MEMORANDUM
RM-3329-PR
OCTOBER 1962

THE METHOD OF LEAST SQUARES AND OPTIMAL FILTERING THEORY

Y. C. Ho

PREPARED FOR:
UNITED STATES AIR FORCE PROJECT RAND

The RAND Corporation
SANTA MONICA • CALIFORNIA

MEMORANDUM
RM-3329-PR
OCTOBER 1962

THE METHOD OF LEAST SQUARES AND OPTIMAL FILTERING THEORY

Y. C. Ho

This research is sponsored by the United States Air Force under Project RAND – Contract No. AF 49(638)-700 – monitored by the Directorate of Development Planning, Deputy Chief of Staff, Research and Technology, Hq USAF. Views or conclusions contained in this Memorandum should not be interpreted as representing the official opinion or policy of the United States Air Force. Permission to quote from or reproduce portions of this Memorandum must be obtained from The RAND Corporation.

The RAND Corporation
1700 MAIN ST. • SANTA MONICA • CALIFORNIA

PREFACE

This Memorandum reports on part of RAND's continuing work on Guidance and Orbit Mechanics. It is a theoretical discussion of methods of estimating unknown parameters in the presence of noise, and can be applied to problems of estimating or predicting space trajectories or orbits. It should be of interest to mathematical statisticians, communications engineers, and those concerned with orbit mechanics.

SUMMARY

This Memorandum demonstrates the correspondence between the well-known method of least squares and the more recent optimal filtering theory of Kalman. It shows that via a simple lemma on matrix inversion, most of the results in linear filtering and prediction theory can be easily derived. The connection of the least square method with the so-called Duality Principle in optimal control theory is also discussed. This connection places in evidence the mathematical similarities between problems of control and problems of prediction. The Memorandum concludes with a proposed application for orbit determination of a 24-hour satellite using the techniques described. This application is concerned with computing corrections to the satellite's orbital parameters based on noisy observations of azimuth and elevation angles by improving an initial orbital parameter estimation through additional observations. The orbital parameter corrections may then be used as the input to an orbit transfer process or to refine a preliminary orbit.

CONTENTS

LIST OF SYMBOLS

A	matrix of coefficients which relate the control u to the terminal state c
C	matrix of coefficients which relate the satellite orbital parameter to observation
c	terminal state of a dynamic system
H	matrix of coefficients which relate the variable x to the observation z
J	quadratic form to be minimized
K	subscript used to indicate time up to and including t_k
k	subscript used to indicate time instant t_k
L	likelihood function
P_k	defined as the matrix $(H_K' R_K^{-1} H_K)^{-1}$
$P_{k/k}$	covariance matrix of error of estimation at t_k
P_{k+1}	defined as the matrix $(H_{k+1}' R_{k+1}^{-1} H_{k+1} + H_K' R_k^{-1} H_K)^{-1}$
$P_{k+1/k}$	extrapolated covariance matrix at t_{k+1} based on P at t_k
[illegible]	covariance matrix of w
R	covariance matrix of the errors of observation
u	control applied to a dynamic system
v	observation errors
w	input disturbance to the dynamic system
x	unknown variable to be estimated
$\hat{x}$	estimation of x
z	observation data
Δp	correction to the satellite orbital parameter
Δq	difference in the computed and observed satellite position
Φ	transition matrix of a dynamic system

A NOTE ON SYMBOLS

The following conventions for notation are adopted here: all upper case Roman or Greek letters are matrices; lower case Roman letters, vectors. Scalars are denoted by lower case Greek letters. Subscripts are used to denote components of matrices or vectors with the understanding that these components themselves may be matrices or vectors. Often, when it is clear in the context, a subscript (e.g., k) will be used to denote the value of the matrix or vector at the instant of time, t_k. For example, x_k means the value of the vector x at time t_k and x_{ki} is its ith component. The transpose of a matrix or vector is indicated by prime '. The quadratic form $x'Ax$ is often written as $\|x\|_A^2$.

I. INTRODUCTION

The least squares method has been investigated and put to use over a long period of time. A historical survey of the subject going as far back as Gauss is neither appropriate nor possible here. Interested readers are referred to Ref. 1 for a more complete treatment. Recently, due to the increasing emphasis on research in space science such as trajectory determination and optimization, stochastic control theory, etc., there has been considerable renewed interest in the theory of least squares estimation and its applications. Notably, there have been two approaches to the subject. As one approach, we can treat the problem as one of stochastic optimization. Following Wiener, but using the more powerful state-space techniques, we compute the conditional probability distribution function of the unknown variables. From these distributions, the least square estimate can be computed.[2, 3] The problem can also be approached from the statistical estimation point of view. We calculate the maximum likelihood estimates of the unknown variables and show that, in the case of gaussian noise, the estimates are optimal in the least squares sense.[4] * While it is known that these two approaches must yield similar results, so far as is known, the exact correspondence and its various implications have not been stated. It is the purpose of this Memorandum to:

- demonstrate this correspondnece through a very elementary derivation and a remarkable theorem on matrix inversion
- reconcile all known and useful results obtained from the two different approaches

---

*Also, Bryson, A. E. and M. Frazer, private communication, April 1962 o Discuss the implications of these results and their applications to problems of crbit determination for satellites.

