NASA TECHNICAL REPORT

NASA TR R-251

ON THE THEORY AND METHODS OF STATISTICAL INFERENCE

by Gerald L. Smith

Ames Research Center
Moffett Field, Calif.

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION • WASHINGTON, D. C. • APRIL 1967

NASA TR R-251

ON THE THEORY AND METHODS OF STATISTICAL INFERENCE

By Gerald L. Smith

Ames Research Center
Moffett Field, Calif.

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

For sale by the Clearinghouse for Federal Scientific and Technical Information
Springfield, Virginia 22151 – CFSTI price $3.00

ON THE THEORY AND METHODS OF STATISTICAL INFERENCE

By Gerald L. Smith
Ames Research Center

SUMMARY

Statistical inference is the process of intelligently using information from observations and experiments to draw conclusions and make decisions. The class of inference problems includes all the classical statistical problems of point and parameter estimation, regression analysis, and hypothesis testing, as well as the engineering problems of control and filtering.

In order that inference problems may be stated in mathematical terms, it is necessary to utilize mathematical measures of the often intangible entities called knowledge and goodness (or optimality). Probability is the measure utilized for knowledge, and loss functions serve as measures of optimality. The conversion of knowledge and goodness into functional or numerical terms is usually quite subjective and thus to a substantial extent arbitrary. Nevertheless, the conversion is necessary if the language of mathematics is to be employed in the inference logic.

The various mathematical methods developed, under different assumptions, for solving inferential problems are unavoidably related since all are (in principle) reducible to a common form. Some of the most fundamental of these--least squares, minimum variance, and maximum likelihood--are reviewed to show their interrelationships. The newer developments of sequential analysis and filtering are similarly described. All of these are shown to be special cases of the general theory of Bayesian decision making.

The Bayesian decision theory requires the specification of an appropriate loss function. The optimum decision is then defined as that which minimizes the mathematical expectation of this loss. Computation of this expectation requires obtaining the posterior distribution of the unknown state based on prescribed scientific observations.

Under restrictive conditions the computations involved in this general decision-making procedure are reducible to relatively simple forms. For application under more general conditions methods of approximation and efficient computational algorithms are presently being developed.

INTRODUCTION

The methods of statistics and modern filter theory are finding ever increasing application in present-day scientific and engineering problems because of the power and universality of these methods. However, an individual attempting to use statistical

techniques often has not had the opportunity to acquire a knowledge of the underlying principles and thus may find himself bewildered by the vast array of literature from which he seeks to extract techniques suitable for his problem.

One of the situations in which this difficulty has appeared is in the application of the Kalman filter theory to space vehicle navigation (refs. 1 and 2). This is a problem in trajectory determination for the solution of which the methods of least squares and maximum likelihood have also been applied, and it is natural to ask what are the differences between the methods--that is, whether the new method is better than the old. Answers to such questions can be made only by examining the foundations of the methods, which can be shown to belong to the same basic theory. As one investigator (ref. 3) says, "Diversified fields of scientific inquiry demand somewhat different and specialized techniques in particular problems; yet the fundamental principles which underlie all the various methods are identical regardless of the field of application."

It is the purpose of this paper to present a unified summary of several specialized techniques, showing how they are related to the general theory of statistical inference. The paper begins with a brief description of (1) the general principles of statistical inference and areas of their application in science and engineering, and (2) the probability and optimality concepts which are the foundations of the theory. Then the development of the specific techniques is outlined to show their differences and similarities.

The treatment used here does not involve esoteric mathematics and carries no pretense of completeness or rigor. By keeping the development simple, it is hoped that a certain class of readers can be helped to a greater insight, a better feeling for the choice of specialized methods, and a more knowledgeable interpretation of results obtained therefrom. For actual use of the particular techniques, the reader must undertake a more intensive study, using the extensive literature, only a small part of which is referenced herein.

