POOR MAN'S EXPLANATION OF KALMAN FILTERING

OR

HOW I STOPPED WORRYING AND LEARNED TO LOVE MATRIX INVERSION

By
Roger M. du Plessis

June 1967

North American Rockwell
Electronics Group

3370 Miraloma Avenue / Anaheim, California 92803

CONTENTS

ILLUSTRATIONS

INTRODUCTION AND SUMMARY

Dr. R. E. Kalman introduced his concept of optimum estimation in 1960. Since that time, his technique has proved to be a powerful and practical tool. The approach is particularly well suited for optimizing the performance of modern terrestrial and space navigation systems.

Many people not directly involved in systems analysis have heard about Kalman filtering and have expressed an interest in learning more about it. Although attempts have been made to supply such people with simple, intuitive explanations of Kalman filtering, it is this writer's opinion that none of these explanations has been completely successful. Almost without exception, they have tended to become enmeshed in the jargon and state-space notation of the "cult." But matrix notation, regardless of how useful and efficient it may be, does not assist the uninitiated reader to understand the concepts.

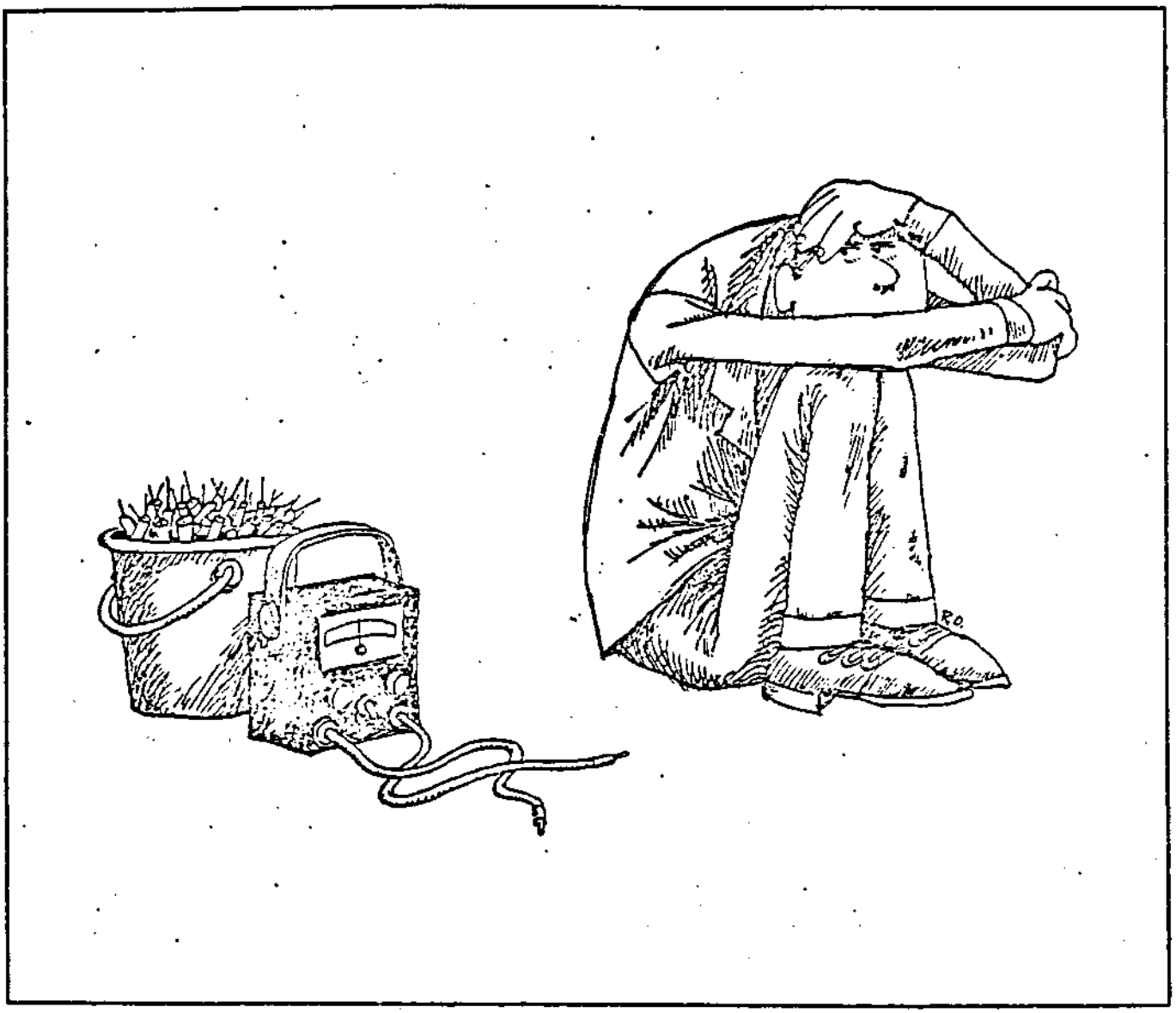
Surprisingly, in spite of all the obscure-looking mathematics (the most impenetrable of which can be found in Dr. Kalman's original paper), Kalman filtering is a fairly direct and simple concept. The full-blown matrix equations can be made intelligible by being presented in a way that appeals to the intuition, and a statistical error analysis or actual system performance data can be viewed with an intuitive understanding of the results.

The present paper presents a straightforward explanation of Kalman filtering that will be accessible to virtually all readers. The most salient features are explained by means of a very simple example, entirely free from matrix notation and complicated dynamics. A second illustrative example is taken from the field of inertial navigation, since this is one of the most fruitful areas of application for Kalman estimation. This slightly more complicated second example is used to describe more of the concepts and terminology of Kalman filters (e.g., "state vector," and so forth). At the same time, it permits a closer look at the actual operations involved in applying Kalman theory. Not all readers will have the familiarity with matrix notation or the patience to unravel every single equation in the discussion of the second example. The writer feels, however, that the gist of this discussion will be grasped by the majority of those readers who are analytically inclined.

The main body of the paper concludes with qualitative comments concerning the practical advantages and disadvantages of the Kalman technique, and the difficulties of applying it in actual systems. Again, navigation systems are particularly cited.

Finally, an appendix defines a criterion for optimum estimation and derives the optimum estimation equations used in the examples. As is the case with the main body of the paper, the appendix is intended to be tutorial and to provide, for those readers who can recall some of their probability theory, a thumbnail derivation of the estimation equations.

THE ELEMENTS OF KALMAN FILTERING



LEAST SQUARES AND MAXIMUM LIKELIHOOD ESTIMATION

In understanding the concept of Kalman filtering, it is helpful to examine the differences between two basic techniques of estimation. These are the so-called "least squares" and the "maximum likelihood" methods.* Toward this end, let us consider the following example of an estimation problem:

> Suppose that we have a bucketful of nominally 100-ohm resistors, accurate to 1 percent RMS. Suppose, further, that from this bucket we select a single resistor and that we want to estimate its actual value. We shall use an ohmmeter with an accuracy of 3 ohms RMS random error on each reading. The error in each reading is statistically uncorrelated with any other. The ohmmeter has no systemic error or biases.

This simple example has most of the elements of a complete Kalman filter, except that (1) it is a one-dimensional problem and hence does not require matrix notation, and (2) there are no dynamics in the problem. If the concepts in the example are understood, their extension to the general case is not difficult.

*These terms involve some semantic difficulties which are briefly mentioned in the appendix.

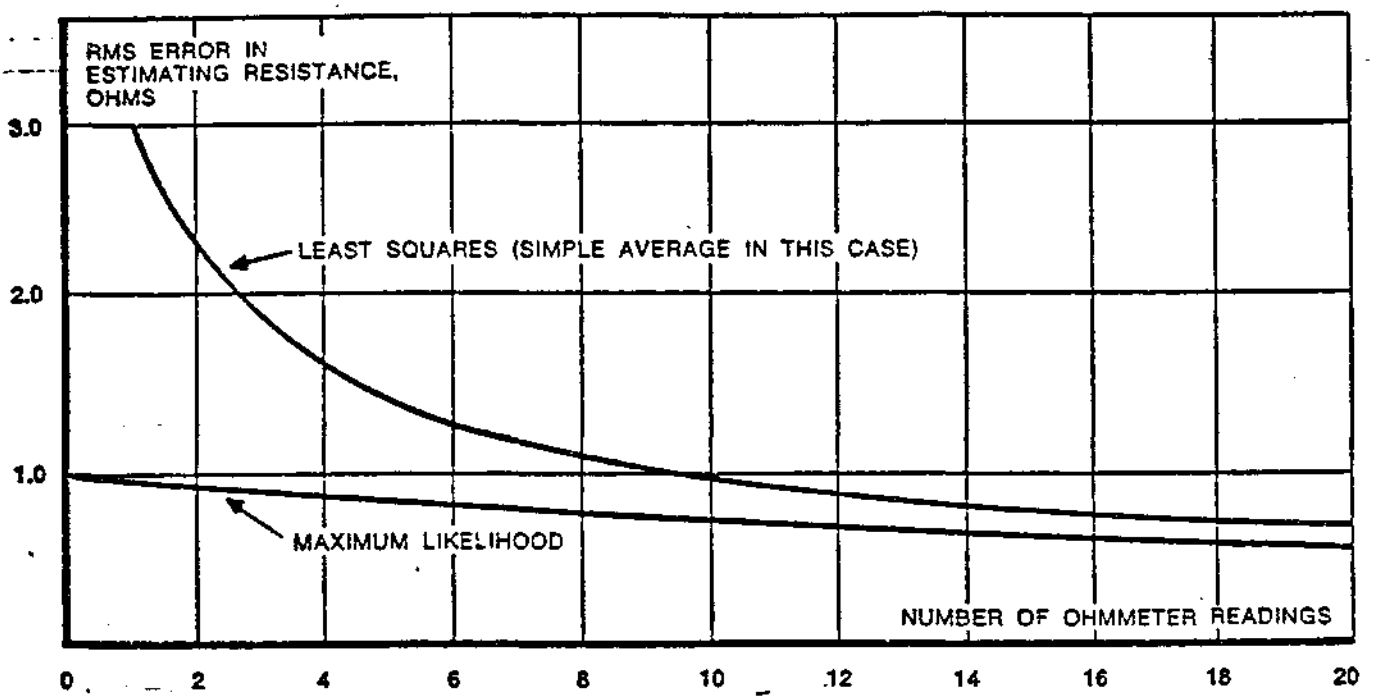


Figure 1. Comparison of Least Squares vs Maximum Likelihood Estimations of the Value of a Resistor with 1 Ohm RMS Tolerance, Using an Ohmmeter with 3 Ohms RMS Random Error

Least Squares Method

Least squares estimation involves fitting a curve to the available data in such a way that the sum square of the residuals to the data points is minimized. For our bucketful-of-resistors example, the curve to be fitted to the ohmmeter readings is simply a constant. It is well known that the least squares curve fit in this case is simply the average of the ohmmeter readings. The RMS error in the estimate is $3/\sqrt{N}$, where N is the number of readings. Until at least one ohmmeter reading is obtained, no curve can fit and no estimate can be made. In general, no least squares curve fit is possible until there are at least as many equations as there are unknowns.