II. A DETERMINISTIC DERIVATION

TIME-INDEPENDENT CASE

We shall start the discussion with a very simple problem. Once the derivation has been carried through for this case, generalization to more involved and practically more useful cases is straightforward. Let us consider the set of equations,

$$H_K x = z_K \tag{1}$$

where x is an unknown n-vector, H_K is a given mxn matrix where $m > n$ and z_K is a given m-vector. Here components of z_K usually play the role of observed data which are related to x, the unknown but desired parameters, through the theoretical relationship of Eq. (1). In general, of course, Eq. (1) does not possess a solution due to inevitable measurement errors, etc. It is thus reasonable to pose the problem of determining $x = \hat{x}_K$ such that the error $v_K = z_K - H_K x$ be minimized in some sense. In particular we consider the quadratic form

$$J = \| H_K x - z_K \|^2_{R_K^{-1}} = (H_K x - z_K)' \, R_K^{-1} \, (H_K x - z_K) \tag{2}$$

where R_K^{-1} is given mxm, positive definite (i.e. R_K also exists) symmetric matrix.*

---

*The interpretation of R_K^{-1} will be given in Section III. Clearly if R_K = Identity matrix = I, then J is simply the sum of the squared errors.

The physical origin of the above problem may be visualized as follows: At various time instants $t_1, t_2, \ldots t_k$ we make corresponding sets of measurements $z_1, z_2, \ldots z_k$. Each set of these measurements is theoretically related to a set of unknown parameters via the relationship

$$H_i x = z_i$$

A collection of these relationships for each set of measurements then yields Eq. (1) where H_K denotes the matrix formed by the sub-matrices $H_1, \ldots, H_k$, and z_K, the vector formed by $z_1, \ldots, z_k$. Now, upon setting grad J = 0, we obtain the well-known result that

$$x = \hat{x}_K = (H_K' R_K^{-1} H_K)^{-1} H_K' R_K^{-1} z_K \tag{3}$$

where it is assumed that the columns of H_K are linearly independent (i.e., $(H_K' R_K^{-1} H_K)^{-1}$ exists).* Now suppose some additional observations have been made, as represented by an r-vector z_{k+1}. We then ask what change $\Delta\hat{x}_{k+1}$ should be made to $\hat{x}_k$ such that the new $\hat{x}_{k+1} = \hat{x}_k + \Delta\hat{x}_{k+1}$ is again optimal with respect to the entire set of data (z_K, z_{k+1}). Equation (1) now has the form

$$\begin{bmatrix} H_K \\ \hline H_{k+1} \end{bmatrix} \begin{bmatrix} \hat{x}_k + \Delta x_{k+1} \end{bmatrix} = \begin{bmatrix} z_K \\ \hline z_{k+1} \end{bmatrix} \tag{4}$$

*If this is not satisfied in practice it usually implies that an observation scheme has some fundamental defects. Again we shall postpone a discussion on this point to later sections.

and it is clear that J should be

$$\|H_K(\hat{x}_k + \Delta x_{k+1}) - z_K\|^2_{R_K^{-1}} + \|H_{k+1}(\hat{x}_k + \Delta x_{k+1}) - z_{k+1}\|^2_{R_{k+1}^{-1}}$$

or,

$$J = \|\Delta x_{k+1}\|^2_{H_K' R_K^{-1} H_K} + \|H_{k+1}(\hat{x}_k + \Delta x_{k+1}) - z_{k+1}\|^2_{R_{k+1}^{-1}} \quad (5)$$

since $H_K' R_K^{-1} H_K x_k - H_K' R_K^{-1} z_K = 0.$* The first term in Eq. (5) has the obvious interpretation of being the cost of deviating from the previously determined optimal $\hat{x}_k$ for the set of data z_K. Now setting $\frac{\partial J}{\partial \Delta x_{k+1}} = 0$ we obtain

$$\Delta x_{k+1} = \Delta\hat{x}_{k+1} = (H_K' R_K^{-1} H_K + H_{k+1}' R_{k+1}^{-1} H_{k+1})^{-1} H_{k+1}' R_{k+1}^{-1} (z_{k+1} - H_{k+1} \hat{x}_k) \quad (6)$$