SYMBOLS

a_{ij}	element in ith row and jth column of matrix A	$f(x,u)$	function of x and u in system equation (56)
A	matrix in linear observation equation (24)	F	matrix in linear system equation (63)
B	arbitrary matrix	$g(x)$	function of x in observation equation (9)
C	arbitrary matrix	G	matrix in linear system equation (63)
$d(x)$	function of x in l	$h(x,e)$	function of x and e in observation equations (62) and (37)
D	normalizing constant in normal probability density function	I	identity matrix
		J	optimality criterion functional
e	observation error vector	K	gain matrix in Kalman filter equation (70)
$E(\ )$	expectation of ( )		

$l(x,\delta)$	loss function (sometimes used without arguments)
N	number of observations
$p(x)$	probability density function of x
$p(x,y)$	joint probability density function of x and y
$p(x \mid y)$	conditional probability density function of x given y
P	prior covariance matrix of x
Q	covariance matrix of e; $E(eeT)$
R	covariance matrix of u; $E(uuT)$
u	random disturbance or control input to system
W	positive semidefinite weighting matrix
x	state or unknown parameter or data vector
X	matrix (or vector) composed of a set of x_i
y	observation or data vector
Y	matrix (or vector) composed of a set of y_i
α	unknown parameter vector
δ	decision
Λ	posterior covariance matrix of x
μ	mode of a distribution
σ	standard deviation
Ω_x	state space

Notation

$\arg\min_{\delta}(\ )$	set of δ generating the minimum $(\ )$
$(\ )^T$	transpose of the matrix $(\ )$
$(\ )^{-1}$	inverse of the matrix $(\ )$
$(\hat{\ })$	estimate of $(\ )$
$(\tilde{\ })$	error in estimation of $(\ )$

Subscripts

i,j,m,n	general indices
k	time index
opt	optimal

BASIC CONCEPTS IN STATISTICAL INFERENCE

The Nature of Statistical Inference

Conclusions regarding the nature of our environment have been derived from myriad observations, or physical sensations, induced by our contact with the physical world; the process of logically interpreting empirical data is termed statistical inference.

Scientific knowledge is never exact--that is, we may assume there exists some true state of things, but it is never perfectly discernible. Our "laws" and compilations of "facts" are then simply estimates of the true state made up of economical condensations of as much as possible of our observational data. The continuing purpose of scientific

effort is to test these laws and facts by means of further experiments and by the process of statistical inference to refine and improve our knowledge. As Jeffreys says (ref. 4), "Scientific progress never achieves finality; it is a method of successive approximation."

Applied science is concerned with the problems of utilizing scientific knowledge in the making of decisions regarding courses of action. Clearly, since decisions depend upon knowledge and knowledge depends upon empirical data, decision making can be regarded as the direct utilization of observations for the determination of courses of action, with or without the intermediate step of estimating the state of nature. Thus, if decisions are thought of as inferences, then statistical inference encompasses decision making as well as estimation. In fact, by regarding estimation as "deciding" what is the true state of things, one concludes that decision making and statistical inference are synonymous in the broad sense.

By the foregoing argument, it is seen that all scientific effort involves statistical inference. Classical statistics, including such subjects as hypothesis testing, regression analysis, estimation, and the analysis of variance, has to do with determining the true state of things. Filter theory has to do with determining the messages contained in received signals. Control theory has to do with the design of automated systems which have the job of "deciding," on the basis of measurements of the environment, what control should be applied to a given plant. Operations research deals with decisions regarding resource allocation, scheduling, inventory control, and the like. And game theory treats the problem of decision making in contests with intelligent adversaries.

To apply the methods of statistical inference to any particular scientific problem, it is first necessary to express the problem in mathematical language. This is often the most difficult step because knowledge and utility concepts--and even sometimes the data itself--frequently are not given in numerical form. In the following sections some of the difficulties of this conversion are discussed.

Probability as the equivalent of empirical knowledge.--Any numerical measure of empirical knowledge must reflect the uncertainty inherent in such knowledge. Considering that empirical knowledge consists, not in facts, but rather of sets of propositions with varying degrees of assurance as to their correct representation of the facts, the assignment of specific numerical values to these degrees of assurance will constitute an appropriate measure, provided, of course, that such assignment is logically consistent.

Historically, the first such assignment of numerical values to degrees of assurance occurred in connection with empirical data generated by repeated random experiments, which it was observed exhibit a significant regularity. For instance, in many throws of a true die, face A is found to turn up almost exactly 1/6 of the time. It is appropriate then to assume that the assurance, or probability, of obtaining result A in one throw is 1/6, and the number 1/6 then represents the empirical knowledge gained from the die experiments. The extension of the concept of probability to represent in a general way "degrees of reasonable belief" which are not necessarily based on such direct frequency ratio measurements is fairly natural. It may be observed that there is a notable lack of agreement amongst mathematicians as to the legitimacy of this more general interpretation of probability. However, the arguments are strictly philosophical, and here we choose to employ the general definition because it permits us to characterize all empirical knowledge in the same way so that we may then apply the entire body of probability theory to the general problem of statistical inference.

