DATA FUSION III: Estimation Theory

Date: February 23, 2006 **Time:** 5:00 – 7:30 PM
Instructor: Dr. James K Beard
Credits: 1 **Course Code:** 0901-501-05 **Registration Number:** 13460

Today's topics:

1 Definition of Estimation theory

Estimation theory is the science of solutions to the problem of estimating the value of unknown parameters from noisy measurements of related quantities. This includes the definition and analysis of estimation methods, their accuracy, and their use in practical problems.

A function of data is called a *statistic.* A function of the data which approximates an unknown parameter is an *estimator*. Thus, estimation theory is the design, analysis, evaluation, and use of estimators, which are statistics computed from available data.

2 Course Overview

We have only five sessions, and we must define the scope of our material carefully to provide a clear treatment of selected topics in that time. We wil

- Define estimation theory and its basic types.
- Define and explain the differences in quality and accuracy of the most common estimators.
- Show examples of estimators that you can use in your work.
- Show how to devise and evaluate estimators for new problems
- Show solutions to problems specific to data fusion issues

3 Accuracy and Confidence Limits

A *confidence limit* is a probability that is set as part of an experiment design as an *a priori* threshold on the significance of results of the experiment. For example, if we want to be 95% sure that our data is biased before we declare that our experiment has shown that it is biased, then we will devise an estimator for the bias, find the probability distribution function for the estimator under the condition that a bias is not present, and set thresholds at the 5% points in this distribution. Then, the results of an ensemble of experiments will declare the data unbiased 95% of the time when it is, indeed, unbiased.

Setting confidence limits requires that the requirements of the experiment be understood. The impact of incorrect interpretation of the experiment are part of the process of setting confidence limits.

4 Properties of Estimators

4.1 The Unbiased and Consistency Properties

An *unbiased estimator* is one whose ensemble mean is equal to the value of the unknown parameter.

A *consistent estimator* is one which approaches the value of the unknown parameter as the number of measurements increases without bound.

4.2 Sufficiency

A *sufficient estimator* is one that uses all the information in the available data that applies to estimating the value of the unknown parameters.

The *Neyman-Fisther Factorization Theorem* says that an estimator $\underline{\phi}(\underline{x})$ is sufficient if, and only if, the conditional probability density of the data given the states can be factored in this way:

$$p\left(\underline{y}\middle|\underline{x}\right) = g\left(\underline{x},\underline{\phi}\right)\cdot h\left(\underline{y}\right) \tag{4.1}$$

where *h(y)* is independent of $\underline{x}$. This is important in mathematical proofs of sufficiency. We will not have the time in this course to delve into the mathematics of estimation

theory to the extent that we will use this but you should remember it as background, and try to understand what it means in terms of the property of sufficiency.

4.3 Invariance

An estimator $\underline{\phi}(\underline{x})$ is *invariant* if, when applied to a nonlinear but single-valued function of the parameters $\underline{f}(\underline{x})$, the result is $\underline{f}\left(\underline{\phi}(\underline{x})\right)$. A sufficient estimator will always be invariant.

4.4 Efficiency

An estimator is *efficient* (or, *statistically efficient*) if the variance of the estimator is the smallest that is obtainable by any estimator. If an estimator is sufficient and unbiased, it will be efficient.

A term that I sometimes use that is not found in the literature as a standard term is *statistical efficiency*. Occasionally an *ad hoc* estimator that is commonly used has a variance that is demonstrably less than the Cramer-Rao bound, or that of an efficient estimator that is available. An example is averaging the absolute values of the errors to estimate standard deviation instead of averaging the squared errors to find the sample variance. I call the ratio of the variances the statistical efficiency of the estimator. When the statistical efficiency is 100%, the estimator is efficient in the classical usage of the term "efficient estimator."

5 Types of Estimators

5.1 Ad Hoc

Estimators that seem "correct" but that have no provenance in estimation theory are common. An example is estimating the number of fish of various types that exist in a lake. The technique that is normally used is to band or mark a small number and return them, then look at the number of fish that are caught of each type. For each type, the ratio of fish that are caught that are banded to those that are not banded is considered to be the best estimate of the number that were banded and thrown back and the rest in the lake.