The RMS error vs the number of readings is shown in the upper curve of Figure 1.

Maximum Likelihood Method and the Kalman Concept

BASIC MAXIMUM LIKELIHOOD TECHNIQUE

Maximum likelihood estimation introduces the new factor of "weighting" into the estimation process, in order to make allowances for the accuracy of the measuring instrument (in our example, an ohmmeter), as well as the accuracy of the thing being measured (in our example, a resistor). The precision of the meter may be known, possibly, from calibration data; that of the resistor may have been learned from vendor's specifications, sampled or 100-percent-inspection data, or engineering judgment based on experience or analysis of similar resistors. We thus have available to us statistics concerning both the *quantity being estimated* and the *errors in measurement*.

Now, suppose that, in our bucketful-of-resistors example, a reading of 95 ohms is indicated on the face of the ohmmeter. It is extremely unlikely that this quantity can actually be the *exact* value of the resistance. But a reading of "95" for the nominally "100-ohm" resistor could well be expected with this meter, in view of its 3-ohm RMS error. In maximum likelihood estimation, meter readings are

weighted with a factor which takes into account the relative quality of the meter, and obtains an *optimum* estimate of the actual resistance.

In explaining how the weighting factor "works," a few definitions are necessary.

Let

x = value of the resistance for the particular resistor in question (the quantity to be estimated)

ϵ = error in ohmmeter reading

M = ohmmeter scale factor, i.e., output/input (in the bucketful-of-resistors example, $M = 1$)

y = ohmmeter reading

$= Mx + \epsilon$

The maximum likelihood weighting factor is given by

$$b = \frac{M\sigma_x^2}{M^2\sigma_x^2 + \sigma_\epsilon^2}$$

where

σ_x^2 = the variance or mean square deviation of the resistance being estimated

σ_ϵ^2 = mean square error in the ohmmeter reading

As described in more detail below, the raw ohmmeter reading y is weighted by the factor b in order to get an estimate of the true value of the resistance x. While the derivation of this factor is relegated to the appendix, one can see that b has at least one desirable attribute: it takes into account the relative accuracy of the ohmmeter measurements. Assuming the scale factor of the ohmmeter to be 1, if the ohmmeter were very accurate relative to the precision of the resistors in the bucket (that is, σ_ϵ^2 much less than σ_x^2), then the weighting factor would approach 1. In other words, we simply would use the raw value of the ohmmeter reading in estimating the resistance. If, on the other hand, the resistors were very precise, say 0.1 percent from nominal, and the ohmmeter were very inaccurate and unreliable, the b would approach zero. In this instance, we would reject the ohmmeter readings and simply use the nominal value of the resistance as our estimate of x.

But we have more to go on here than just intuition. The value of b is derived on the basis that it renders the estimate *optimum*. Because of the random errors in the resistors and ohmmeter, "optimum" is taken in a statistical sense and is defined as the condition in which RMS error in the estimate is minimized. Like any other statistically optimum policy, whether in estimation or at the blackjack table, the maximum likelihood method does not yield a uniformly excellent result each and every time. But b "plays the odds" in such a way that, on the average, the best possible estimate is obtained.

We should note that there is a fundamental difference between the two estimation methods—least squares and maximum likelihood—in the latter's introduction of the concept of "weighting," which takes into account the accuracies of both the measuring instrument and the object being measured. Least squares, as we have seen, works by the fitting of data to a curve by minimizing the squared difference between the data and the curve fit—but a good data fit does not necessarily imply a good estimate. Maximum likelihood obtains an "optimum" estimate by minimizing the mean squared error in the estimate—but with bad data there can be a poor fit.

The maximum likelihood estimate (see the appendix for a derivation) is the expected value of the quantity being estimated, given the measurement data. With *no* measurements, the expected value of the resistance of our particular resistor is nominal, or 100 ohms. We have no basis for stating that it is above or below nominal, and our *best* estimate is, indeed, 100 ohms. There is an error in this estimate due to the deviation of the resistor from nominal; the RMS value of the error in estimation is the resistor's "rated" accuracy—1 ohm.

Now suppose a reading is taken of the resistance using the ohmmeter. This reading will be weighted by the factor

$$b = \frac{M\sigma_x^2}{M^2\sigma_x^2 + \sigma_\epsilon^2}$$

$$= \frac{1 \cdot 1^2}{1^2 \cdot 1^2 + 3^2}$$

$$= 0.1$$

The estimate, based on first reading y_1, is given by

$$\hat{x}_1 = 100 + b(y_1 - 100)$$

where

$\hat{x}_1$ = the estimate of the resistance x based on one ohmmeter reading. (The ^ symbol is frequently reserved for statistical estimates.)

Suppose, for example, the reading on the ohmmeter was 95 ohms. This value of resistance would be highly unusual from a bucket of 1-percent RMS resistors. It would represent a 5σ deviation from normal. However, the maximum likelihood estimate for this particular situation would be

$$\hat{x}_1 = 100 + 0.1(95 - 100)$$
$$= 99.5 \text{ ohms}$$

a result which is not at all unreasonable. Least squares, on the other hand, would minimize the sum square of the residuals by estimating 95 ohms. This gives a good data fit, but most probably is not a very accurate estimate.

Naturally, there are some errors in maximum likelihood estimates. The formula for the mean square error in a maximum likelihood estimate is given by

$$\sigma^2_{\hat{x}} = \frac{\sigma_x^2 \sigma_\epsilon^2}{M^2\sigma_x^2 + \sigma_\epsilon^2}$$
$$= (1 - bM)\sigma_x^2$$

This equation is derived in the appendix, but it can be observed that at least in one regard the formula seems reasonable. Assuming again that $M = 1$, and supposing that we have a relatively precise measurement (that is, σ_ϵ^2 is much less than σ_x^2), then the RMS error reduces to σ_ϵ, or the RMS error in the estimate is simply the RMS error in the measurement. At the other extreme, suppose the measurement was very inaccurate relative to the tolerance on the resistors; then the RMS error in the estimate is σ_x, the RMS tolerance of the resistance. In this extreme case, we have rejected the data ($b = 0$), and we are no smarter after the measurement than we were before.

For this example, the mean square error in the estimate is given by

$$\sigma_{\hat{x}}^2 = (1 - bM)\sigma_x^2$$
$$= (1 - 0.1)1^2$$
$$= 0.9 \text{ ohm}^2$$
$$\sigma_{\hat{x}} = 0.95 \text{ ohm RMS}$$

The RMS error has been reduced from 1 ohm with no measurements, to 0.95 ohm after one measurement. This is not much of an improvement, and is a reflection of the relative inaccuracy of the ohmmeter.

THE ITERATIVE NATURE OF THE KALMAN FILTER

Now suppose that a second measurement is obtained on the ohmmeter. It is at this point that the second fundamental concept of the Kalman Filter—its recursive or iterative nature—rears its head.* It is this concept, brought forth by Dr. R. E. Kalman in 1960, that has rendered maximum likelihood estimation a feasible and practical technique for use on small digital computers in real-time application. As a result of our first reading, we have an estimate of x, which we call $\hat{x}_1$, and can calculate the mean square error in the estimate, or 0.9 ohm^2. Now, when the second reading is made, we recognize similarity between this situation and the previous situation, when our estimate was a nominal 100 ohms with a mean square error of 1 ohm^2. We proceed to weight the data as before, but realizing that we have a better handle on the resistor than we did before. The new weighting factor is

$$b = \frac{M\sigma^2_{\hat{x}_1}}{M^2\sigma^2_{\hat{x}_1} + \sigma_\epsilon^2}$$
$$= \frac{1(0.9)}{1(0.9) + 3^2}$$
$$= 0.091$$

which is slightly less than the previous weighting factor of 0.1. This is because, as a result of the first ohmmeter reading, we have a more accurate estimate of x, and we will tend to weight the second ohmmeter reading more lightly. Using this new value of b, the estimate based on the two ohmmeter readings is

$$\hat{x}_2 = \hat{x}_1 + b(y_2 - M\hat{x}_1)$$

where $b = 0.091$. In this equation, $(y_2 - M\hat{x}_1)$ is the difference between the actual reading y_2 and the value of the ideal reading which would have been obtained if $\hat{x}_1$ had been perfect. It is very important to note that the form of this equation is identical to the form of the equation used to make the estimate with the first reading. It is not necessary to save the first reading y_1; the only information re-

*This is *not* to imply that least squares cannot be made iterative—it can. In the example given here, a new least squares curve fit can be derived by updating the old average to include the new data point:

$$\hat{x}_n = \frac{n-1}{n}\hat{x}_{n-1} + \frac{1}{n}y_n.$$

The result can be generalized.

quired to make the second optimum estimate is the first optimum estimate, the computed mean square error in the estimate $\sigma_{\hat{x}_1}$, and the mean square ohmmeter error σ_ϵ^2. Again, a mean square error in the estimate can be computed thus:

$$\begin{aligned}\sigma_{\hat{x}_2} &= (1 - bM)\sigma_{\hat{x}_1} \\ &= (1 - 0.091)(0.9) \\ &= 0.82 \text{ ohm}^2\end{aligned}$$

or 0.90 ohm RMS.