A complete set of probabilistic statements (or likelihood measures) is called a probability distribution and is, of course, defined on the set of all possible states. The idea of a distribution is illustrated in figure 1, which represents a continuous probability density

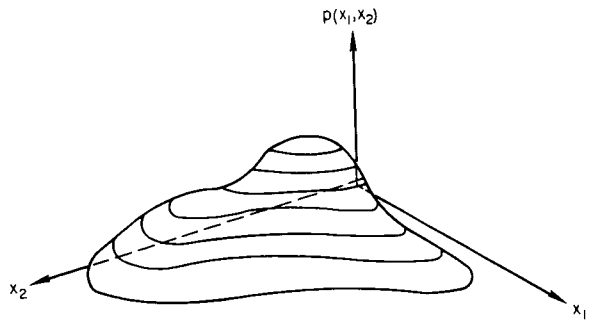

Figure 1 - Probability density function

function. The bounded space in the horizontal (x_1, x_2) plane represents the set of all possible conditions, that is, the state space. Elevation of the surface above the plane represents the degree of confidence one has that the state lies in the neighborhood of that point directly under the surface, and the surface is designated the probability density function, $p(x)$. Probability statements are in the form $p(x)dx$, which is the probability that x lies in the infinitesimal region dx. The $\int_{\Omega_x} p(x)dx$ over the whole state space is the sum of all these probabilities, and since the whole state space represents certainty, the integral is equal to unity.

Evidently, a probability distribution defined on a given state space contains all the knowledge we can express regarding the particular set of circumstances represented by the state space. Hence, probability is the full equivalent of empirical knowledge. As such, it is clear that probability is constantly subject to change as new information is gained by scientific investigation.

Relative to some specific set of experiments, it is logical to refer to prior and posterior probabilities as the summation of knowledge respectively before and after the experiments. Posterior probability is also termed conditional probability, that is, one speaks of the probability of a state x conditioned upon observations y. The conventional notation employed for the posterior (or conditional) density function is $p(x|y)$. Obviously, the posterior probability for one set of experiments becomes the prior probability for a succeeding set. In fact, practically speaking every probability is posterior inasmuch as it summarizes a certain body of empirical knowledge, which may be said to "condition" said probability.

Specification of prior probabilities.--Consider now how one obtains the prior distribution. That is, where does one find the table of statements of relative belief one has in various conditions, or the curve or function representing degrees of belief in a continuum of possible conditions? This specification of prior probabilities, together with the specification of an optimality criterion, is one of the more difficult aspects of statistical inference problems because it so often entails the conversion from the vague realm of beliefs to the specific realm of numbers.

The specification of prior probabilities is simple when the investigation pertains to a situation in which the results of repeated experiments in the past have been recorded and summarized, such as in the tossing of a coin or shooting at a target or in the testing

of items from a production line. In such cases it is known that the results were sometimes state A, sometimes state B, etc., and the prior distribution is given directly by the relative frequencies of the various occurrences.

What does one do, however, in a less objective investigation, such as determining whether or not there is life on Mars? Our prior belief (or disbelief) in life on Mars is real enough, being based on quantities of information obtained in assorted observations. This information has to be sensibly organized and analyzed to make the required probabilistic statements. If the data are insufficient to perform the analysis, we could simply make a knowledgeable judgment. Raiffa and Schlaifer (ref. 5) state that, "In such situations the prior distribution represents simply and directly the betting odds with which the responsible person wishes his final decision to be consistent."

This goes against the grain for many meticulous scientists, who often say that no probability statements can be made in such circumstances. But surely this is unduly pessimistic. The summation of our information at any particular point in scientific investigations can not be ignored, and probability statements are the only appropriate way of reflecting this knowledge.

As pointed out by Raiffa and Schlaifer, the psychological difficulty of assigning "betting odds" to express the investigator's belief in the state of affairs is so great that he will generally be reluctant to specify the distribution in any more detail than by a few summary measures, such as the mean and a percentile or two. From this point on, he is likely to settle for a representation of the distribution by any reasonable form which fits these minimum specifications. This practice accounts, in part at least, for the widespread use of simple distributions such as the normal, which is so easily specified by two parameters (i.e., the mean and variance).