Equation (6) suggests a successive improvement scheme for determining the unknown x as more and more observations are made. If we define

$$(H_K' R_K^{-1} H_K)^{-1} = P_k \ , \quad (H_K' R_K^{-1} H_K + H_{k+1}' R_{k+1}^{-1} H_{k+1})^{-1} = P_{k+1}$$

then

$$P_{k+1}^{-1} = P_k^{-1} + H_{k+1}' R_{k+1}^{-1} H_{k+1} \quad (7)$$

$$\Delta\hat{x}_{k+1} = P_{k+1} H_{k+1}' R_{k+1}^{-1} (z_{k+1} - H_{k+1} \hat{x}_k) \quad (8)$$

---

*Equation (5) is actually an implicit statement of the Principle of Optimality and can lead to a dynamic programming solution of the problem, as was first pointed out by Bellman (Ref. 5).

are the desired recursion formulas. This is the stagewise estimation scheme suggested by Swerling[4] and more recently but independently by Bryson and Fraser.* However, the computational scheme of Eqs. (7) and (8) has one drawback--the inversion of the nxn matrix P_{k+1} must be carried out at every stage. This is a straightforward but nevertheless somewhat unpleasant task. However, this difficulty can be circumvented by the following remarkable Matrix Inversion Lemma which is well known in numerical analysis but not in control engineering.

Lemma: If $P_{k+1}^{-1} = P_k^{-1} + H'_{k+1} R_{k+1}^{-1} H_{k+1}$ where P, R are symmetric and nonsingular, and P_k, $R > 0$

then P_{k+1} exists and is given by

$$P_{k+1} = P_k - P_k H'_{k+1} (H_{k+1} P_k H'_{k+1} + R_{k+1})^{-1} H_{k+1} P_k \tag{9}$$

Proof: by direct substitution we have (dropping unnecessary subscripts)

$$P_{k+1}^{-1} P_{k+1} = I + H' R^{-1} H P_k - H' R^{-1} H P_k H' (H P_k H' + R)^{-1} H P_k - H' (H P_k H' + R)^{-1} H P_k$$

$$= I + H' \left[R^{-1} - R^{-1} (H P_k H' + R) (H P_k H' + R)^{-1} \right] H P_k = I \quad \text{Q.E.D.}$$

Now Eq. (9) can take the place of Eq. (7) with an important difference. The matrix $(H_{k+1} P_k H'_{k+1} + R_{k+1})$ is of dimension rxr where r is the number of new observations. Since we are at liberty to process new data, one or a group at a time, r can be simply taken as 1. Thus, the matrix inversion problem has been eliminated in the stagewise correction scheme.

---

*Bryson, A. E. and M. Frazer, private communication, April 1962.

TIME-DEPENDENT CASE

So far, we have treated the problem assuming that the elements of the matrix H_K and H_{k+1} are given. It is usual to postulate (with good justification in many cases) that the vector x is the state of a force-free linear dynamic system and that it changes with time according to some deterministic and known relationship.

$$x_{k+1} = \Phi(t_{k+1}, t_k)\, x_k = \Phi\, x_k$$

Thus, considering the dynamic system as shown in Fig. 1 we have at every instant the set of theoretical relationships

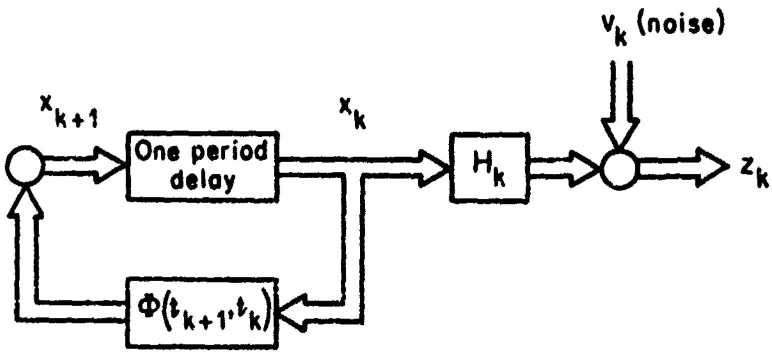

Fig. 1 — Dynamic system representation of observations

$$H_1\, x_1 = z_1$$

$$\vdots$$

$$H_k\, x_k = z_k$$

$$H_{k+1}\, x_{k+1} = z_{k+1}$$

Since the equations for the dynamic system are known, knowledge of the state of the system at any one time implies knowledge of the state at any other time. Now suppose that we already have an estimate $\hat{x}_k$ based on $z_1, \ldots, z_k$ and wish to obtain an estimate $\hat{x}_{k+1}$ by incorporating the new data z_{k+1}. The above equations can then be rewritten as