It can be seen that a recursive or iterative procedure is developing. Each time a reading is taken a new weighting factor b is computed. It is applied against the data to yield a new estimate x, and a mean square error in the estimate is computed. The latter quantity is necessary in computing b for the next reading, but past data need never be saved. The equations involved are summarized as shown here.

1. Compute weighting coefficient:

$$b_n = \frac{M\sigma^2_{\hat{x}_{n-1}}}{M^2\sigma^2_{\hat{x}_{n-1}} + \sigma_\epsilon^2}$$

2. Use weighting coefficient to make new estimate:

$$\hat{x}_n = \hat{x}_{n-1} + b_n(y_n - M\hat{x}_{n-1})$$

3. Update mean square error in estimation:

$$\sigma^2_{\hat{x}_n} = (1 - b_nM)\sigma^2_{\hat{x}_{n-1}}$$

To start this cyclic process, three quantities are needed: 1) the expected value of the item to be measured, prior to the taking of any readings; 2) the mean square deviation about nominal value of the item; and 3) the mean square error in the measuring instrument's indication.

The importance of the Kalman process' recursive nature cannot be overemphasized. Review of the equations presented above shows that no past data need ever be stored. Each estimation is identical in procedure to all those which took place before it, but each has a new weighting factor b computed to take into account the sum total effect of all the previous estimates.

If our resistor example is carried out for successive measurements and estimations and the RMS is computed, the results point out (Figure 1) that maximum likelihood Kalman estimation is always more accurate than the least squares method. Of course, the bucketful-of-resistors example which we have cited here is a highly artificial one. It is intended only to serve to bring out the basic concepts of Kalman estimation, and no general quantitative conclusions should be inferred from the information presented in Figure 1.

Effect of Erroneous Statistics for Resistor Example

In the preceding discussion, it was noted that three quantities had to be known before the cyclic application of the Kalman filter equations could be initiated. For the bucketful-of-resistors example,

1. The expected value of the resistance x prior to the taking of any readings—namely, 100 ohms
2. The mean square deviation about nominal —namely, 1 ohm^2
3. The mean square error in the ohmmeter readings—namely, 3 ohms^2

In any actual situation, these three statistics are only approximately known, and are often, at best, only estimates or educated guesses. Let us now take a look at the consequences of using vague or erroneous statistics in the estimation process.

For simplicity, we will assume that only the second of the three quantities listed above is in error; that is, the RMS deviation of the resistors which is assumed in the filter equations is 1 ohm, but the actual RMS deviation of the resistors in the bucket is some value other than 1 ohm. The effect on the errors in estimation for various values of the true RMS deviation will now be examined.

For illustration, let us assume that the true RMS deviation of the resistors is 2 ohms. Using the formula previously presented for b, the optimum weighting coefficient is

$$\begin{aligned}b &= \frac{M\sigma_x^2}{M^2\sigma_x^2 + \sigma_\epsilon^2} \\ &= \frac{1 \cdot 2^2}{1^2 \cdot 2^2 + 3^2} \\ &= 0.308\end{aligned}$$

This weighting coefficient will not actually be applied, however. Under the erroneous assumption that the bucket of resistors tolerance is 1 percent RMS, rather than the actual 2 percent RMS, it was

calculated in the previous section of this paper that the coefficient to be used in the estimation process was 0.100. This coefficient is too small and would result in the placing of too little weight on the ohmmeter reading. The RMS error in estimating the resistance is computed to be 1.82, in contrast to the error of 1.66 which would have resulted if the optimum weighting of 0.308 had been used.

To explore the situation more completely, Figure 2 was generated. Based on the best available data or engineering estimates, the resistor tolerance was assumed by the filter to be 1 ohm RMS, and the ohmmeter accuracy to be 3 ohms RMS. The effect of the real world's difference from the artificial world assumed in the model is shown by the various curves in Figure 2, plotted for the five conditions that actual variation in the bucket of resistors is ⅓, 1, 1½, 2, and 3 ohms RMS. Two general cases can be observed in Figure 2:

Case 1. In this case, the resistors are much more out-of-tolerance (relative to the assumed ohmmeter accuracy) than assumed, and it is possible for the Kalman procedure to yield estimation errors worse than those obtainable with ordinary least squares estimation. In other words, the weighting coefficients were too small, and we should have given more credence to the ohmmeter readings.

Case 2. In this case, the actual variation of the resistors is smaller than assumed, and the estimation errors are still better than the "optimum." However, by having been too pessimistic in the assumed tolerance on the resistors, we took too little advantage of their inherent precision. We would have been better off just to use the nominal value of 100 ohms and to forget the ohmmeter readings altogether.

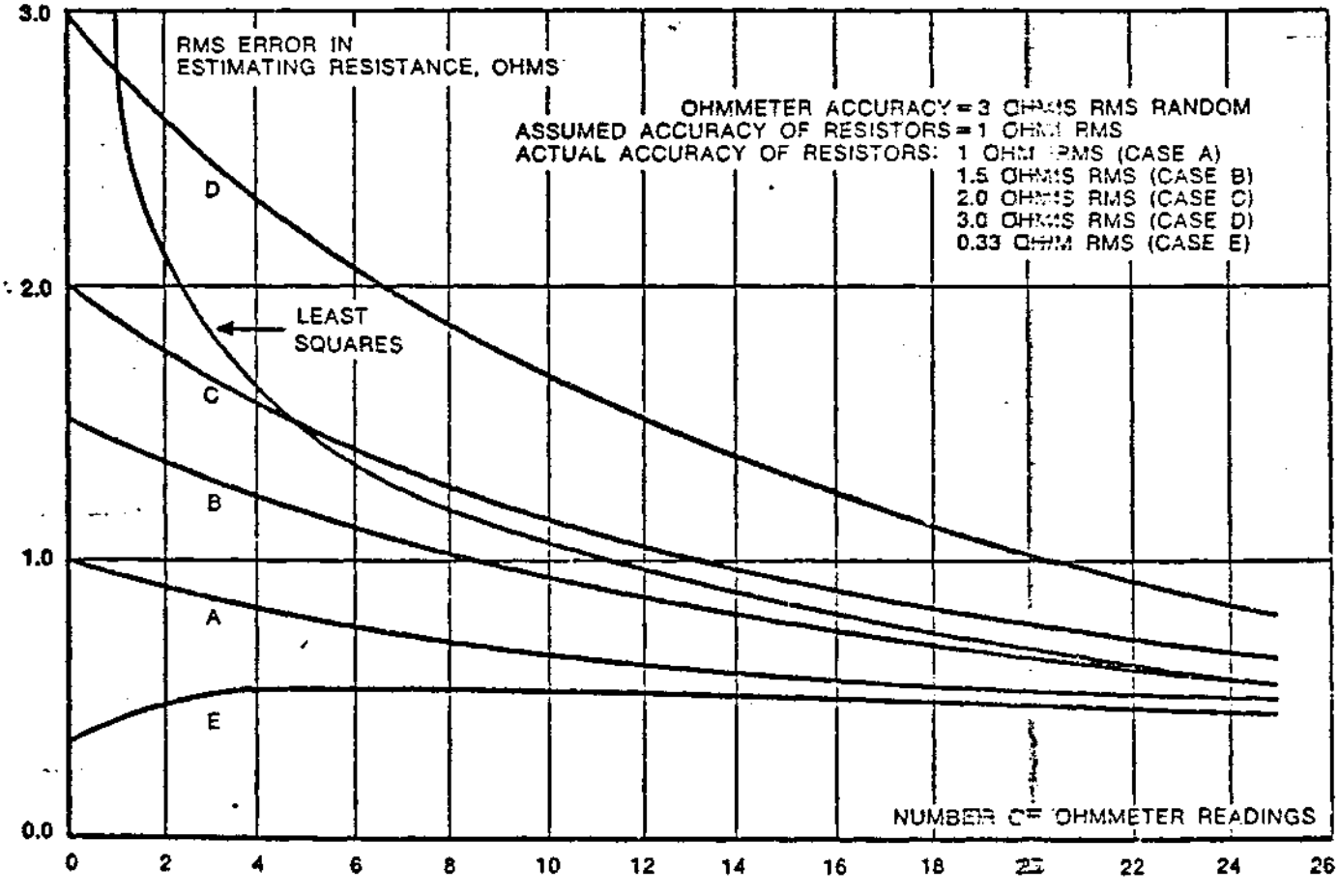


Figure 2. Comparison of the Effects of Erroneous Statistics on Least Squares and Maximum Likelihood Methods for Estimating Resistor Value