The wide variety of special methods for the solution of statistical inference problems which fill the literature reflects a variety of different assumptions regarding prior probability. These range from complete ignorance of the distribution, as is typical in regression analysis and least-squares techniques, to complete knowledge, that is, the prior distribution completely specified. Neither of these assumptions is ever wholly correct. Total ignorance is implausible in the statement of any problem, and perfect knowledge is equally implausible from the argument of subjectivity which implies that nothing is ever known perfectly. However, the assumptions can usually be rationalized in terms of approximations to the true situation. When prior knowledge is assumed totally lacking what one really means is that the bulk of the information regarding the state is contained in the observations, so that for practical purposes prior knowledge may as well be ignored. When some type of "perfect" prior knowledge is assumed, one really means that the prior knowledge is already so good with respect to the information contained in the observations that the latter cannot be expected materially to improve the estimate.

One often has to be cautious, of course, in assuming any kind of "perfect" prior knowledge. Even though prior knowledge may be extremely good with respect to a certain body of data, there may always come a time with continued experimentation that the assumption is no longer valid. For example, in the determination of the trajectory of a space vehicle from tracking data, the astrodynamic constants (e.g., mass of the earth, astronomical unit, etc.) may be considered to be known perfectly at first, relative to a small amount of data. But after much tracking data has been accumulated, it will be found that errors in the constants can be detected, that is, the estimates of the constants can be improved.

Posterior probabilities.--Mathematically, the transition from the prior to the posterior (also called conditional distribution) is expressed by Bayes' rule (given here in terms of the density functions):

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)} \tag{1}$$

Here, x represents the state and y the observations, $p(x)$ is the prior density function, and $p(x|y)$ the posterior density function. The transition is illustrated in figure 2.

It is seen from equation (1) that one simply multiplies $p(x)$ by $p(y|x)/p(y)$ to obtain $p(x|y)$. One thus needs to have $p(y|x)$ (which is called the likelihood function) and $p(y)$. The observation, y, is a function of the state, x, and observation error, e; that is, $y = h(x,e)$, where e is, like x, an uncertain quantity. Knowledge regarding e is embodied in its probability distribution, which may be presumed given in any specific problem.[1] Then, in principle, the distribution density functions $p(y|x)$ and $p(y)$ may be derived from the state and error distributions. In processing data, $p(y|x)$ and $p(y)$ are evaluated for the specific data obtained, and (1) is then used to "update" $p(x)$. Some of the mathematical details of this procedure will be given in a later section.

The updating of the distribution of x to include the new information contained in y constitutes the most basic principle of statistical inference. Because the posterior distribution contains all the knowledge, both old and new, that we have regarding x, the determination of this distribution via equation (1) is referred to as the *general estimation problem* (ref. 6). When a decision is to be made on the basis of this totality of information another step is required which involves the use of an optimality criterion. The specification of such criteria is discussed in the next section.

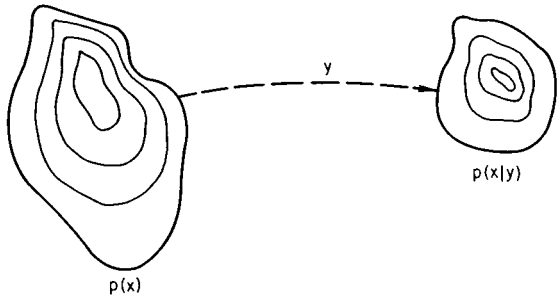

Figure 2 - Transition from prior to posterior distribution

Optimality Criteria

Before one can answer questions regarding how best to use data for drawing inferences, one must define what is meant by "best" and translate the concept of "goodness" into the mathematical language of estimation and decision problems.

[1] It is assumed here that x and e are statistically independent. If they are not, the joint density $p(x,e)$ must be used instead of the separate $p(x)$ and $p(e)$ functions.