5.2 Averages

Simply averaging noisy quantities to obtain an estimate of their ensemble average is a common example. This method is easily analyzed in terms of the accuracy of the estimate.

5.3 Least Squares

Least squares is a simple technique that was first used by Gauss to determine orbits of celestial bodies and the Earth from astronomical observations. An example we commonly use is a way to fit a polynomial to a set of data points. If we denote the set of *M* data points as

$$\left(x_i, y_i\right),\ i = 1 \ldots M \tag{5.1}$$

and wish to fit a polynomial $p_N(x)$ of degree N to those points, we have

$$\begin{aligned} \tilde{y}_i &= p_N\left(x_i\right) \\ &= \sum_{j=0}^{N-1} a_j \cdot x_i^j \end{aligned} \tag{5.2}$$

We define a cost function as the total squared error:

$$J = \sum_{i=1}^{M} \left(\tilde{y}_i - y_i\right)^2 \tag{5.3}$$

and find the polynomial coefficients that minimize the total squared error J. We can pose the problem in vectors and matrices and write the solution immediately. We collect the polynomial coefficients a_j, the powers of the independent variables x_i, and the dependent variables y_i as vectors:

$$\underline{a} = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_{N-1} \end{bmatrix} \tag{5.4}$$

$$\underline{x}_i = \begin{bmatrix} 1 \\ x_i \\ \vdots \\ x_i^{N-1} \end{bmatrix} \tag{5.5}$$

$$\underline{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_M \end{bmatrix} \tag{5.6}$$

The data equations become

$$\underline{\tilde{y}} = X \cdot \underline{a} \tag{5.7}$$

where the rows of the matrix X are the vectors $\underline{x}_i$:

$$X = \begin{bmatrix} \underline{x}_1^T \\ \underline{x}_2^T \\ \vdots \\ \underline{x}_M^T \end{bmatrix} \tag{5.8}$$

The error vector $\underline{y}_e$ is

$$\underline{y}_e = \underline{\tilde{y}} - \underline{y} \tag{5.9}$$

and the total squared error J is

$$\begin{aligned} J &= \underline{y}_e^T \cdot \underline{y}_e = \left(\tilde{\underline{y}} - \underline{y}\right)^T \cdot \left(\tilde{\underline{y}} - \underline{y}\right) \\ &= \left(X \cdot \underline{a} - \underline{y}\right)^T \cdot \left(X \cdot \underline{a} - \underline{y}\right) \end{aligned} \tag{5.10}$$

We find the value of $\underline{a}$ that minimizes the total squared error J with simple calculus. We take the gradient of J with respect to $\underline{a}$ to get N equations in N unknowns and solve:

$$\frac{\partial J}{\partial \underline{a}} = 2 \cdot X^T \cdot \left(X \cdot \underline{a} - \underline{y}\right) = \underline{0} \tag{5.11}$$

The solution is

$$\underline{a} = \left(X^T \cdot X\right)^{-1} \cdot X^T \cdot \underline{y} \tag{5.12}$$

The matrix $X^T \cdot X$ is very interesting and, because it occurs in this very common problem, it has been studied quite a bit. It is of the form

$$X^T \cdot X = \begin{bmatrix} M & \sum_{i=1}^{M} x_i & \sum_{i=1}^{M} x_i^2 & \cdots & \sum_{i=1}^{M} x_i^{N-1} \\ \sum_{i=1}^{M} x_i & \sum_{i=1}^{M} x_i^2 & \sum_{i=1}^{M} x_i^3 & \cdots & \sum_{i=1}^{M} x_i^{N} \\ \sum_{i=1}^{M} x_i^2 & \sum_{i=1}^{M} x_i^3 & \sum_{i=1}^{M} x_i^4 & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \\ \sum_{i=1}^{M} x_i^{N-1} & \sum_{i=1}^{M} x_i^{N} & \cdots & & \sum_{i=1}^{M} x_i^{2N-2} \end{bmatrix} \tag{5.13}$$

When $M=N$ then X is square and we can simply invert X to find a polynomial that passes through all the points; in this case, X is a Vandermonde matrix, which occurs also in other problems. The Vandermonde matrix is known to be ill-conditioned – that is, practical problems in accuracy are often encountered in inverting it. When $M>N$, we are doing a least-squares fit of data to a polynomial, which is a smoothing operation, not a polynomial fit. The matrix for this operation as given by (5.13) is less troublesome than the Vandermonde matrix, and problems are rarely encountered when $M>>N$.