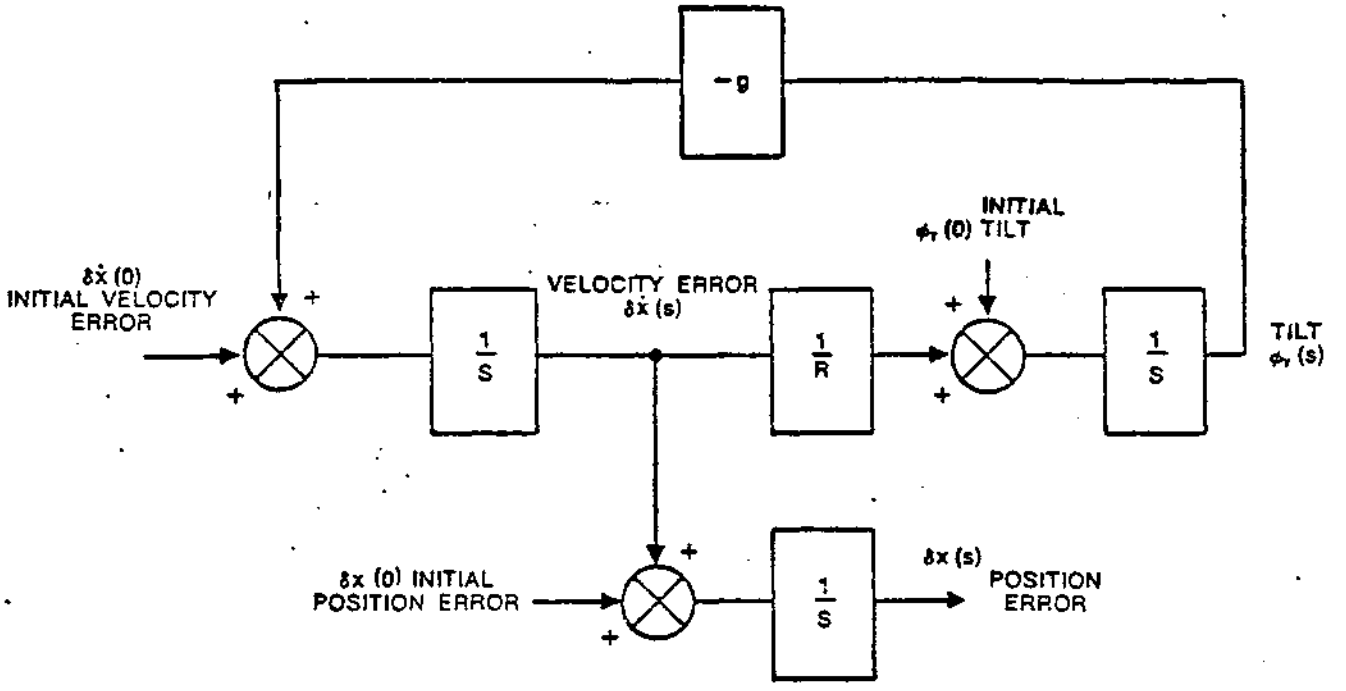


Figure 3. Simplified Schuler Loop

The moral to our story: Real-world statistics and dynamics will generally differ somewhat from the conditions assumed in the Kalman filter coefficients. Before the Kalman procedure is used in a particular application, a sensitivity analysis of the sort described above should be made to assess the risks involved in the event the real world should happen to differ appreciably from that assumed in the model. (If one is concerned about the possibility that the real world has statistics other than those assumed, one might be tempted to consider "biasing" the statistics to be employed in the filter, thereby hedging against this possibility. To accomplish this biasing, it would be necessary to alter *a priori* statistics in the proper direction to reduce the risk that would be involved if the real world were different from the model. However, it can be shown that tricks of this sort are pretty futile. The best policy is to use the expected values of the σ's, based on the available data, and to let it go at that.)

APPLICATION OF KALMAN FILTER TO DYNAMIC SYSTEMS

The bucket-of-resistors example illustrates some salient features of the Kalman procedure, but includes no dynamics. Also, it is a one-dimensional problem. The following example illustrates application of the Kalman technique to a dynamic, multi-dimensional system.

The example under consideration involves a simplified inertial navigation system.* The navigation system is described by the loop shown in Figure 3 and is subject to only three error sources: initial position error, initial velocity error, and initial tilt. We will assume that the system has available external position fixes every 10.5 minutes (or every one-eighth Schuler period), and these position fixes can be used to update system accuracy. The total assumed RMS errors acting on the system are thus:

Error	*Magnitude*	
$\delta x(0)$, initial position error	1,000 ft RMS	All uncorrelated
$\delta\dot{x}(0)$, initial velocity error	6 ft/sec RMS	
$\phi_T(0)$, initial tilt	0.1 mr RMS	
r, position-fix error	1,000 ft RMS	

*The reader who is unfamiliar with inertial navigation can ignore the physical interpretation of this example, yet benefit just as well. Figure 3 can be taken at face value.

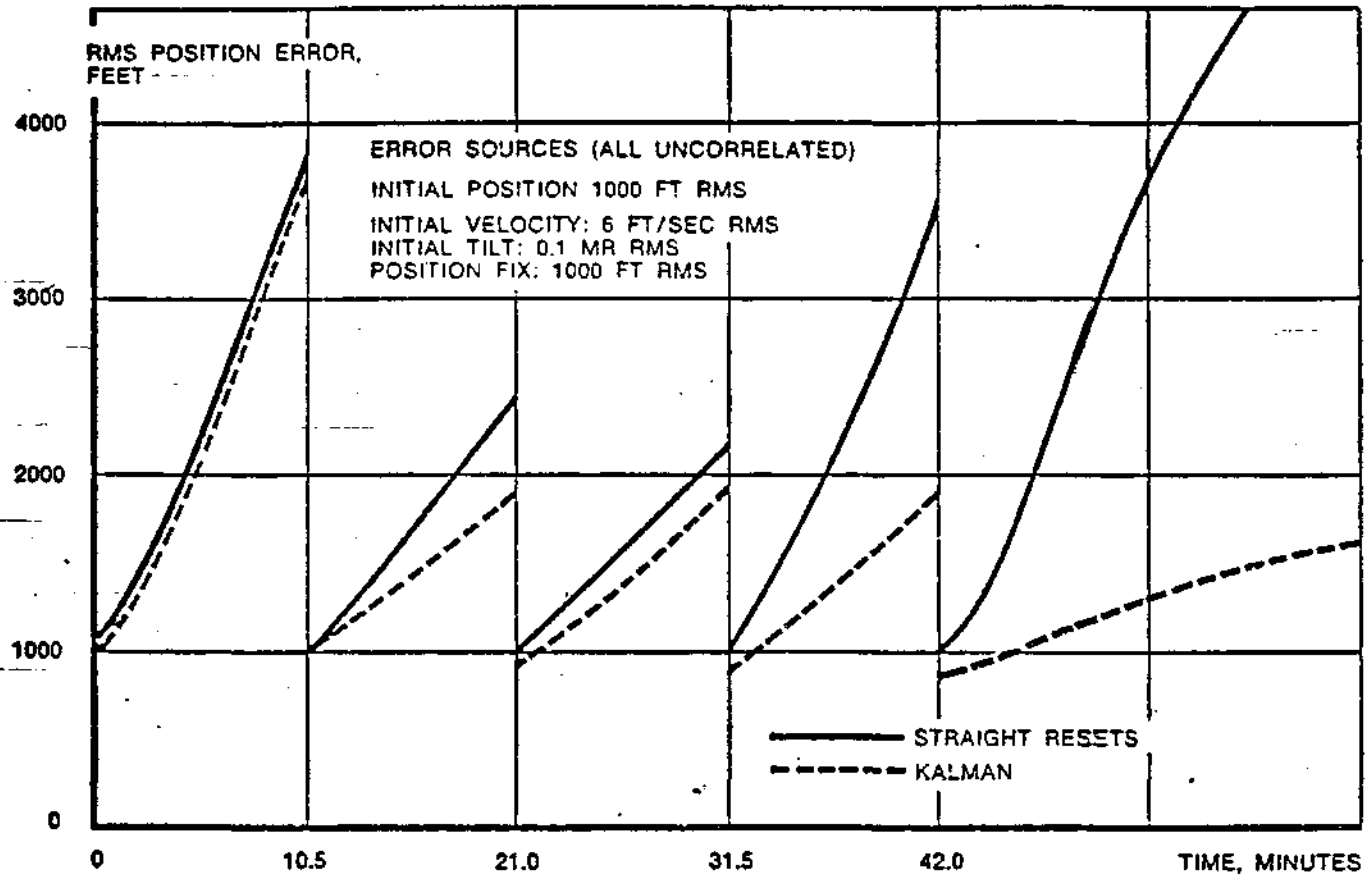


Figure 4. Comparison of Straight Position Resets with Kalman Resets

Our example will consider two cases. Case 1 involves simply correcting the inertial navigator position error with the external position-fix data. Physically, this might just mean adjusting the position display to agree with the externally indicated position. The velocity and tilt errors are unaffected by this procedure, and will continue to propagate away, unabated.

Detailed analysis of this case would do nothing to promote better understanding of Kalman and is, hence, omitted here. Position error as a function of time is presented in Figure 4 to provide a basis for comparison with the Kalman case, which is also shown in the same figure.

Case 2, employing the Kalman approach, will be treated in some detail. In order to deal with the dynamic, multi-dimensional aspects of the problem, the concepts of "state vector," "transition matrix," "measurement matrix," and "covariance matrix" will be introduced. It is easily verified that the position, velocity, and tilt errors as functions of time are given by

$$\begin{aligned}\delta x(t) &= \delta x(0) + \frac{s}{\omega}\delta\dot{x}(0) - R(1-c)\phi_y(0)\\ \delta\dot{x}(t) &= c\delta\dot{x}(0) - R\omega s\phi_y(0) \qquad (2)\\ \phi_y(t) &= \frac{s}{R\omega}\delta\dot{x}(0) + c\phi_y(0)\end{aligned}$$

in which $s = \sin\omega t$ and $c = \cos\omega t$, and $\omega = \sqrt{g/R}$. This can be written in matrix notation as

$$\begin{bmatrix}\delta x(t)\\ \delta\dot{x}(t)\\ \phi_y(t)\end{bmatrix} = \begin{bmatrix}1 & \frac{s}{\omega} & -R(1-c)\\ 0 & c & -R\omega s\\ 0 & \frac{s}{R\omega} & c\end{bmatrix}\begin{bmatrix}\delta x(0)\\ \delta\dot{x}(0)\\ \phi_y(0)\end{bmatrix} \qquad (3)$$

By a change of notation, (3) can be written

$$x(t) = \Phi(t)x(0) \qquad (4)$$

where the symbols of (4) denote the corresponding matrices in (3). The quantity $x(t)$ is called the system

"state vector." It is nothing more nor less than a convenient notational form. Its use results in neat equations such as (4), which aid and assist comprehension, and avoids messy equations such as (2), or, worse yet, complicated equations involving double and triple summations. The dynamics of the system are represented by $\Phi(t)$, which is called the "transition matrix." As will be seen, $\Phi(t)$ is used to a considerable extent in the Kalman filter.