The principle of least squares was perhaps the earliest development in the concept of optimality. It was originally developed as a means of solving two types of problems. Problems of the first type arise when repeated measurements of some phenomenon do not always give the same result, and one wishes to remove the apparent inconsistencies in the data, such as in the determination of the orbits of planets from a large number of line-of-sight observations. Problems of the second type are simply curve-fitting situations (or regression analysis), where it is desired to summarize data in the more manageable form of simple functional relationships (often linear) or at least in the form of a smooth curve. In these applications "best" means that the solution minimizes the sum of squares of deviations between data points and corresponding points derived from the solution.

The difficulty with least squares is that its use is not based on a clear-cut optimality principle but rather on the fact that its use leads to simple linear operations upon the data. There is no reason per se why a "better" solution might not be obtained by minimizing, for instance, the sum of absolute values of the deviations. Of course, it can be argued that the latter approach requires far more difficult computations, so that "best" in the broad sense should take this into account and assign a plus value to the least squares procedure.

A more precise definition of optimality is obtained by invoking the concept of the cost, l, associated with operations upon data. The principle involved is to design one's *modus operandi* for handling data (either for forming an estimate or for making an action decision) so as to minimize the average (or expected) cost, $E(l)$, of a single operation. A *modus operandi* so selected is optimal in the sense that the total cost of a large number of operations of this type would in all probability be less than for any other *modus operandi* of the same class.

Optimality is now dependent upon the definition of l for the particular application one has in mind. For instance, consider the problem of designing an operational system for tracking a space vehicle, determining its trajectory, and providing control over in-flight maneuver, possible abort, reentry, and recovery aspects of the mission. Such factors as the accuracy of the estimate, the speed with which it may be obtained, reliability and cost of complex tracking and computing equipment, the flexibility of recovery forces, etc., are all significant and may be combined in a variety of ways to define an optimal system. The weighting applied to each factor must be decided by engineers with, at best, an imperfect knowledge of the true importance of each factor. Obviously, one is faced here with the same difficulty encountered in specifying prior probabilities--namely, the conversion of subjective quantities into numerical or functional form. Thus, the practical engineer, knowing that he can only approximate, is justified in arbitrarily choosing cost representations which are mathematically tractable. Of course, any system so designed is optimal only with regard to the more or less arbitrary criterion so established.

This situation has led to a dominant use of quadratic type criteria, whose tractability is indisputable. Such criteria call for the data operations to minimize functionals of the form

$$J = E\left(\xi^T W \xi\right) \tag{2}$$

where W is a positive semidefinite weighting matrix, and ξ is usually a zero mean random variable defined appropriately for the problem. For instance, in the comparatively simple problem of pure estimation (i.e., where the only significant cost is due to inaccuracy in the estimate), ξ would be the estimation error, $\xi = \tilde{x} = x - \hat{x}$, so that J is a measure of

the concentration of estimates $\hat{x}$ about the true value x. Indeed, J is a linear function of the second-order moments (covariances) of the $\tilde{x}$ distribution, hence the name "minimum variance" applied to one of the estimation techniques which uses this type of optimality criterion.

Whatever objections may be raised to this pre-eminent use of quadratic criteria (which reduce to the familiar mean-square error criterion for scalar x), we are faced with the fact that most of the useful analytical methods are based on such criteria. Specialized analytical results occasionally have been obtained for other criteria (refs. 7 and 8), and there has been at least one successful attempt to generalize the minimum variance solution to a broader class of criteria (refs. 6 and 9) which will be discussed later. Beyond this, to utilize an arbitrary criterion one would be forced to use numerical methods and a large scale computer.

THE MATHEMATICS OF STATISTICAL INFERENCE

In the previous sections we have identified the basic principles of the universal notion of statistical inference. We now proceed to a review of some of the explicit mathematical methods of statistical inference. Emphasis is on showing the relation of each method to the other methods and to the general theory. It is assumed in all of the following that probability distributions, cost functions, and functional relationships between variables are given when needed--that is, the problem of translating scientific questions into suitable mathematical language has been surmounted.

Least Squares

The two types of problems ordinarily treated by least squares were described in the previous section. These problems are actually not distinct if one adopts the proper point of view, a conclusion supported by the results developed below.

Reconciliation of inconsistent data.--In this case it is assumed that the body of data to be analyzed consists of "noisy" measurements of some physical, economic, biologic, etc., phenomenon. The problem is to find an estimate of the underlying phenomenon which is consistent with the data and matches it best in a least-squares sense, under the assumption that the functional relation is known between the unknown quantities and the observables being measured. That is, the mathematical model of the system is given as

$$y_i = g(x) + e_i, \qquad i = 1, 2, \ldots N$$

where the y_i are the measurements, x is the unknown quantity, e_i is the "noise" or error in the ith measurement, and g is a known function.