$$\left\{ \begin{bmatrix} H_1 \Phi(t_1, t_{k+1}) \\ \cdot \\ \cdot \\ \cdot \\ H_k \Phi(t_k, t_{k+1}) \\ \hline H_{k+1} \end{bmatrix} \right\} H_K \Phi(t_k, t_{k+1}) \left[\Phi(t_{k+1}, t_k) \hat{x}_k + \Delta x_{k+1} \right] = \left\{ \begin{bmatrix} z_1 \\ \cdot \\ \cdot \\ \cdot \\ z_k \\ \hline z_{k+1} \end{bmatrix} \right\} z_K \qquad (10)$$

where we have utilized the properties of the transition matrix

$$\Phi(t_k, t_1) = \Phi(t_k, t_r)\, \Phi(t_r, t_1)$$

and $\qquad \Phi(t_r, t_r) = 1$ for all $k, r \geq 0$

The complication we have added is the fact that x is now varying from one instant to another through the transformation $\Phi = \Phi(t_{k+1}, t_k)$. Equation (10) correctly reflects this fact since we still have $H_K x_K \approx z_K$. Now, a little reflection shows that we wish to minimize

$$J = \| H_{k+1} (\Phi \hat{x}_k + \Delta x_{k+1}) - z_{k+1} \|^2_{R^{-1}_{k+1}}$$

$$+ \| H_K \Phi^{-1} (\Phi \hat{x}_k + \Delta x_{k+1}) - z_K \|^2_{R^{-1}_K}$$

The second term in J measures the cost of deviating from the previous optimum x_k by the amount Δx_{k+1} for the data set z_K. A change Δx_{k+1} at t_{k+1} is equivalent to $\Phi^{-1} \Delta x_{k+1}$ at t_k, hence the inclusion of the factor Φ^{-1} in the expression. Minimizing J, we get

$$\Delta\hat{x}_{k+1} = P_{k+1} H'_{k+1} R^{-1}_{k+1} (z_{k+1} - H_{k+1} \Phi \hat{x}_k) \tag{11}$$

$$P^{-1}_{k+1} = \Phi^{-1'} P^{-1}_k \Phi^{-1} + H'_{k+1} R^{-1}_{k+1} H_{k+1} \tag{12}$$

A simple generalization of the matrix inversion lemma yields (again drop-dropping the obvious subscripts for simplicity)

$$P_{k+1} = \Phi P_k \Phi' - \Phi P_k \Phi' H' (H\Phi P_k \Phi' H' + R)^{-1} H\Phi P_k \Phi'$$

or

$$P_{k+1/k+1} = P_{k+1/k} - P_{k+1/k} H' (H P_{k+1/k} H' + R)^{-1} H P_{k+1/k} \tag{13}$$

Here we can think of $P_{k+1/k} = \Phi P_k \Phi'$ as the simple extrapolation of P_k to P_{k+1}, while $P_{k+1/k+1}$ represents the value of P_{k+1} when the r additional measurements are incorporated.

Furthermore, Eq. (11) can be rewritten as

$$\Delta\hat{x}_{k+1} = P_{k+1/k} H' (H P_{k+1/k} H' + R)^{-1} (z_{k+1} - H \Phi \hat{x}_k) \tag{14}$$

by showing that

$$P_{k+1/k} H' (H P_{k+1/k} H' + R)^{-1} = P_{k+1/k+1} H' R^{-1}$$

Equations (13) and (14) are computationally slightly more convenient that Eqs. (11) and (13).

It will now be shown that the purely deterministic derivation given above has a natural probabilistic interpretation which easily connects and reconciles the various currently known results in estimation.

III. PROBABILISTIC INTERPRETATION

To get a probabilistic interpretation of our results, we proceed to make one additional assumption. We shall assume the discrepancies in Eq. (4) are caused by measurement noises and that the measurement noise v_k can be adequately described by the joint gaussian probability distribution function

$$P(v_{k1}, \ldots, v_{kr}) = \frac{1}{(2\pi)^{r/2} \det (R^{-1})^{1/2}} \exp\left[-\frac{1}{2} v_k' R_k^{-1} v_k\right] \quad (15)$$

with zero mean and covariance matrix R_k, and that every set of r measurements are statistically independent. Thus, the joint probability density of the set of measurement $z_1, z_2, \ldots, z_k$ is simply a product of k expressions similar to Eq. (15). Now if we make the change of variable for the problem in Section II,

$$v_K = z_K - H_K x \quad (16)$$

and noting that the Jacobian of the transformation is 1, we obtain the joint probability density function for v_K as,

$$L(z_K, x) = \frac{1}{(2\pi)^{m/2} (\det R_K^{-1})^{1/2}} \exp\left[-\frac{1}{2} (z_K - H_K x)' R_K^{-1} (z_K - H_K x)\right] \quad (17)$$