In the resistor example, the ohmmeter measurement y was related to the resistance x by the relation

$$y = Mx + \epsilon \tag{5}$$

where M was the meter scale factor (M = 1 in the resistor example). In multi-dimensional problems, it is convenient to have a similar matrix equation, in which the measurement is related to the system state. In the present example, the measurement y is a position fix. It is equal to the actual position error δx plus a position fix error ϵ:

$$y = \delta x + \epsilon \tag{6}$$

Recall from above that

$$x = \begin{bmatrix} \delta x \\ \delta \dot{x} \\ \phi_y \end{bmatrix}$$

We can obtain the relationship (6) in the format of (5) by writing

$$y = (1\,0\,0)\begin{bmatrix} \delta x \\ \delta \dot{x} \\ \phi_y \end{bmatrix} + \epsilon \tag{7}$$

One can readily see that (7) is identical with (6).

By defining

$$M = (1\,0\,0) \tag{8}$$

equation (7) can be written in matrix notation:

$$y = Mx + \epsilon$$

M is called the "measurement matrix." It simply denotes the part or the component of the state vector x that is being measured.

In the bucketful-of-resistors example, it was seen that a fundamental part of the filter was the mean square error in the value of the resistors, namely (1 ohm)^2. Similarly, in multi-dimensional dynamic systems, such a quantity is required to compute the optimum weighting coefficients. This quantity is a matrix called the "covariance matrix." For our example, it is made up thus:

$$P_x = \begin{bmatrix} \text{Mean square position error} & \text{Cross-correlation between position and velocity errors} & \text{Cross-correlation between position and tilt errors} \\ \text{Cross-correlation between position and velocity errors} & \text{Mean square velocity error} & \text{Cross-correlation between velocity and tilt errors} \\ \text{Cross-correlation between position and tilt errors} & \text{Cross-correlation between velocity and tilt errors} & \text{Mean square tilt error} \end{bmatrix}$$

Initially, it was assumed that the mean-square position, velocity, and tilt errors were (1,000 ft)^2, (6 ft/sec)^2, and (0.1 mr)^2, respectively—all uncorrelated. Hence, the initial value of P_x is given by

$$P_x = \begin{bmatrix} (1000)^2 & 0 & 0 \\ 0 & (6)^2 & 0 \\ 0 & 0 & (0.0001)^2 \end{bmatrix} \tag{9}$$

As with the resistor example, the value of P_x must be known ahead of time. It can be determined from physical tests on similar systems, by error analysis, or, if need be, by best engineering judgment.

Now the covariance matrix P_x is not constant with time. The initial velocity error will start propagating into position and tilt error, etc., according to equations (2) or (3). It can easily be shown that the covariance matrix at some time t can be determined from $P_x(0)$ by

$$P_x(t) = \Phi(t)\,P_x(0)\,\Phi(t)^T \tag{10}$$

$\Phi(t)$ was discussed above and defined in equations (3) and (4). (The symbol $\Phi(t)^T$ means "the transpose of $\Phi(t)$" and simply denotes the exchange of the rows and columns of $\Phi(t)$. Transposition appears frequently in error analysis equations in matrix form, but can be ignored as far as understanding the gist of the equations.) Equation (10), while not derived here, certainly appears to be reasonable. In other words, if Φ propagates x, it is reasonable to expect that Φ "squared" would propagate the covariance of x. Consideration of a simple scalar example would illustrate this.

It must be assumed that, in addition to the inertial navigation hardware, there is a digital computer to perform the calculations described below. The computer is initialized at the initial value of P_x. At t = 10.5 sec, the first checkpoint is obtained. At this time, the computer will use the checkpoint data to make an estimate of the position, velocity, and tilt errors in the system. In making the estimate, the computer will consider the relative accuracy of the position fix data and the inertial navigation system data. To do this, the computer requires a current measure of the accuracy of the inertial navigation system at t = 10.5—namely, the covariance. The computer obtains this by extrapolating $P_x(0)$, using equation (10) or its equivalent.

$$P_x(t) = \begin{bmatrix} 1 & \frac{s}{\omega} & -R(1-c) \\ 0 & c & -R\omega s \\ 0 & \frac{s}{R\omega} & c \end{bmatrix} \begin{bmatrix} (1000)^2 & 0 & 0 \\ 0 & (6)^2 & 0 \\ 0 & 0 & (10^{-4})^2 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ \frac{s}{\omega} & c & \frac{s}{R\omega} \\ -R(1-c) & -R\omega s & c \end{bmatrix}$$

$$= \begin{bmatrix} 0.129^{15} & 0.154^{13} & 0.537^{10} \\ 0.154^{13} & 0.211^{12} & 0.597^{7-3} \\ 0.537^{10} & 0.597^{7-3} & 0.339^{7-7} \end{bmatrix} \tag{11}$$

The diagonal elements are the mean square position, velocity, and tilt errors, respectively. The off-diagonal elements are cross-correlations between these same three quantities.

It is instructive to examine the detailed expressions for a couple of these elements, just to see that equation (10) gives reasonable results. For example, the upper left-hand element of (11) is mean square position error and is given by

Mean square position error at t = 10.5:

$$= (1000)^2 + \left(\frac{s}{\omega}\right)^2 (6)^2 + R^2(1-c)^2(10^{-4})^2$$

$$= 0.129^{15} \text{ ft}^2$$

$$= 3{,}590 \text{ ft RMS}$$

By examining this expression, we can see the initial position error; the effects of the initial velocity error, which propagates into position as a sine s/ω; and of the initial tilt error, which propagates into position as a one minus cosine, $-R(1-c)$. Had the initial conditions been correlated, as is often the case, the expression would have been more complicated. It would have had terms involving the cross-correlations of the initial errors, indicating that the effects of these errors tended to add or cancel.

The element in the top row, second column of (11) is the cross-correlation between position error and velocity error. When written out in detail it is

Cross-correlation between velocity and position errors at $t = 10.5$:

$$= \frac{s}{\omega} c(6)^2 + [-R(1-c)][-R\omega s](10^{-4})^2$$

$$= 0.154^{45} \text{ ft ft/sec}$$

The first term is due to the initial velocity error, which propagates as a sine into position s/ω, but at the same time diminishes as a cosine c. The second term is due to the initial tilt error, which propagates simultaneously into position and velocity error as $-R(1-c)$ and $-R\omega s$, respectively, and hence contributes the cross-correlation as shown. Again, if the initial errors had been correlated, the expressions would have been more complicated, but the idea is the same.

In the resistor example, the data y was weighted by a coefficient b given by

$$b = \frac{M\sigma_x^2}{M^2\sigma_x^2 + \sigma_\epsilon^2} \qquad (12)$$

The same thing will be done here, except that the coefficient is a matrix given by

$$b = P_x M^T (M P_x M^T + P_\epsilon)^{-1} \qquad (13)$$

In equation (13), P_x is the covariance of x, and is analogous to σ_x^2 in (12). M is the measurement matrix, and is analogous to the scale factor M in (12). P_ϵ is the covariance of the position fix error, and is analogous to σ_ϵ^2 in equation (12). The superscript $^{-1}$ denotes matrix inversion, and is analogous to division (and in this example it will be seen that it *is* division). As remarked above, the transposition superscript T can be ignored as far as understanding the gist of these equations is concerned. After taking into account the idiosyncrasies in notation, it can be seen that (13) is very similar in appearance to (12). But, more importantly, *(13) is identical in concept to (12).* All the discussion regarding b in the resistor example applies to this example. If one understands the gist of (12), then (13) is no longer a mysterious, unintelligible mass of matrices.

Equation (13) will be given a more detailed examination. In doing so, it will be convenient to write P_x and P_ϵ in terms of their elements.

Using the definition of M and this notation for P_x and P_ϵ, equation (13) can be written out

$$P_x = \begin{bmatrix} P_{11} & P_{12} & P_{13} \\ P_{21} & P_{22} & P_{23} \\ P_{31} & P_{32} & P_{33} \end{bmatrix}$$

$$P_\epsilon = [P_{\epsilon\epsilon}]$$

where the numerical values of the elements of P_x were computed in (11), and $P_{\epsilon\epsilon}$ is simply $(1{,}000 \text{ ft})^2$.

Using the definition of M and this notation for P_x and P_ϵ, equation (13) can be written out:

$$\mathbf{b} = \mathbf{P}_x\mathbf{M}^T(\mathbf{M}\mathbf{P}_x\mathbf{M}^T + P_{\epsilon\epsilon})^{-1}$$

$$= \begin{bmatrix} P_{11} & P_{12} & P_{13} \\ P_{21} & P_{22} & P_{23} \\ P_{31} & P_{32} & P_{33} \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \left[\begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} P_{11} & P_{12} & P_{13} \\ P_{21} & P_{22} & P_{23} \\ P_{31} & P_{32} & P_{33} \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + [P_{\epsilon\epsilon}] \right]^{-1}$$

$$= \begin{bmatrix} P_{11} \\ P_{21} \\ P_{31} \end{bmatrix} [P_{11} + P_{\epsilon\epsilon}]^{-1} \qquad (14)$$

$$= \begin{bmatrix} \dfrac{P_{11}}{P_{11} + P_{\epsilon\epsilon}} \\ \dfrac{P_{21}}{P_{11} + P_{\epsilon\epsilon}} \\ \dfrac{P_{31}}{P_{11} + P_{\epsilon\epsilon}} \end{bmatrix} = \begin{bmatrix} 0.93^{10} \\ 0.11^{10-2} \\ 0.39^{10-7} \end{bmatrix}$$

Matrix inversion in this case involved only scalars $P_{11} = 0.129^{18}$ and $P_{\epsilon\epsilon} = 0.01^{18}$. Hence, inversion was in reality division by 0.139^{18}.