It is desired to determine that value of x for which the quantity

$$J = \sum_{i=1}^{N} [y_i - g(x)]^2 \tag{3}$$

is a minimum. If y_i is a vector (i.e., several different quantities measured), then equation (3) should be written as

$$J = \sum_{i=1}^{N} [y_i - g(x)]^T [y_i - g(x)] \tag{4}$$

Of course, the unknown x may also be a vector.

To determine the minimum, equation (4) is differentiated with respect to x, obtaining

$$\frac{\partial J}{\partial x} = -2 \sum_{1}^{N} \left(\frac{\partial g}{\partial x}\right)^T [y_i - g(x)] \tag{5}$$

For J to be minimum, equation (5) is set equal to zero; then

$$N\left(\frac{\partial g}{\partial x}\right)^T g(x) = \left(\frac{\partial g}{\partial x}\right)^T \sum_{1}^{N} y_i \tag{6}$$

It is not possible to write an explicit solution for the minimizing x from this equation unless an analytic expression for g is given; even then the set of equations is generally nonlinear, hence often difficult to solve. However, in the common situation in which g is linear (e.g., $g(x) = Ax$) the solution is straightforward. Since in this case $\partial g/\partial x = A$, equation (6) reduces to

$$N(A^T A)\, x = A^T \sum_{1}^{N} y_i \tag{7}$$

and the least-squares estimate of x is

$$\hat{x} = (A^T A)^{-1} A^T \cdot \frac{1}{N} \sum_{1}^{N} y_i \tag{8}$$

where it is seen that $\frac{1}{N} \sum_{1}^{N} y_i$ is simply the numerical average of the measurements.

The above solution is by no means the most general which can be handled by least squares. For instance, very commonly one has situations in which the functional relation g between the state and observables may be different (although still assumed known) for each measurement. The direct way to treat such a case is to consider the entire body of data as a single vector observation, say y. Then the system equation is

$$y = g(x) + e \tag{9}$$

In the linear case $y = Ax + e$. The equation to be solved is

$$(A^T A)x = A^T y \tag{10}$$

and the least-squares solution is simply

$$\hat{x} = (A^T A)^{-1} A^T y \tag{11}$$

The equations represented in matrix form in equations (7) and (10) are the so-called normal equations. It should be noted that they do not necessarily have a unique solution, depending upon whether or not the normal matrix $A^T A$ is singular. Even so, a least-squares "best" estimate can always be obtained by means of the generalized inverse (see ref. 10).

Curve-fitting and regression analysis.--Here one is given a set of data, such as that shown in figure 3, which represents the simultaneous measurement of related quantities x and y. The object is to find a function, or curve, which best fits the data in a least-square sense. This is the usual regression or curve-fitting problem in its simplest form.

If the function for fitting the data is designated as $y = g(x)$, then the least-squares principle requires that we pick g to minimize the quantity

$$J = \sum_{1}^{N} [y_i - g(x_i)]^2 \tag{12}$$

where N is the number of data points. Generalization of the problem to include a number of variables is easily made by regarding y_i and x_i both as vectors (not necessarily of the same dimension). In fact, just as before, the entire body of data can be represented by the composite vectors x and y, and the quantity to be minimized is represented by

$$J = [y - g(x)]^T [y - g(x)] \tag{13}$$

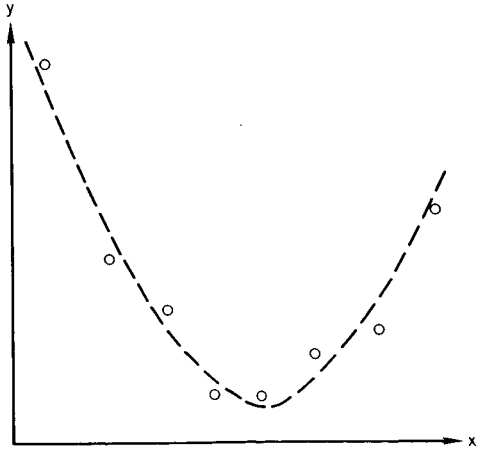

Figure 3 - Least squares curve fitting

Now, it is easy to see, using figure 3 as an example, that the curve-fitting problem makes sense only if smoothing is employed. That is, obviously the curve consisting of straight line segments connecting the data points make J zero, and no better fit can be found. But this does not effect any condensation of the data whatsoever and is thus pointless. What we need is to restrict our choice of function to a class that can be described by a number of parameters smaller than N--preferably enough smaller to obtain a significant condensation. This difference between the "dimensions" of the data and of the fitting function is often referred to as the "degrees of freedom." As the number of degrees of freedom is increased the minimum obtainable J generally increases.