where R_K^{-1} is a block diagonal consisting of $R_1^{-1}, \ldots, R_k^{-1}$ and m is the dimension of z_K. In order to maximize the likelihood (hence the symbol L for Eq. (17)) that the given z_K are produced by a certain x we choose $x = \hat{x}_k$ so as to minimize the exponent in Eq. (17), i.e., minimize $J = \|z_K - H_K x\|^2_{R_K^{-1}}$ which, of course, leads to the solution $x_k = (H_K' R_K^{-1} H_K)^{-1} H_K' R_K^{-1} z_K$. But by Eq. (16) we get

$$x_k = x + (H_K' R_K^{-1} H_K)^{-1} H_K' R_K^{-1} v_K \tag{18}$$

which leads to the expression that

$$E(\hat{x}_k) = x \tag{19}$$

$$Cov(x_k) = E\left[(x - \hat{x}_k)(x - \hat{x}_k)'\right] = (H_K' R_K^{-1} H_K)^{-1} = P_k \tag{20}$$

by direct calculation. When r more measurements are incorporated, we have in addition to Eq. (16)

$$H_{k+1} x + v_{k+1} = z_{k+1} \tag{21}$$

Now, because of the independence of the observed errors (v_K, v_{k+1}) the likelihood function for the data (z_K, z_{k+1}) and the unknown parameters $x = \hat{x}_k + \Delta x_{k+1}$ (of Eq. (5)) is

$$L(z_{k+1}, z_K, x) = \frac{1}{(2\pi)^{r/2} (\det (R_{k+1}^{-1}))^{1/2}} \exp -\left[(z_{k+1} - H_{k+1} (\hat{x}_k + \Delta x_{k+1}))' R_{k+1}^{-1} (z_{k+1} - H_{k+1} (\hat{x}_k + \Delta x_{k+1}))\right]$$

$$\cdot \frac{1}{(2\pi)^{m/2} (\det (P_k^{-1}))^{1/2}} \exp\left[- (x - \hat{x}_k)' P_k^{-1} (x - \hat{x}_k) \right] \tag{22}^*$$

*Since P_k measures the covariance of the error of the estimate x_k (Eq. (19)), the second half of the term is the likelihood of the error Δx of x_k. For example if P_k is diagonal and P_{11} is very small (implying considerable confidence in the estimate x_{11}), then a larger penalty is imposed on guessing Δx_1 away from zero. See also the deterministic interpretation in Section II.

Maximizing Eq. (22) with respect to Δx_{k+1} is equivalent to minimizing the sum of the exponents. We are then led to the same expression for $\Delta\hat{x}_{k+1}$ (Eq. (8)). The new estimate $\hat{x}_{k+1} = \hat{x}_k + \Delta\hat{x}_{k+1}$ after various manipulations using Eqs. (9) and (18) yields

$$\hat{x}_{k+1} = x + P_{k+1} H_K' R_K^{-1} v_K + P_{k+1} H_{k+1}' R_{k+1}^{-1} v_{k+1} \tag{23}$$

Thus we have

$$E(\hat{x}_{k+1}) = x \tag{24}$$

$$\begin{aligned} \text{Cov}\ (\hat{x}_{k+1}) &= E\left[(x - \hat{x}_{k+1})\ (x - \hat{x}_{k+1})'\right] \\ &= P_{k+1} (H_{k+1}' R_{k+1}^{-1} H_{k+1} + H_K' R_K^{-1} H_K)\ P_{k+1} \\ &= P_{k+1} \end{aligned} \tag{25}$$

Consequently, we arrive at the following well-known results that in the case of least square estimates with gaussian noise we have:

1. The least squares estimate x as given by Eq. (8) and calculated via Eq. (9) is a gaussian random vector (since from Eq. (23) we know it is merely a linear combination of other gaussian variables) with mean, x, and covariance of the error of estimation P_{k+1}.

2. The least squares estimate is unbiased and efficient (i.e., it has minimum variance which is merely the statistical way of saying "least squares").

3. The conditional probability of x given z_K and z_{k+1} is a gaussian random vector and has mean x and covariance P_{k+1}, (this

is evident from Eq. (23)). Thus, the conditional mean of x <u>is</u> the least squares estimate.

Nothing is changed if we consider the time-dependent case. All the formulas in Section II apply without modification. At any time instant t_k we have,

1. A current best estimate x_k which is also the conditional mean of x_k given $z_0, z_1, \ldots, z_k$.

2. The covariance matrix of the error, $x_k - \hat{x}_k$, $P_{k/k}$.

3. The extrapolated error covariance matrix of $x_{k+1} - \Phi\hat{x}_k = \Phi(x_k - \hat{x}_k)$ which is $P_{k+1/k} = \Phi\, P_{k/k}\, \Phi'$.

When a new set of measurements z_{k+1} is made we have

4. The updated estimate $\hat{x}_{k+1} = \Phi\, \hat{x}_k + \Delta\hat{x}_{k+1}$. (Eq. (11) or (14))

5. The updated error covariance matrix $P_{k+1/k+1}$ given by Eq. (13).

Furthermore, a simple generalization is possible. Suppose we consider the case as shown in Fig. 2.