Also instructive is the product $X \cdot \underline{y}$: that we use when $M>N$:

$$X^T \cdot \underline{y} = \begin{bmatrix} \sum_{i=1}^{M} y_i \\ \sum_{i=1}^{M} x_i \cdot y_i \\ \sum_{i=1}^{M} x_i^2 \cdot y_i \\ \vdots \\ \sum_{i=1}^{M} x_i^{N-1} \cdot y_i \end{bmatrix} \tag{5.14}$$

Note that these sums can be accumulated as data is accrued, so that only the partial sums need be kept in memory as the data is collected. Since the element at row *p* and column *q* of $X^T X$ is given by

$$\left(X^T \cdot X\right)_{p,q} = \sum_{i=1}^{M} x_i^{p+q-2} \tag{5.15}$$

This allows these terms to be accumulated as data is accrued also.

5.4 Maximum Likelihood

The method of maximum likelihood is based on a simple principle. In general, we can usually write the conditional probability density function of the data, given the states:

$$Likelihood\left(\underline{x}\right) = p\left(\underline{y}\middle|\underline{x}\right) \tag{5.16}$$

This is called the *likelihood equation* because at the time that R. A. Fisher devised it, the common term for the probability density function was the likelihood function. This method is simple to use because the likelihood function can be written easily whenever the probability density function of the errors in the measurements can be characterized, which is nearly all the time. Estimators found in this way are called *maximum likelihood estimators* (MLEs). We will present the method of maximum likelihood in greater detail later today.

5.5 Maximum a Posteriori

We can write a conditional probability distribution of the states $\underline{x}$, given the measurements $\underline{y}$, as a *posterior probability* of $\underline{x}$ given $\underline{y}$:

$$PP\left(\underline{x},\underline{y}\right) = p\left(\underline{x}\middle|\underline{y}\right) \tag{5.17}$$

If we find the states $\underline{x}$ that maximize this as a function of the observed data $\underline{y}$, this is the *maximum a posteriori* (MAP) estimator.

From Bayes' rule, we know that conditional probabilities are given in terms of joint probabilities according to the rule

$$p\left(\underline{x},\underline{y}\right) = p\left(\underline{x}\middle|\underline{y}\right)\cdot p\left(\underline{y}\right) = p\left(\underline{y}\middle|\underline{x}\right)\cdot p\left(\underline{x}\right) \tag{5.18}$$

where the probability densities of separate sets of variables are given by

$$\begin{aligned} p\left(\underline{x}\right) &= \int p\left(\underline{x},\underline{y}\right)\cdot d\underline{y} \\ p\left(\underline{y}\right) &= \int p\left(\underline{x},\underline{y}\right)\cdot d\underline{x} \end{aligned} \tag{5.19}$$

Thus the posterior probability function can be written in three forms:

$$PP\left(\underline{x},\underline{y}\right) = p\left(\underline{x}\middle|\underline{y}\right) = \frac{p\left(\underline{y}\middle|\underline{x}\right)\cdot p\left(\underline{x}\right)}{p\left(\underline{y}\right)} = \frac{p\left(\underline{x},\underline{y}\right)}{p\left(\underline{y}\right)} \tag{5.20}$$

Thus the MAP estimator can be applied whenever we can write any one of these.

When the *a priori* probability density function of $\underline{x}$ is unknown without other information, as is often the case, and $p(\underline{x})$ is taken as uniform, the MAP estimator reduces to the MLE.

5.6 Bayesian Mean

If we look at the probability distribution of the states $\underline{x}$ given the measurements $\underline{y}$, and we use it to compute the mean of the states given the measurements $\underline{y}$,

$$\hat{\underline{x}} = \int \underline{x} \cdot p\left(\underline{x} \middle| \underline{y}\right) \cdot d\underline{x} \tag{5.21}$$

is called the *Bayesian mean* estimator. This estimator can be shown to be minimum variance, and thus is sufficient, and often it produces an efficient estimator. In addition, it is unbiased. These properties make is a method of choice in mathematical developments but in real-world cases the difficulty in writing $p\left(\underline{x} \middle| \underline{y}\right)$ and then performing the integral make its use less practical than the MLE.