Before discussing a little of the significance of (14), let us go through the mechanics of making the Kalman estimate. Suppose the position fix indicated a position error of −4,000 ft. Then, the computer makes an estimate of position, velocity, and tilt error

$$\hat{\mathbf{x}} = \mathbf{b}y$$

$$= \begin{bmatrix} 0.93 \\ 0.11^{10-2} \\ 0.39^{10-7} \end{bmatrix} (-4000)$$

$$= \begin{bmatrix} -3710 \text{ ft} \\ -4.4 \text{ ft/sec} \\ -0.154 \text{ mr} \end{bmatrix}$$

In the resistor example, it was necessary to account for the nominal value, or mean value before a measurement was made, of 100 ohms. In this example, the mean errors are zero (although, of course, the mean square errors are not zero).

The first element of b is given by

$$b_1 = \frac{P_{11}}{P_{11} + P_{\epsilon\epsilon}}$$

in which P_{11} is the mean square position error at $t = 10.5$ and $P_{\epsilon\epsilon}$ is the mean square position fix measurement error. Hence b_1 is identical in form and meaning to the weighting coefficient in the resistor example. This reinforces the statement that the matrix equation (13) is identical in concept to (12). All the discussion of the resistor example carries over completely here and does not need to be repeated.

The second element of b is given by

$$b_2 = \frac{P_{21}}{P_{11} + P_{\epsilon\epsilon}} \qquad (15)$$

and involves the cross-correlation P_{21} between position error and velocity error; b_2 is used to weight position error to yield an estimate of velocity error.

If the reader is still somewhat confused as to how the Kalman filter can simultaneously estimate three quantities—namely, position, velocity and tilt—from a noisy measurement of only one quantity, perhaps the following discussion will help.

Suppose that in a certain area annual rainfall, the height of corn, and farmer's annual income are observed to be related; they all tend to increase or decrease together. Then, by observing only one of them, it is clearly possible to get an estimate of the

other two, using empirically derived relationships. If there were only one dominant causal factor, say annual rainfall, then this estimate might be a pretty good one. If other independent random factors existed, such as market conditions, then the estimate might be less accurate.

In an analogous manner, the position, velocity, and tilt errors can be simultaneously estimated from a single data point. Instead of an empirical relationship such as the one mentioned in the paragraph above, the proportionality factor b is computed using the covariance. The factor b_2 has the units of, and can roughly be interpreted as, "velocity error per unit observed position error."

After making the optimum estimate of position, velocity, and tilt, the mean square errors will be reduced. In the resistor example, the error in estimating the resistance was given by

$$\sigma_{\hat{x}}^2 = (1 - bM)\sigma_x^2$$

In the multi-dimensional case, the corresponding formula is

$$\mathbf{P}_{\hat{x}} = (\mathbf{I} - \mathbf{bM})\mathbf{P}_x$$

Expanding this equation in terms of its elements

$$\mathbf{P}_{\hat{x}} = \left[\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \begin{bmatrix} \dfrac{P_{11}}{P_{11}+P_{ee}} \\ \dfrac{P_{21}}{P_{11}+P_{ee}} \\ \dfrac{P_{31}}{P_{11}+P_{ee}} \end{bmatrix} (1\ 0\ 0)\right] \begin{bmatrix} P_{11} & P_{12} & P_{13} \\ P_{21} & P_{22} & P_{23} \\ P_{31} & P_{32} & P_{33} \end{bmatrix}$$

$$= \begin{bmatrix} 1 - \dfrac{P_{11}}{P_{11}+P_{ee}} & 0 & 0 \\ -\dfrac{P_{21}}{P_{11}+P_{ee}} & 1 & 0 \\ -\dfrac{P_{31}}{P_{11}+P_{ee}} & 0 & 1 \end{bmatrix} \begin{bmatrix} P_{11} & P_{12} & P_{13} \\ P_{21} & P_{22} & P_{23} \\ P_{31} & P_{32} & P_{33} \end{bmatrix} \tag{16}$$

$$= \begin{bmatrix} \dfrac{P_{11}P_{ee}}{P_{11}+P_{ee}} & P_{12} - \dfrac{P_{12}P_{11}}{P_{11}+P_{ee}} & P_{13} - \dfrac{P_{13}P_{11}}{P_{11}+P_{ee}} \\ P_{12} - \dfrac{P_{12}P_{11}}{P_{11}+P_{ee}} & P_{22} - \dfrac{P_{12}^2}{P_{11}+P_{ee}} & P_{23} - \dfrac{P_{12}P_{23}}{P_{11}+P_{ee}} \\ P_{13} - \dfrac{P_{13}P_{11}}{P_{11}+P_{ee}} & P_{23} - \dfrac{P_{12}P_{23}}{P_{11}+P_{ee}} & P_{33} - \dfrac{P_{13}^2}{P_{11}+P_{ee}} \end{bmatrix}$$

The diagonal elements of equation (16) are the mean square errors in estimating position, velocity

and tilt. The numerical values of the P's can be obtained from equation (11). Using these values, the following table can be constructed.

Quantity	*RMS Error before Optimum Estimate*	*RMS Error after Optimum Estimate*
Position	3,600 ft	960 ft
Velocity	4.59 ft/sec	1.97 ft/sec
Tilt	0.184 mr	0.114 mr

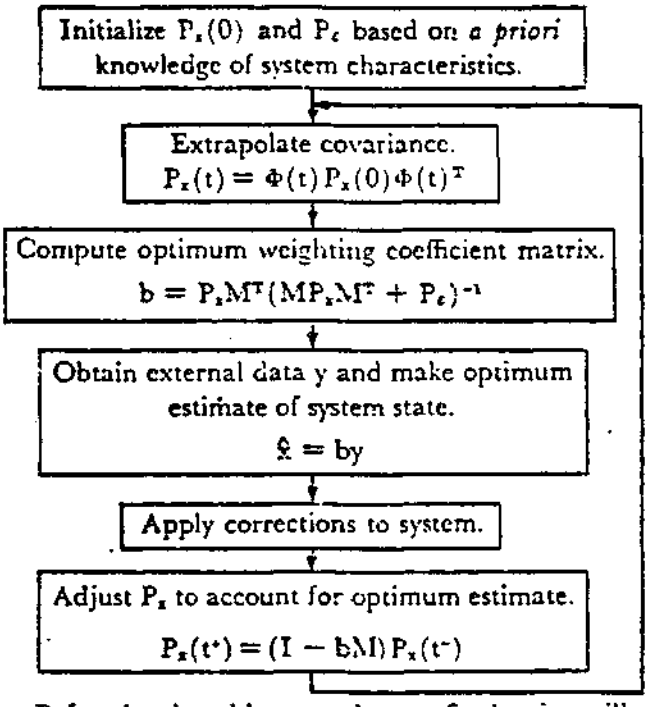
There is a considerable improvement not only in position, which was estimated slightly more accurately than the 1,000 ft RMS position fix, but also in velocity and tilt. This is because there is strong correlation between position error and velocity error, and between position error and tilt error; that is, P_{12} and P_{13} are large. A more detailed analysis reveals that initial velocity error is the dominant contributor to the RMS position, velocity, and tilt errors existing at $t = 10.5$.

Error Source	*Effect on Errors at t = 10.5* *Position*	*Velocity*	*Tilt*
Initial position, 1000 ft RMS	1000 ft	0 ft/sec	0 mr
Initial velocity, 6 ft RMS	3400	4.23	0.17
Initial tilt, 0.1 mr	600	1.73	0.07
RSS Totals	3600	4.59	0.184

The computer, based on the covariance matrix it has updated, knows that the velocity and tilt errors in the system are both roughly proportional to the observed position error and can make a relatively good estimate of them.

After the computer makes the optimum estimate, it can update the covariance as indicated in equation (16). It is now ready to extrapolate the covariance as in equation (11), until the next position fix is obtained. Again, as in the resistor example, a cyclic or recursive process can be seen to evolve.

The question arises—what should be done with the optimum estimates? The computer has two options. One is to do the obvious: Adjust the position and velocity registers and physically "torque out" the estimated tilt. The other option is to leave the Schuler loop alone, and to compute corrections to the inertial navigation system outputs as time progresses. But, this involves extrapolating the corrections using equation (3), and is an avoidable computational complexity. By choosing the former option, the mean value of the system errors is always zero, and the Kalman equations are simplified. They are summarized as follows:



Before leaving this example, one final point will be made to compare, once again, the maximum likelihood estimation with an ordinary least squares curve fit. Suppose that, shortly after $t = 0$, two position fixes are obtained with random errors of 1,000 ft RMS. Clearly, if we attempt to make a least squares curve fit of an 84-minute sine and cosine to these closely spaced, relatively inaccurate data points, very large errors would result. But, the Kalman filter recognizes this problem automatically. When the covariance matrix is updated, the cross-correlation between position and velocity is found to be very small:

$$P_{12} = \frac{s}{\omega} c(6)^2 + [-R(1 - c)][-R\omega s](10^{-4})^2 \approx 0$$

This is due to the $\sin \omega t$ and $(1 - \cos \omega t)$ factors, which for short interval t are approximately zero. The resulting b matrix is

$$b = \begin{bmatrix} \dfrac{P_{11}}{P_{11} + P_{\epsilon\epsilon}} \\ \dfrac{P_{12}}{P_{11} + P_{\epsilon\epsilon}} \\ \dfrac{P_{13}}{P_{11} + P_{\epsilon\epsilon}} \end{bmatrix} \approx \begin{bmatrix} \frac{1}{2} \\ 0 \\ 0 \end{bmatrix}$$

With the b matrix so computed, the computer will make a position correction equal to ½ the observed position fix data, and will make almost zero corrections to tilt and velocity. The RMS position error after this procedure will be reduced by $1/\sqrt{2}$, and RMS velocity and tilt errors will be unaffected.

In general, when properly mechanized, the Kalman filter will not introduce into a system errors which are larger—statistically speaking—than the errors which existed prior to making the estimate. This is true regardless of how poor the reference data are. It is definitely not true of least squares, as was illustrated in the bucketful-of-resistors example.