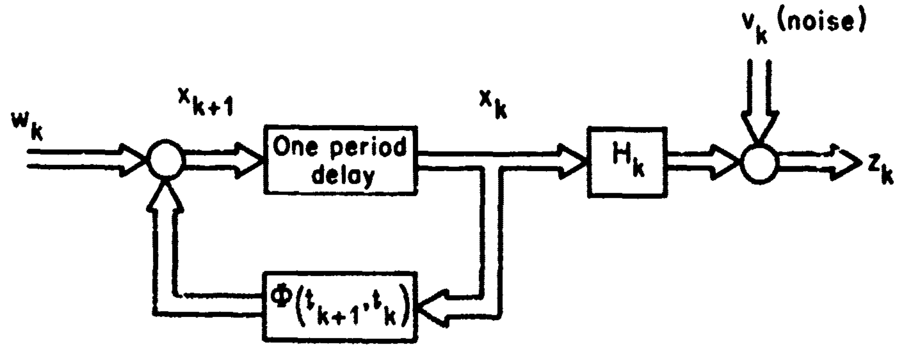

Fig. 2 — Dynamic system representation of observations with input disturbance

Where the dynamic system is subject to some random disturbances, let

$$E(w_k) = 0$$

$$E(w_k' w_j) = Q\delta(k - j)$$

$$E(w_k' v_j) = 0 \text{ for all } k, j$$

i.e., w is a white gaussian random process independent with v. It is clear that the effect of w at each step is to insert an uncertainty into the value of x at the next step. Thus, if the covariance of $\hat{x}_k - \hat{x}_k$ at x_k is $P_{k/k}$ then the extrapolated covariance of $x_{k+1} - \Phi \hat{x}_k$ is simply

$$P_{k+1/k} = \Phi P_{k/k} \Phi' + \Gamma Q \Gamma' \tag{26}$$

The covariances add since we assume w to be white. Then if $P_{k+1/k}$ from Eq. (26) is used in Eq. (13) for the updating of $P_{k+1/k+1}$, all other equations hold with no modifications. Equations (26), (14) and (13) are exactly what were derived in Eq. (2) via the Wiener-Kalman approach.*

Consequently, we have established the exact correspondence between the two seemingly different approaches to estimation. The missing link between the two is provided by the Matrix Inversion Lemma.

---

*Kalman assumed an even more general model where w and v are correlated. But the complication can be reduced to a case with no correlation, as also was shown in Ref. 2. In all other respects, the notations here and those in Ref. 2 are the same.

The main computational contribution of the Wiener-Kalman method is the relative simplicity in the calculation of the covariance matrices $P_{k+1, k+1}$. However, the original derivation of the variance Eq. (12) is considerably more involved without the use of the Lemma. Conceptually, the Wiener-Kalman approach via the calculation of the conditional probability distribution of the unknown parameters is more general. Once the conditional probabilities are computed, optimal estimates for other criteria can be calculated. In fact, it has been pointed out (Ref. 2) that the conditional mean $\dot{x}_k$ is optimal for many other criterion functions than the least squares one. Furthermore, the likelihood approach is not particularly hampered by the presence of nonlinear functions in the case of least squares estimates with gaussian noise. We shall attempt to discuss some aspects of this nonlinear estimation problem in another Memorandum. The present Memorandum is aimed primarily at the linear case where the two approaches are identical.

It is also clear that the estimation scheme Eqs. (26), (13) and (14) remain valid if w is nonstationary, i.e., Q is time-varying. Furthermore, the initial estimate $\hat{x}_o$ and its covariance P_o must be provided as part of the startup data. Usually we assume $x_o = 0$ and choose the best guess for P_o. An alternate procedure is to make at least n measurements first before producing an estimate and then let $P_o = (H_1' R_1^{-1} H_1)^{-1}$.

IV. DUALITY WITH CONTROL PROBLEM

There is an interesting mathematical similarity between the problem discussed above and the well-known quadratic error control problem as first discovered by Kalman.[3] In the present context this so-called duality is even more transparent. Recall that in Section II we first consider the problem of

$$\left\{ \begin{array}{l} \text{Minimize the error } J = v_K' R_K^{-1} v_K \\ \text{for the set of equations} \\ \begin{bmatrix} H_K \end{bmatrix} [x] = \begin{bmatrix} z_K \end{bmatrix} - \begin{bmatrix} v_K \end{bmatrix} \end{array} \right\} \text{(Problem A)}$$

Let us now consider the so-called dual problem of

$$\left\{ \begin{array}{l} \text{Minimize } J = u_K' R_K u_K \\ \begin{bmatrix} A_K \end{bmatrix} \begin{bmatrix} u_K \end{bmatrix} = [c] \end{array} \right\} \qquad \text{(Problem B)}$$

In Problem (A) we have more equations than unknowns and one tries to minimize the square error in the resultant set of equations. In (B) we have more unknowns than equations and one tries to minimize the square of the solution to the equations.