The MLE produces estimators that are the same as those produced as the Bayseian mean except in cases where nonlinearities exist between the measurements and the states. When this occurs, the MLE is biased while the Bayesian mean is not, while both are minimum variance estimators. The difficulty in removing the mean from the MLE is typically less than that in implementing a Bayesian mean estimator and produces an equivalent result.

6 The Method of Maximum Likelihood

6.1 Definition and properties

The method of maximum likelihood is based on a simple principle. In general, we can usually write the conditional probability density function of the data, given the states:

$$Likelihood\left(\underline{x}\right) = p\left(\underline{y} \middle| \underline{x}\right) \tag{6.1}$$

This is called the *likelihood equation* because at the time that R. A. Fisher devised it, the common term for the probability density function was the likelihood function.

The method of maximum likelihood is based on the idea that if you find the states $\underline{x}$ that maximize the value of the likelihood function for the observed values of the measurements $\underline{y}$, then you would have a "good" estimator. The importance of maximum likelihood estimators (MLEs) lies in the fact that these properties have been proven for it:

- MLEs are sufficient and consistent.
- MLEs are asymptotically unbiased.
- The errors in MLEs are asymptotically Gaussian.
- MLEs are asymptotically efficient.
- If an unbiased, efficient estimator exists, the MLE will be that estimator.

The MLE is useful in our work is because

- Nearly all common estimators can be shown to be found by the method of maximum likelihood, including the examples given
- The method of maximum likelihood is easily stated and applied for nearly all practical problems.

- When the measurement errors are distributed by a rule other than Gaussian random variables, the method of maximum likelihood can be easily applied without approximation.
- MLEs are usually simple to write, implement, and analyze.

6.2 The Likelihood Equations and the MLE

Most probability density functions are given primarily in terms of an exponential. Examples are Gaussian, Rayleigh, and Poisson. An exception is the Cauchy distribution. As such, we usually proceed with the log likelihood function,

$$L\left(\underline{x}\right) = \ln\left(p\left(\underline{y}\middle|\underline{x}\right)\right) \tag{6.2}$$

We then find the maximum of this function as the states $\underline{x}$ are varied through ordinary calculus. We take the gradient of the log likelihood function $L(\underline{x})$ and set it equal to zero to form the *likelihood equation*:

$$l\left(\hat{\underline{x}}\right) = \frac{\partial L\left(\underline{x}\right)}{\partial \underline{x}} = \underline{0} \tag{6.3}$$

This is N equations in N unknowns. Much is said by some that the generality of this equation makes it suspect, particularly in the fact that such an equation can have many roots. In practical cases this happens only in cases where the result is predictable and easily handled, such as in phase ambiguities of 2π radians. When this equation does have multiple solutions, then the solution or solutions with the largest value of the log likelihood function are the solution. If there are multiple such solutions, the ambiguous results are part of the solution to the estimation problem.

The solution to the likelihood equation (6.3) is the MLE. Now, we need to know the covariance of the estimate. The Fisher Information Matrix is the second gradient of the log likelihood function:

$$F = -\left.\frac{\partial^2 L\left(\underline{x}\right)}{\partial \underline{x}^2}\right|_{\underline{x}=\hat{\underline{x}}} \tag{6.4}$$

Fisher showed in his landmark papers (see the Library) that the asymptotic value of the covariance of the estimate is

$$P = F^{-1} \tag{6.5}$$

This value of the covariance matrix is called the *Cramer-Rao Bound*.

If the likelihood equation (6.3) is linear in the states, the MLE is unbiased and statistically efficient, and its covariance is given by (6.5). If the likelihood equation is nonlinear, the MLE is sufficient, asymptotically unbiased, and asymptotically efficient.

