdf. The combined knowledge of this system is provided by multiplying these two pdfs together.



[FIG5] Shows the new pdf (green) generated by multiplying the pdfs associated with the prediction and measurement of the train's location at time $t = 1$. This new pdf provides the best estimate of the location of the train, by fusing the data from the prediction and the measurement.

At $t = 1$, we also make a measurement of the location of the train using the radio positioning system, and this is represented by the blue Gaussian pdf in Figure 4. The best estimate we can make of the location of the train is provided by combining our knowledge from the prediction and the measurement. This is achieved by multiplying the two corresponding pdfs together. This is represented by the green pdf in Figure 5.

A key property of the Gaussian function is exploited at this point: the product of two Gaussian functions is another Gaussian function. This is critical as it permits an endless number of Gaussian pdfs to be multiplied over time, but the resulting function does not increase in complexity or number of terms; after each time epoch the new pdf is fully represented by a Gaussian function. This is the key to the elegant recursive properties of the Kalman filter.

The stages described above in the figures are now considered again mathematically to derive the Kalman filter measurement update equations.

The prediction pdf represented by the red Gaussian function in Figure 3 is given by the equation

$$y_1(r;\mu_1,\sigma_1) \triangleq \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(r-\mu_1)^2}{2\sigma_1^2}}. \quad (8)$$

The measurement pdf represented by the blue Gaussian function in Figure 4 is given by

$$y_2(r;\mu_2,\sigma_2) \triangleq \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{(r-\mu_2)^2}{2\sigma_2^2}}. \quad (9)$$

The information provided by these two pdfs is fused by multiplying the two together, i.e., considering the prediction and the measurement together (see Figure 5). The new pdf representing the fusion of the information from the prediction and measurement, and our best current estimate of the system, is therefore given by the product of these two Gaussian functions

$$\begin{aligned} &y_{\text{fused}}(r;\mu_1,\sigma_1,\mu_2,\sigma_2) \\ &= \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(r-\mu_1)^2}{2\sigma_1^2}} \times \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{(r-\mu_2)^2}{2\sigma_2^2}} \\ &= \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2}} e^{-\left(\frac{(r-\mu_1)^2}{2\sigma_1^2}+\frac{(r-\mu_2)^2}{2\sigma_2^2}\right)}. \end{aligned} \quad (10)$$

> **A KEY PROPERTY OF THE GAUSSIAN FUNCTION IS EXPLOITED AT THIS POINT: THE PRODUCT OF TWO GAUSSIAN FUNCTIONS IS ANOTHER GAUSSIAN FUNCTION.**

The quadratic terms in this new function can expanded and then the whole expression rewritten in Gaussian form

$$\begin{aligned} &y_{\text{fused}}(r;\mu_{\text{fused}},\sigma_{\text{fused}}) \\ &= \frac{1}{\sqrt{2\pi\sigma_{\text{fused}}^2}} e^{-\frac{(r-\mu_{\text{fused}})^2}{2\sigma_{\text{fused}}^2}}, \end{aligned} \quad (11)$$

where

$$\begin{aligned} \mu_{\text{fused}} &= \frac{\mu_1\sigma_2^2+\mu_2\sigma_1^2}{\sigma_1^2+\sigma_2^2} \\ &= \mu_1 + \frac{\sigma_1^2(\mu_2-\mu_1)}{\sigma_1^2+\sigma_2^2} \end{aligned} \quad (12)$$

and

$$\sigma_{\text{fused}}^2 = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2} = \sigma_1^2 - \frac{\sigma_1^4}{\sigma_1^2+\sigma_2^2}. \quad (13)$$

These last two equations represent the measurement update steps of the Kalman filter algorithm, as will be shown explicitly below. However, to present a more general case, we need to consider an extension to this example.

In the example above, it was assumed that the predictions and measurements were made in the same coordinate frame and in the same units. This has resulted in a particularly concise pair of

equations representing the prediction and measurement update stages. It is important to note however that in reality a function is usually required to map predictions and measurements into the same domain. In a more realistic extension to our example, the position of the train will be predicted directly as a new distance along the railway line in units of meters, but the time of flight measurements are recorded in units of seconds. To allow the prediction and measurement pdfs to be multiplied together, one must be converted into the domain of the other, and it is standard practice to map the predictions into the measurement domain via the transformation matrix $\mathbf{H}_t$.

We now revisit (8) and (9) and, instead of allowing y_1 and y_2 to both represent values in meters along the railway track, we consider the distribution y_2 to represent the time of flight in seconds for a radio signal propagating from a transmitter positioned at x $=$ 0 to the antenna on the train. The spatial prediction pdf y_1 is converted into the measurement domain by scaling the function by c, the speed of light. Equations (8) and (9) therefore must be rewritten as

$$y_1(s;\mu_1,\sigma_1,c) \triangleq \frac{1}{\sqrt{2\pi(\frac{\sigma_1}{c})^2}} e^{-\frac{\left(s-\frac{\mu_1}{c}\right)^2}{2(\frac{\sigma_1}{c})^2}} \tag{14}$$

and

$$y_2(s;\mu_2,\sigma_2) \triangleq \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{(s-\mu_2)^2}{2\sigma_2^2}}, \tag{15}$$

where both distributions are now defined in the measurement domain, radio signals propagate along the time "s" axis, and the measurement unit is the second.

Following the derivation as before we now find

$$\frac{\mu_{\text{fused}}}{c} = \frac{\mu_1}{c} + \frac{\left(\frac{\sigma_1}{c}\right)^2\left(\mu_2 - \frac{\mu_1}{c}\right)}{\left(\frac{\sigma_1}{c}\right)^2 + \sigma_2^2}$$

$$\Rightarrow \mu_{\text{fused}} = \mu_1 + \left(\frac{\frac{\sigma_1^2}{c}}{\left(\frac{\sigma_1}{c}\right)^2 + \sigma_2^2}\right)\cdot\left(\mu_2 - \frac{\mu_1}{c}\right). \tag{16}$$

Substituting $H = 1/c$ and $K = (H\sigma_1^2)/(H^2\sigma_1^2 + \sigma_2^2)$ results in

$$\mu_{\text{fused}} = \mu_1 + K\cdot(\mu_2 - H\mu_1). \tag{17}$$

Similarly the fused variance estimate becomes

$$\frac{\sigma_{\text{fused}}^2}{c^2} = \left(\frac{\sigma_1}{c}\right)^2 - \frac{\left(\frac{\sigma_1}{c}\right)^4}{\left(\frac{\sigma_1}{c}\right)^2 + \sigma_2^2}$$

$$\Rightarrow \sigma_{\text{fused}}^2 \quad \sigma_1^2 \quad \frac{c^{\frac{2}{1}}}{\sigma\sigma\mu\sigma\sigma_{\frac{2}{2}}} \quad c^{\frac{2}{1}}$$