Assume u_K has m components and A_K is nxm where m > n, then it is well-known that the solution to (B) is

$$u_K = R_K^{-1} A_K' (A_K R_K^{-1} A_K')^{-1} c \qquad (27)$$

Now if r more free components, and r more columns are added to u_K and A_K, respectively, then we have,

$$\left[A_{k+1} \,\vdots\, A_K\right] \begin{bmatrix} u_{k+1} \\ \hdashline u_K \end{bmatrix} = \left[c\right] \tag{28}$$

and the optimal solution to (B) is

$$u_{k+1} = R_{k+1}^{-1} A'_{k+1} \left(A_{k+1} R_{k+1}^{-1} A'_{k+1} + A_K R_K^{-1} A'_K\right)^{-1} c$$

$$u_K = R_K^{-1} A'_K \left(A_{k+1} R_{k+1}^{-1} A'_{k+1} + A_K R_K^{-1} A'_K\right) c \tag{29}$$

The additional r components of u_{k+1} are given by a formula very similar to the formula for yielding the correction $\Delta\hat{x}_{k+1}$ in Eq. (6). The striking thing is, of course, the matrix $(A_{k+1} R_{k+1}^{-1} A'_{k+1} + A_K R_K^{-1} A'_K)^{-1}$ If we define $P_k^* = (A_K R_K^{-1} A'_K)^{-1}$

$$P_{k+1}^* = (A_K R_K^{-1} A'_K + A_{k+1} R_{k+1}^{-1} A'_{k+1})^{-1}$$

$$\text{or} \quad P_{k+1}^* = P_k^{*-1} + A_{k+1} R_{k+1}^{-1} A'_{k+1}$$

then we have via the Matrix Inverse Lemma

$$P_{k+1}^* = P_k^* - P_k^* A_{k+1} (A'_{k+1} P_k^* A_{k+1} + R_{k+1})^{-1} A'_{k+1} P_k^* \tag{30}$$

which is exactly the same as Eq. (12) if we identify H_{k+1} with A'_{k+1}. This is really the essence of duality.

Problem (B), of course, can be given a control interpretation. Vector c plays the role of desired terminal state, components of u_{k+1}, u_K, the control efforts at different times, and A_{k+1} A_K, the matrix of influence coefficients which relates the effort at one instant to the state at the terminal time. The following more or less obvious statements concerning problems (A) and (B) further illustrate the concepts of duality:

(A)	(B)
For an infinite stage estimation process with stable dynamics, we can identify the true x with increasing precision.	For an infinite step control process with stable dynamics we can achieve the desired state with decreasing effort.
When no observation has been made we have no confidence in our estimates $P_0 = \infty$.	When there are no control steps left, the desired state must be achieved with infinite energy $P_0^* = \infty$.
The columns of H_1 must be linearly independent so that $(H'_K R_K^{-1} H_K)^{-1}$ exists. This is called observability.	The rows of A_1 must be linearly independent so that $(A_K R_K^{-1} A'_K)^{-1}$ exists. This is called controllability.

The extension of the above ideas to the time-dependent case of Figs. 1 and 2 is straightforward. In fact, it is well-known that for a discrete dynamic system

$$x_{k+1} = \Phi^* (t_{k+1}, t_k) x_k + H_k^* u_k$$

$$z_k = H_k^* x_k + v_k = y_k + v_k \tag{31}$$

and criterion function

$$J = \| x_N \|^2_{P_0} + \sum_{i=1}^{N-1} \| u_i \|^2_R + \| y_i \|^2_Q$$

the optimal control sequence is given by (via dynamic programming see Refs. 3 and 6)

$$u_i = -(H^{*'} P^*_{N-i-1} H^* + R)^{-1} H^{*'} P^*_{N-i-1} \Phi^* x_i \tag{32}$$

where

$$P^*_{N-i/N-i-1} = P^*_{N-i-1} - P^*_{N-i-1} H^* (H^{*'} P^*_{N-i-1} H^* + R)^{-1} H^{*'} P^*_{N-i-1} \tag{33}$$

$$P^*_{N-i} = \Phi^{*'} P^*_{N-i/_{N-i-1}} \Phi^* + \Gamma^{*'} Q \Gamma^* \tag{34}$$

Equations (34) and (33) are exactly the same as (26) and (13) if we identify

Γ^*	with	Γ'	
H^*	with	H'	
Φ^*	with	Φ'	
P^*_{N-i}	with	$P_{k/k}$	$i = N-1, \ldots, 0$ $k = 1, \ldots, N$

and hence, from this comes the name Duality Principle.