SOME SALIENT GENERAL FACTS REGARDING KALMAN FILTERS

1. Kalman techniques apply in a practical sense principally to linear systems.
2. The Kalman filter requires use of the covariance matrix with calculations. In effect then, the computer has a capability for real-time on-board *error analysis*. In fact, this error analysis capability uses somewhat sophisticated techniques which were not even available several years ago
3. The Kalman filter accepts various external data and makes corrections to the system "state," which in the case of inertial navigation systems might include position, velocity, tilt, gyro bias, azimuth error, etc. This makes it, in effect, an alinement process. In fact, there is no essential *difference between the navigation* and alinement mechanizations when using a Kalman filter, and the various modes such as "coarse aline," "fine aline," and "navigate" are all done away with as far as the filter is concerned.
4. We saw in the example above that the measurement matrix was given as

$$M = (1\ 0\ 0)$$

when the external data was a position fix. In case reference velocity had been available, the matrix would have been

$$M = (0\ 1\ 0)$$

In other words, by simply using the appropriate M matrix the computer can accept any form of external data that are related to the state variables. All the equations remain otherwise the same.

5. The Kalman filter can make use of human judgment. For example, we saw that the computer made use of the mean square position-fix error in computing the filter coefficients. Suppose that an airplane pilot were involved in the position-fix observation. It is possible he would have much higher confidence in some fixes than in others, say with a sighting of the airfield tower from 500 feet compared to a landmark observation from 15,000 feet. If the control panel were so arranged, it would be possible for him to indicate a rough confidence level in the fix when he entered the data into the computer. The computer might use a preset of table of RMS fix errors, corresponding to whether the pilot indicated the fix was "excellent," "good," or "fair."
6. Perhaps the biggest computational problem involves updating the covariance matrix. In the example, it was stated that it is updated by the formula

$$P_x(t) = \Phi(t) P_x(0) \Phi(t)^T$$

This would be a relatively simple calculation if Φ were available. But, Φ is hard to obtain when the system is time-varying, as is the case with an aircraft. The direct method is to numerically integrate the dynamical equations describing the system, assuming unity initial conditions. If the system is of nth order, then this involves integrating n simultaneous differential equations. This uses up much computer time and space, and gives rise to the possibility of round-off errors.

Another approach is to update the covariance by steps, by repeated application of the equation above. If the time increment is small enough, then an approximate expression for the transition matrix can be obtained analytically by assuming the parameters are constant in that interval. For a specific example, assume the system is an aircraft inertial navigation system. Then, if the time interval is short enough, we can assume that all the parameters, such as latitude, are constant, and the transition matrix takes on a form similar to equation (3) in the example. (The aircraft's maneuvers can be accounted for in a short time interval by accumulating velocity pulses from the accelerometer and

coupling azimuth error into velocity error through the velocity increment.) But this procedure replaces the problems encountered above with the problems of many matrix multiplications, again using up computer time and space, and generating the distinct possibility of round-off errors. If the round-off errors are large enough, some of the elements in the covariance could possibly be of the wrong sign, which might create an unstable loop.

There are other approaches to updating the transition matrix. But, whatever the technique, updating the covariance matrix remains a difficult problem at best.

7. Another problem of appreciable magnitude is that of properly modeling the system. This problem takes on several aspects. These are discussed below.

   a. There is a tradeoff between complexity of the system model versus computer complexity versus accuracy. As a specific example, consider the dynamics of an inertial navigation system. It is possible to derive an extremely complex model of such a system, considering only the linear aspects. One could include the Schuler loop, 24-hour effects, various instrument parameters and sensitivities, including instrument biases, scale factors, misalinements, random effects, trends, g and g^2 sensitivities, various servo phenomena, random driving functions such as gravity anomalies, phase of the moon, complex modeling of the reference data, etc., etc. Clearly, it is impractical to include this model in the on-board computer. It is necessary then to make tradeoff studies to get the right balance between improvement in system accuracy and available computer capability.

   b. Another difficult problem arises when the exact model of the system is not even well known.

      An example might occur in the determination of the orbit of a satellite. Suppose the computer thinks the only errors are uncertainties in initial position and velocity of the satellite. After obtaining sufficient external data, the computer thinks it has trimmed up the orbit *à la* Kalman, and the b coefficients approach zero. But, if there are [illegible] and unmodeled forces such as solar pressure the actual orbit of the satellite may start to diverge from the computed orbit. The external data will now be rejected and the orbit will go uncorrected.

      By judicious modeling techniques such problems can be reduced or eliminated.

   c. The problem of erroneous statistics was discussed in connection with the bucket of resistors example. There, it was seen that if the resistors were much worse than assumed in the model, poor estimates were achieved. If the ensemble of systems is the same as that assumed in the model, then the Kalman procedure will obtain optimum performance out of this ensemble. But the Kalman filter cannot make good systems out of bad ones.

8. Implementation of the Kalman filter can impose—depending on the application—heavy demands on the computer. The practical engineer will explore various schemes which can provide the essential benefits of the Kalman approach, but which will provide relief with regard to computer time and space. Besides the modeling approaches discussed above, other compromises with the full-blown theory should be explored:

   a. A body of theory has been developed regarding what are called "sub-optimal filters." These are Kalman-type filters, except that certain portions of the system state are assumed to be uncoupled from the remainder. For example, it might be assumed that the x and y channels of an inertial navigation system are uncoupled: this would ignore 24-hour and azimuth channel effects. It turns out that this type of approach yields two or more Kalman-type problems for the computer to solve, except that the sum of the parts is less than the whole. In other words, it is easier for a computer to deal simultaneously with two third-order problems than with just one sixth-order problem.

   b. When data come in at a high rate, it becomes impossible for the computer to process the data with the complete Kalman equations. One approach is to prefilter the data by analog or digital means and to have the computer deal with the prefiltered data at a much

slower rate. Another approach is to have the Kalman filter ignore the prefiltered data altogether, and to have the high-speed data enter the system by some more conventional means, and the slower-speed data processed in the Kalman manner.

c. In general, when the data vector y consists of n components, theory says that the Kalman filter has to invert an $n \times n$ matrix to derive the filter matrix b. But matrix inversion often is a very difficult process to do in real time, even for small n, such as 2 or 3. One device to avoid this problem, when the input data rate is low, is to consider the components of the data vector to be coming in singly. For example, suppose the data consist of two components, latitude and longitude. The computer could be instructed to process just the latitude data first, and it would have only to "invert" a 1×1 matrix which, translated, means it only has to perform a simple division. Having processed the latitude data, it can accordingly adjust the covariance matrix to account for them, and then process longitude.

d. When the mission is known beforehand, consideration should be given to precalculating the b coefficients. This can yield considerable computer time and, perhaps, space savings.

In short, the full-blown, doctrinaire approach to Kalman could well lead to impractical demands on the computer. An inventive, shortcut, hybrid approach might yield the benefits of Kalman without overloading the computer.

APPENDIX

The purpose of this appendix is to provide an abbreviated tutorial derivation of the equations appearing in the two examples cited in the main body of the paper. This, in turn, should reveal the principles underlying maximum likelihood estimation.

Inasmuch as the purpose of the paper is tutorial, and since general derivations are available in numerous other places, general equations will not be derived. Some difficulties in the manipulation of matrices is avoided in this way and, at the same time, little is sacrificed with respect to understanding the concepts. To assist the reader whose knowledge of probability notation may be hazy, the derivation given below makes use of specific references to the bucketful-of-resistors example.

We start out by defining a criterion for optimality. Given some quantity x (e.g., unknown value of a certain resistor) which we want to estimate, the symbol $\hat{x}$ is assigned to denote this estimate. Suppose we have a measurement y (e.g., an ohmmeter reading). The error in estimation is then given by $x - \hat{x}$. This error has associated with it some loss which we arbitrarily define as

$$Q = (x - \hat{x})^2 K^2$$

where

Q = loss (e.g., in units of dollars) associated with the error in estimating the value of the first resistor

$(x - \hat{x})^2$ = error squared in estimating the resistance

K^2 = positive constant (e.g., in units of dollars/ohms^2) which converts squared estimation error into units of loss

Now, we would like somehow to minimize Q, given the measurement y. But there is clearly no way to guarantee this, because of the random nature of the process; Q has a minimum value of 0, and to guarantee this would imply we could guarantee a perfect estimate of x. We must resort to a statistical definition of optimum. If the statistics of x and y are known, then the expected or mean value of Q can be determined. $\hat{x}$ is defined as being optimum if, given the data y, the expected value of Q is minimized. Written in symbols:

Choose $\hat{x}$ so that

$$Q = E[(x - \hat{x})^2 K^2 | y] \text{ is minimized}$$

(The symbol $E[A|B]$ denotes the expected value of A, given B.) This is our criterion for optimum estimation.

We can determine the minimum expected value of the loss by setting its derivative with respect to $\hat{x} = 0$:

$$\frac{\partial E[Q]}{\partial \hat{x}} = 2E[(x - \hat{x})K^2 | y]$$

$$= 2K^2 E[x|y] - 2K^2 E[\hat{x}|y]$$

The symbol $E[\hat{x}|y]$ denotes "the expected value of our estimate, given the data y." But, we are free to choose $\hat{x}$ in any manner we choose, and the expected value of $\hat{x}$ is simply whatever we choose it to be. In other words,

$$E[\hat{x}|y] = \hat{x}$$

Putting this into the expression for the derivative and setting the result equal to zero yields the following expression for the optimum estimate

$$\hat{x} \text{ optimum} = E[x|y] \qquad \text{(A-1)}$$

This is a very general and important result. To retrace the steps above:

1. Given a weighted, squared estimation error type of loss function
2. Given *a priori* statisics on the quantity being estimated x and the measurement y
3. Given some measurement data y
4. Then the expected value of x given these data can be computed (just how will be explained below)
5. The optimum estimate $\hat{x}$ is equal to this expected value.

It can be noted that the weighting constant K^2 drops out and does not enter into the determination of $\hat{x}$. This is perhaps intuitively obvious for this simple scalar case. It is also true for the multi-dimensional case in which K^2 is replaced by a weighting matrix.