6.3 The MLE for Mean and Covariance of Correlated Gaussian Variables

Suppose that we have a series of samples of a distribution that is Gaussian, but we do not know the mean or covariance of the distribution. Thus, our measurements are

$$\underline{y} = \begin{bmatrix} \underline{y}_1 \\ \underline{y}_2 \\ \vdots \\ \underline{y}_M \end{bmatrix}, \quad \underline{y}_i = \underline{m} + \underline{v}, \text{Cov}\{\underline{v}\} = R \tag{6.6}$$

and our states are

$$\underline{x} = \begin{bmatrix} \underline{m} \\ (\text{Elements of } R) \end{bmatrix} \tag{6.7}$$

The likelihood function is

$$p\left(\underline{y}\middle|\underline{x}\right) = \prod_{j=1}^{M} p\left(\underline{y}_j\middle|\underline{x}\right) \tag{6.8}$$

where each set of measurements $\underline{y}_j$ has the probability distribution

$$p\left(\underline{y}_j\middle|\underline{x}\right) = \frac{1}{(2\pi)^{N/2}\cdot|R|^{1/2}}\cdot\exp\left(-\frac{1}{2}\cdot\left(\underline{y}_j-\underline{m}\right)^T\cdot R^{-1}\cdot\left(\underline{y}_j-\underline{m}\right)\right) \tag{6.9}$$

The log likelihood function is

$$L\left(\underline{x}\right) = -\frac{M\cdot N}{2}\cdot\ln\left(2\pi\right)-\frac{M}{2}\cdot\ln\left(|R|\right)-\frac{1}{2}\cdot\sum_{j=1}^{M}\left(\underline{y}_j-\underline{m}\right)^T\cdot R^{-1}\cdot\left(\underline{y}_j-\underline{m}\right) \tag{6.10}$$

We find separate likelihood functions for $\underline{m}$ and R, solving for $\underline{m}$ first:

$$l\left(\hat{\underline{m}}\right) = \frac{\partial L\left(\underline{x}\right)}{\partial \underline{m}} = \sum_{j=1}^{M} R^{-1}\cdot\left(\underline{y}_i-\hat{\underline{m}}\right) = \underline{0} \tag{6.11}$$

The solution to this one is simple:

$$\hat{\underline{m}} = \frac{1}{M}\cdot\sum_{j=1}^{M}\underline{y}_j \tag{6.12}$$

We find the covariance R by using the property of consistency, and solving for R^{-1} instead. We do this taking the gradient of the log likelihood equation with respect to the matrix R^{-1}. We need a couple of identities from our work with gradients first, though. See Gelb's Equation (2.1-75) page 23 for

$$\frac{\partial\ln\left(\left|R^{-1}\right|\right)}{\partial R^{-1}} = \frac{1}{\left|R^{-1}\right|}\cdot\frac{\partial\left|R^{-1}\right|}{\partial R^{-1}} = R \tag{6.13}$$

and Gelb's Equation (2.1-72) e\page 23 for

$$\frac{\partial\left(\underline{y}_i-\underline{m}\right)^T\cdot R^{-1}\cdot\left(\underline{y}_i-\underline{m}\right)}{\partial R^{-1}} = \left(\underline{y}_i-\underline{m}\right)\cdot\left(\underline{y}_i-\underline{m}\right)^T \tag{6.14}$$

With these two equations, we have the likelihood equation for the covariance as

$$l\left(R^{-1}\right) = \frac{\partial L\left(\underline{x}\right)}{\partial R^{-1}} = \frac{M}{2}\cdot\hat{R}-\frac{1}{2}\cdot\sum_{j=1}^{M}\left(\underline{y}_i-\hat{\underline{m}}\right)\cdot\left(\underline{y}_i-\hat{\underline{m}}\right) = 0 \tag{6.15}$$

This has the solution

$$\hat{R} = \frac{1}{M}\cdot\sum_{j=1}^{M}\left(\underline{y}_i-\hat{\underline{m}}\right)\cdot\left(\underline{y}_i-\hat{\underline{m}}\right)^T \tag{6.16}$$