V. APPLICATION

To illustrate the application of least squares estimation theory to trajectory determination, we consider the case of a 24-hour satellite in a circular orbit. With passage of time, the satellite tends to drift out of the orbit due to various causes such as gravitational perturbations, etc. Noisy observations are made on the satellite from the ground. It has been shown (Eq. (8)) that under various linearity assumptions, the following relationship is true

$$\Delta q = C \; \Delta p \tag{35}$$

where

$$\Delta q = \begin{bmatrix} \Delta A_1 \cos h_1 \\ \Delta h_1 \\ \Delta A_2 \cos h_2 \\ \Delta h_2 \\ \cdot \\ \cdot \\ \cdot \\ \Delta A_m \cos h_m \\ \Delta h_m \end{bmatrix}$$

$$\Delta p = \begin{bmatrix} \Delta a/a \\ \Delta(e \cos E_o) \\ \Delta(e \sin E_o) \\ \Delta\tilde{u}_o \\ \Delta\tilde{v}_o \\ \Delta\tilde{w}_o \end{bmatrix}$$

$$C_{ij} = C_{ij}\ (\theta_t, A_r, h_r, \Phi, \lambda)$$

The definitions of the variables are:

- A_r = azimuth angle of the satellite if it is exactly in the circular orbit
- h_r = elevation angle of the satellite if it is exactly in the circular orbit
- ΔA_i = measured difference between the observed azimuth and A_r at time t_i
- Δh_i = measured difference between the observed elevation and h_r at time t_i
- θ_t = time of observation
- Φ = longitude of observation point
- λ = latitude of observation point
- p = set of orbital parameters
- Δp = change of the orbital parameters from the reference values of the perfect circular orbit

Since the elements of C are independent of the observations and are calculable for all time, we can numerically solve the variance equation beforehand. If we define the successive rows of C as c_k' then

$$P_{k+1} = P_k - P_k\, c_{k+1}\, c'_{k+1}\, P_k\, (c_{k+1}\, P_{k-1}\, c'_{k+1} + \sigma^2)^{-1} \qquad (36)$$

Where σ_k^2 is the variance of the observation errors in azimuth and elevation for the kth observation. Equation (35) can be computed beforehand for as many steps as needed. A rough estimate of the number of steps (i.e., observations) that are needed can be obtained by checking the size of $Tr(P_k)$, since $Tr(P_k)$ is the sum of the variance of the error of the estimates Δp. The initial condition for Eq. (36) can be taken to be any reasonable guess Δp_o. A good way of estimating P_o would be to compute a few extra rows of C, say, c'_{-1}, c'_{-2}, ..., c'_{-n} and form the matrix C^*. Then we can let

$$P_o = \frac{1}{\sigma^2}\,(C^{*'}\; C^*)^{-1} \qquad (37)$$

When we have finished computing P_k's for all k, then the estimate in the change in orbital parameters is simply given by

$$\Delta p_k = \Delta p_{k-1} + \frac{1}{\sigma^2}\, P_k\; c_k\, (\Delta q_k - C\Delta p_{k-1}) \qquad (38)$$

where Δq_k is the difference in measured and computed azimuth or elevation. Equation (38) is the only equation that needs to be processed in real time as observations are made.

In practice, of course, the above procedure is applicable only when the true change Δp is small so that the linearity assumptions hold. The case when the change in orbit parameter is large so that Eq. (35) is not a good approximation will require different treatment. We shall discuss this in a later Memorandum.

REFERENCES

1. Linnik, Y. V., Method of Least Squares and Principles of the Theory of Observations, trans. by R. C. Elandt, Pergamon Press, 1961.

2. Kalman, R. E., "New Methods and Results in Linear Prediction and Filtering Theory," RIAS Report 61-1, to appear in the Proc. of Symposium on the Application of Random Function Theory and Probability, John Wiley and Sons, 1962.

3. Kalman, R. E. and R. C. Bucy, "New Results in Linear Optimal Filtering and Prediction," J. of Basic Eng., Series D, Trans. of ASME, March 1961.

4. Swerling, P., "First Order Error Propagation in a Stagewise Smoothing Procedure for Satellite Observations," J. Astron. Sci., Vol. 6, No. 3, Autumn 1959.

5. Bellman, R., Introduction to Matrix Analysis, McGraw Hill, 1960, pp. 154-155.

6. Smith, F. T., "Optimization of a Multistage Transfer Process by Dynamic Programming," ARS Journal, November 1961, pp. 1553-1559.

7. Smith, F. T. and J. H. Hutcheson, An Orbital Control Process for a 24-Hour Communication Satellite, The RAND Corporation, RM-2809-NASA, October 1961.