Presuming that a functional form for g has been selected (or perhaps is given as part of the problem statement), describable by a certain number of parameters which are to be determined, the minimization procedure can proceed. Here we will consider only the case of simple linear regression, that is, g is linear in x, or $g(x) = Ax$. (It is assumed here, without loss of generality, that the data have been translated, so that the origin is at the point $[(1/N)\Sigma y_i,\ (1/N)\Sigma y_i]$ so as to simplify the mathematical expressions which follow.) In this case we have

$$J = \sum_{1}^{N} [y_i - Ax_i]^T [y_i - Ax_i] \tag{14}$$

where it is assumed that y_i is an m vector (i.e., has m components), and x_i is an n vector; A is therefore an $m \times n$ matrix of unknown regression coefficients which are to be determined to minimize J. The minimization is achieved in the usual way by differentiating J with respect to A and setting the result equal to zero, where $\partial J/\partial A$ is defined as the matrix

$$\frac{\partial J}{\partial A} = \begin{bmatrix} \frac{\partial J}{\partial a_{11}} & \frac{\partial J}{\partial a_{12}} \cdots\cdots \frac{\partial J}{\partial a_{1n}} \\ \\ \frac{\partial J}{\partial a_{m1}} \cdots\cdots\cdots\cdots \frac{\partial J}{\partial a_{mn}} \end{bmatrix} \tag{15}$$

Following this procedure, we obtain

$$\frac{\partial J}{\partial A} = -2 \sum_{1}^{N} (y_i - Ax_i) x_i^T = 0 \tag{16}$$

and from this we must have

$$A \sum_{1}^{N} x_i x_i^T = \sum_{1}^{N} y_i x_i^T \tag{17}$$

Note that to obtain a unique solution for A the matrix $\sum_{1}^{N} x_i x_i^T$ must be nonsingular. This requires that N, the number of data points, must be at least as great as the dimensionality of the x_i.

If the vectors x_i and y_i are arrayed in columns to form matrices as follows,

$$\left.\begin{aligned} X &= \begin{bmatrix} x_{11} & x_{21} & \cdots & x_{N_1} \\ x_{12} & x_{22} & & \vdots \\ \vdots & \vdots & & \vdots \\ x_{1n} & x_{2n} & & x_{Nn} \end{bmatrix} \\ Y &= \begin{bmatrix} y_{11} & \cdots & y_{N1} \\ \vdots & & \vdots \\ y_{1m} & & y_{Nm} \end{bmatrix} \end{aligned}\right\} \tag{18}$$

then equation (17) can be rewritten in the matrix form

$$AXX^T = YX^T \tag{19}$$

or

$$(XX^T)\,A^T = XY^T \tag{20}$$

and

$$A^T = (XX^T)^{-1}\,XY^T \tag{21}$$

Note the similarity of equation (20) to the normal equation (10) for the point estimation problem: The roles of the independent variables and the regression coefficients are simply reversed. That is, in the regression problem the X and Y are given, and we are to estimate A, whereas in the point estimation problem A and y are given and we are to estimate x.

Reference 11 (pp. 179-183) develops the result (eq. (21)) utilizing the principle of maximum likelihood and the assumption that the data points (x_i, y_i) come from a multivariate normal distribution, and reference 12 (pp. 577-579) shows how to obtain the same result by means of the Gauss-Markov theorem. However, as has been shown here, the least-squares method does not depend on any probabilistic statements.