Note that the estimator for $\hat{\underline{m}}$ is linear, so it is statically efficient and unbiased. However the estimator for $\hat{R}$ involves $\hat{\underline{m}}$ so that this part of the estimator is nonlinear. The inclusion of $\hat{\underline{m}}$ can be seen to make the sum have one less degree of freedom, and to make the variation in each sample smaller by a factor of *(M-1)/M*. A more rigorous analysis will bear out that an unbiased estimator of the covariance is

$$\hat{R}_{UNBIASED} = \frac{1}{M-1} \cdot \sum_{j=1}^{M} \left(\underline{y}_i - \hat{\underline{m}} \right) \cdot \left(\underline{y}_i - \hat{\underline{m}} \right)^T \tag{6.17}$$

6.4 Covariance of Estimates

We can use (6.4) and (6.11) to obtain the Cramer-Rao bound for the covariance of $\hat{\underline{m}}$. The Fisher information matrix for $\hat{\underline{m}}$ is

$$F_M = -\frac{\partial^2 L(\underline{x})}{\partial \underline{m}^2} = M \cdot R^{-1} \tag{6.18}$$

from which we have the covariance of the estimate $\hat{\underline{m}}$:

$$P_m = \frac{1}{M} \cdot R \approx \frac{1}{M} \cdot \hat{R} \tag{6.19}$$

Since the estimate of $\underline{m}$ is linear, this covariance, the Cramer-Rao bound, is applicable because the estimator is efficient.

The remainder of the covariance is more complex. The estimator for $\hat{R}^{-1}$ is nonlinear, and in fact the state vector contains both the elements of $\underline{m}$ and R, a total of M^2+M elements, not just M.

7 The Kalman Update as an MLE

We can easily pose the Kalman update as an MLE by focusing on the meausurement data and the extrapolated state vector. To avoid a notation clash here, we will denote the measurement for the Kalman filter update as $\underline{yy}$ and reserve the standard notation for measurements $\underline{y}$ as our measurements for the MLE. The data is then

$$\underline{y} = \begin{bmatrix} \underline{yy} \\ \underline{\tilde{x}} \end{bmatrix} \tag{7.1}$$

and the covariance of the measurements is

$$\mathrm{Cov}\left\{ \underline{y} \right\} = \begin{bmatrix} R & 0 \\ 0 & \tilde{P} \end{bmatrix} \tag{7.2}$$

The mean values of the measurements are

$$E\left(\underline{y} \right) = \begin{bmatrix} \underline{h}(\underline{x}) \\ \underline{x} \end{bmatrix} \tag{7.3}$$

We can write the conditional probability as

$$p\left(\underline{y},\underline{\tilde{x}}\middle|\underline{x}\right)=\frac{1}{\left(2\pi\right)^{(M+N)/2}\cdot\left|R\right|^{1/2}\cdot\left|\tilde{P}\right|^{1/2}}\cdot\exp\left(-\frac{1}{2}\cdot\begin{pmatrix}\left(\underline{yy}-\underline{h}\left(\underline{x}\right)\right)\cdot R^{-1}\cdot\left(\underline{yy}-\underline{h}\left(\underline{x}\right)\right)\\+\left(\underline{\tilde{x}}-x\right)^{T}\cdot\tilde{P}^{-1}\cdot\left(\underline{\tilde{x}}-x\right)\end{pmatrix}\right) \quad (7.4)$$

The log likelihood equation is

$$\begin{aligned}L\left(\underline{x}\right)=&-\frac{M+N}{2}\cdot\ln\left(\pi\right)-\frac{1}{2}\cdot\ln\left(\left|R\right|\cdot\left|\tilde{P}\right|\right)\\&-\frac{1}{2}\cdot\left(\left(\underline{yy}-\underline{h}\left(\underline{x}\right)\right)\cdot R^{-1}\cdot\left(\underline{yy}-\underline{h}\left(\underline{x}\right)\right)+\left(\underline{\tilde{x}}-x\right)^{T}\cdot\tilde{P}^{-1}\cdot\left(\underline{\tilde{x}}-x\right)\right)\end{aligned} \quad (7.5)$$

The likelihood equation is

$$l\left(\underline{\hat{x}}\right)=\frac{\partial L\left(\underline{\hat{x}}\right)}{\partial\underline{\hat{x}}}=H^{T}\cdot R^{-1}\cdot\left(\underline{yy}-\underline{h}\left(\underline{\hat{x}}\right)\right)+\tilde{P}\cdot\left(\underline{\tilde{x}}-\underline{\hat{x}}\right)=\underline{0} \quad (7.6)$$