Now suppose the quantities x and y are gaussian. If this is the case, then we can obtain a special form for $\hat{x}$. This form is the same as was shown in the examples.

Recall that the general form of a gaussian probability density function is as shown below:

$$p(r) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma_r} e^{-\frac{(r-m_r)^2}{2\sigma_r^2}} \qquad \text{(A-2)}$$

where

r = gaussian-distributed random variable

σ_r = standard deviation about the mean

m_r = mean value of r

Then let the probability density function of x be as follows:

$$p(x) = \frac{1}{\sqrt{2\pi}\,\sigma_x} e^{-\frac{(x-m_x)^2}{2\sigma_x^2}} \qquad \text{(A-3)}$$

where

σ_x = standard deviation of x (= 1 ohm in example)

m_x = mean value of x (= 100 ohms in example)

Let

$$y = Mx + \varepsilon \qquad \text{(A-4)}$$

where M = scale factor (= 1 in example)

and ε = measurement error

Assume that ε is gaussian with zero mean and independent of x:

$$p(\varepsilon) = \frac{1}{\sqrt{2\pi}\,\sigma_\varepsilon} e^{-\frac{\varepsilon^2}{2\sigma_\varepsilon^2}} \qquad \text{(A-5)}$$

where σ_ε = standard deviation of ε (= 3 in example).

Now, according to the general formula (A-1), we require an expression for $E[x|y]$. This can be obained if an expression for the conditional probability density function $p(x|y)$ can be obtained—by averaging $p(x|y)$, $E(x|y)$ results. But, what is $p(x|y)$? In words, it is the probability density function of x, given the data y. This is very much different from $p(x)$ with no data y. For example, suppose x is the height of all males in the U.S.A. The probability density function might have mean value 5 feet 11 inches and standard deviation of 3 inches. With no measurements available, we would estimate the height of a given male as 5 ft 11 in. The RMS error would be 3 in. But given the measurement y made to 0.2 in. RMS on the height of that male, the probability distribution of x is very much altered.

To determine $p(x|y)$, we can make use of Bayes' Rule. This is easily derived from the following statement:

> The probability that both x and y jointly occur is equal to the probability that y happens, times the conditional probability that y having happened, x will happen.

In symbols

$$p(x \text{ and } y) = p(y)p(x|y)$$

Obviously, x and y can be interchanged on the right:

$$p(x \text{ and } y) = p(x)p(y|x)$$

By combining these two equations, Bayes' Rule results:

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)} \qquad \text{(A-6)}$$

On the left is the quantity we desire to know. On the right, $p(x)$ is known from (A-3). The probability function of y can easily be obtained. The mean value of y is given by

$$E[y] = E[Mx + \varepsilon]$$

$$= E[Mx] + E[\varepsilon]$$

$$E[y] = M m_x$$

The standard deviation squared of y is given by

$$E[(y - m_y)^2] = E[Mx + \epsilon - Mm_x)^2]$$
$$= E[(M(x - m_x) + \epsilon)^2] \qquad (A\text{-}7)$$

or $$\sigma_y^2 = M^2\sigma_x^2 + \sigma_\epsilon^2$$

where the last step depends on the assumption that the ϵ and x are uncorrelated. We now have the mean value and standard deviation of the gaussian variable y. This is all that is required to write p(y):

$$p(y) = \frac{1}{\sqrt{2\pi}\,\sigma_y} e^{-\frac{(y - Mm_x)^2}{2\sigma_y^2}} \qquad (A\text{-}8)$$

where σ_y is given by A-7).

In the same way, the conditional probability density p[y|x] can be determined. If x is given and known, say x = 100.6732 ohms, then the probability that y has some value, say 101.0000 ohms, is just the probability that ϵ = 101.0000 − 100.6732. In other words, the probability density function p[y|x] is the probability density function $p(\epsilon = y - Mx)$. It has standard deviation σ_ϵ and mean value Mx. Putting this in the form of (A-2),

$$p(y|x) = \frac{1}{\sqrt{2\pi}\,\sigma_\epsilon} e^{-\frac{1}{2\sigma_\epsilon^2}(y - Mx)^2} \qquad (A\text{-}9)$$

Putting (A-3), (A-7), and (A-8) into Bayes' Rule gives

$$p(x|y) = \frac{1}{\sqrt{2\pi}} \frac{\sigma_y}{\sigma_x\sigma_\epsilon} \exp\left[-\frac{1}{2}\left(\frac{(x - m_x)^2}{\sigma_x^2} + \frac{(y - Mx)^2}{\sigma_\epsilon^2} - \frac{(y - Mm_x)^2}{\sigma_y^2}\right)\right] \qquad (A\text{-}10)$$

If the expression in brackets is multiplied out and terms are collected, a perfect square results:

$$\left[-\frac{1}{2}\left(\frac{(x - m_x)^2}{\sigma_x^2} + \frac{(y - Mx)^2}{\sigma_\epsilon^2} - \frac{(y - Mm_x)^2}{\sigma_y^2}\right)\right]$$
$$= -\frac{1}{2}\left[\frac{\sigma_y^2}{\sigma_\epsilon^2\sigma_x^2}x^2 - 2x\left(\frac{m_x}{\sigma_x^2} + \frac{My}{\sigma_\epsilon^2}\right)\right.$$
$$\left. + \frac{1}{\sigma_y^2}\left(\frac{M^2\sigma_x^2}{\sigma_\epsilon^2}y^2 + 2m_xMy + \frac{\sigma_\epsilon^2}{\sigma_x^2}m_x^2\right)\right] \qquad (A\text{-}11)$$
$$= -\frac{1}{2}\frac{\sigma_y^2}{\sigma_\epsilon^2\sigma_x^2}\left[x - \frac{1}{\sigma_y^2}(M\sigma_x^2y + \sigma_\epsilon^2m_x)\right]^2$$
$$= -\frac{1}{2}\frac{\sigma_y^2}{\sigma_\epsilon^2\sigma_x^2}\left[x - \left(m_x + \frac{M\sigma_x^2}{\sigma_y^2}(y - Mm_x)\right)\right]^2$$

In reducing (A-11), use is made of (A-7). If the factored expression (A-11) is substituted into (A-10), there results

$$p(x|y) = \frac{1}{\sqrt{2\pi}} \frac{\sigma_y}{\sigma_\epsilon\sigma_x} \exp\left[-\frac{1}{2}\frac{\sigma_y^2}{\sigma_\epsilon^2\sigma_x^2}\left(x - \left[m_x - \frac{M\sigma_x^2}{\sigma_y^2}(y - Mm_x)\right]\right)^2\right] \qquad (A\text{-}12)$$

(A-12) is of the form of the standard gaussian probability distribution (A-2). Comparing (A-12) with (A-2), the expected or mean value of $p(x|y)$ can be determined, which corresponds to m_r in (A-2). This is

$$E[x|y] = m_x - \frac{M\sigma_x^2}{\sigma_y^2}(y - Mm_x) \quad \text{(A-13)}$$

This is the optimum estimate of x, given the data y. By using (A-7), (A-13) can be rewritten

$$\hat{x}\ \text{optimum} = E[x|y]$$

$$= m_x - \frac{M\sigma_x^2}{M^2\sigma_x^2 + \sigma_z^2}(y - Mm_x) \quad \text{(A-14)}$$

The weighting coefficient b can be seen to be

$$b = \frac{M\sigma_x^2}{M^2\sigma_x^2 + \sigma_z^2}$$

which is the same as that used in the resistor example.

The expression (A-14) appears reasonable. m_x is the best estimate of x given no data. Mm_x is the reading we would get if x were indeed m_x. Hence, $y - Mm_x$ represents a sort of error signal. b weights that error signal to account for the relative precision of the measurement data.

One can now see how the term "maximum likelihood" arises. $p(x|y)$ is the probability density of x, given some data y; it might be termed a "likelihood function." Under the definition of optimality given above, the optimum estimate of x, for any type of statistics on x and y, gaussian or not, is $E[x|y]$. But, when gaussian statistics are assumed, $p(x|y)$ takes on the familiar bell-shaped form, and $E[x|y]$, is at the peak value, or point of maximum likelihood of that curve. For other types of statistics, this may or may not be the case. (A semantics problem arises in the use of the terms "least squares" and "maximum likelihood." These terms may not have the same meaning to all people. They were used in this paper, perhaps non-rigorously, to denote two estimation techniques. The important thing in this tutorial treatment is to make understandable to the reader the difference in the two estimation techniques, rather than to present a rigorous definition of terms.)

The mean square error in the estimate is given by

$$\sigma_{\hat{x}}^2 = E[(x - \hat{x})^2]$$

If the expression (A-14) is substituted for $\hat{x}$ and it is recalled that ϵ was assumed to be uncorrelated with x, then the expression for ϵ reduces to

$$\sigma_{\hat{x}}^2 = \frac{\sigma_z^2\sigma_x^2}{M\sigma_x^2 + \sigma_z^2} \quad \text{(A-15)}$$

This again is the expression used in the example.

One last remark concludes this abbreviated development of the Kalman filter equations. We have just seen the derivation of the equations for processing a single data point and making an optimum estimate. This optimum estimate is the expected or mean value of x, given the data y. If another data point is obtained, then x is the new expected value of x and will play the role of m_x in the new estimate. The variance of the estimate, given by (A-15), will play the role of σ_x^2 in the new estimate.

$$\hat{x}_2 = \hat{x}_1 + \frac{M\sigma_{\hat{x}_1}^2}{M^2\sigma_{\hat{x}_1}^2 + \sigma_z^2}(y - M\hat{x}_1)$$

Finally, if the system has dynamics, then the estimated state of the system will not "stay put" between estimates. The old estimate and covariance can be extrapolated to account for system dynamics, using the transition matrix described in the second example, or some equivalent process. Thus, having gone through both examples and the derivation of the basic equations for the maximum likelihood estimate, the reader should find the extension of these equations to their recursive form for dynamic systems fairly obvious.