We note that we can use the first two terms of a Taylor series,

$$\underline{h}\left(\underline{\hat{x}}\right)=\underline{h}\left(\underline{\tilde{x}}\right)+H\cdot\left(\underline{\hat{x}}-\underline{\tilde{x}}\right) \quad (7.7)$$

and the likelihood equation becomes

$$l\left(\underline{\hat{x}}\right)=-\left(\tilde{P}^{-1}+H^{T}\cdot R^{-1}\cdot H\right)\cdot\left(\underline{\hat{x}}-\underline{\tilde{x}}\right)+H^{T}\cdot R^{-1}\cdot\left(\underline{yy}-\underline{h}\left(\underline{\tilde{x}}\right)\right) \quad (7.8)$$

or,

$$\left(H\cdot\tilde{P}\cdot H^{T}+\tilde{R}\right)\cdot\left(\underline{\hat{x}}-\underline{\tilde{x}}\right)=H^{T}\cdot R^{-1}\cdot\left(\underline{yy}-\underline{h}\left(\underline{\tilde{x}}\right)\right) \quad (7.9)$$

from which we get the Kalman update:

$$\underline{\hat{x}}=\underline{\tilde{x}}+K\cdot\left(\underline{yy}-\underline{h}\left(\underline{\tilde{x}}\right)\right) \quad (7.10)$$

where the Kalman gain *K* is

$$K=\left(\tilde{P}^{-1}+H^{T}\cdot R^{-1}\cdot H\right)^{-1}\cdot H^{T}\cdot R^{-1} \quad (7.11)$$

The covariance *P* of the estimate is found from

$$\frac{\partial^{2}l\left(\underline{x}\right)}{\partial\underline{x}^{2}}=-P^{-1}=-\tilde{P}^{-1}-H^{T}\cdot R^{-1}\cdot H \quad (7.12)$$

The update, Kalman gain, and covariance of the estimate are as given in the Kalman filter equations.

Note that the Kalman filter extrapolation is an application of the linear covariance propagation equations. Thus, between those and the development of the update equations that we have here, we have a complete derivation of the Kalman filer equations.

8 Estimation Theory Library

8.1 Foundations of Modern Estimation Theory

1) R. A. Fisher, *Theory of Statistical Estimation,* Proc. Cambridge Phil. Soc., v. 22 p. 700 (1925).
2) R. A. Fisher, *On the Mathematical Foundations of Theoretical Statistics,* Phil. Trans. Roy. Soc., London, v. 222, p. 309 (1922).
3) R. A. Fisher, *Two New Properties of Mathematical Likelihood,* Proc. Roy. Soc., London, v. 144, p. 285 (1934).

4) R. A. Fisher, *The Logic of Inductive Inference,* J. Roy. Statist. Soc., v. 98, p. 39 (1935).
5) C. R. Rao, *Information and Accuracy Attainable in the Estimation of Statistical Parameters,* Bull. Calcutta Math. Soc., v. 37, pp 81-91 (1945).
6) Harold Cramer, *Mathematical Methods of Statistics,* Princeton University Press (1946).
7) A. Bhattacharyya, "On Some Analogues of the Amount of Information and their Use in Statistical Estimation, *Sankhya, v. 8, pp, 1, 201, 314 (1946, 1947, 1948).*

8.2 Major References

8) Richard Bellman, *Introduction to Matrix Analysis,* Second Edition, SIAM Press (1997) (reprint from 1984; first edition was in 1960).
9) M. G. Kendall and A. Stuart, *The Advanced Theory of Statistics, Vol. 2, Inference and Relationship,* Hafner, New York (1961).
10) Harry L. Van Trees, *Detection, Estimation, and Modulation Theory Part I,* John Wiley (1968).
11) Harold W. Sorenson, *Parameter Estimation,,* Marcel Dekker (1980).
12) M. B. Priestley, *Spectral Analysis and Time Series,* Academic Press (1981).
13) Louis L. Scharf, *Statistical Signal Processing,* Addison-Wesley (1991)