As a critique of least squares it should be pointed out that the method is relatively simple in that it is independent of probabilistic statements regarding the observation error, e. (It may be argued that the criterion of minimization of the sum of squares of

residuals implies certain statistical properties of e in that only when e has such properties is such a criterion reasonable, but this is not equivalent to a direct *a priori* assumption regarding the distribution of e.) However, the lack of probabilistic statements regarding e carries with it the disadvantage that it is impossible to make any probabilistic statements about the least-squares estimate. For instance, from equation (11) it can be seen that the error in estimate $\tilde{x} = x - \hat{x}$ in the linear case is proportional to e:

$$\tilde{x} = -(A^T A)^{-1} A^T e \tag{22}$$

and without any knowledge of e nothing can be said about $\tilde{x}$. One can, of course, make certain *ex post facto* assessments, such as that if e has zero mean ($E(e) = 0$) and the e are independent, then $\tilde{x}$ has zero mean and the least-squares estimate is "unbiased." If, further, e has the covariance $E(ee^T) = Q = \sigma^2 I$, the covariance of $\tilde{x}$ is

$$E(\tilde{x}\tilde{x}^T) = \sigma^2 (A^T A)^{-1} \tag{23}$$

which provides a measure of the precision of the estimate. There is the additional possibility, for purposes of error analysis, of examining the residual, $[y - g(\hat{x})]$, which constitutes an estimate of e, and from this inferring the distribution of e. However, even this if it is to be done rationally requires some assumptions as to the character of e which are not given in the pure least-squares problem.

Minimum Variance and Weighted Least Squares

The so-called minimum variance estimate is usually associated with the Gauss-Markov theorem (see, e.g., refs. 12 and 13), which may be stated as follows:

> Suppose we have a set of observations y, described as in the previous section by
>
> $$y = Ax + e \tag{24}$$
>
> The matrix A is assumed known; the set of parameters represented by the vector x is unknown and is to be estimated; and e is the set of observation errors with covariance matrix $E(ee^T) = Q$.
>
> The "best" estimate of x in the sense of being that estimate having minimum variance from the class of all linear unbiased estimates, is given by the solution of the normal equation
>
> $$A^T Q^{-1} A x = A^T Q^{-1} y \tag{25}$$
>
> or
>
> $$\hat{x} = (A^T Q^{-1} A)^{-1} A^T Q^{-1} y \tag{26}$$

It can be readily shown that this same equation may be obtained by minimizing in the same manner as in the last section the functional

$$J = [y - Ax]^T Q^{-1} [y - Ax] \tag{27}$$

which is called the "weighted" sum of squares because each element of the sum in J is multiplied by one of the elements of Q^{-1}. Thus, the minimum variance estimate may also be called a weighted least-squares estimate.

The precision of the minimum variance estimate can be expressed easily in terms of the covariance

$$E(\tilde{x}\tilde{x}^T) = (A^T Q^{-1} A)^{-1} \tag{28}$$

which is obtained by substituting equation (24) into equation (26), forming the difference $\tilde{x} = x - \hat{x}$, forming the matrix product $\tilde{x}\tilde{x}^T$, and taking the expectation of the result.

The relation of this estimate to the least-squares estimate of the previous section is easily seen. If Q is a diagonal matrix (i.e., errors in all of the observation components are uncorrelated one with another) and all errors have equal variance σ^2 then $Q = \sigma^2 I$. Substituting this into equation (26), we obtain

$$\hat{x} = (A^T A)^{-1} A^T y \tag{29}$$

which is precisely the least-squares estimate.

Clearly, the linear least-squares estimate can be regarded as a special case of the minimum variance estimate. Many times in applications the least-squares estimate is used simply because it is easier to compute than the minimum variance estimate or because there is some doubt about the correct Q matrix (ref. 13, p. 21). In such a case the method of least squares will still give an unbiased estimate, although it will not be the best obtainable. The loss in accuracy resulting from using least squares instead of minimum variance is analyzed in reference 14.

Note that only linear estimates have been considered here, although the results can also be applied to nonlinear systems which can be successfully linearized (i.e., accurately approximated by linear systems). It also may be noted that only the mean (assumed zero here) and variance of e are given. Other properties of the distribution are immaterial because consideration is limited to linear systems and linear estimates and because the criterion is minimization of the variance of the estimate.

It is seen that with only the first and second moments of the distribution of e given, only the corresponding moments of the distribution of $\hat{x}$ can be obtained. In many practical problems these are adequate, but if more general results are required one must utilize one of the more general methods to be described later.

Maximum Likelihood

The principle of maximum likelihood for use in parameter estimation was introduced by R. A. Fisher in 1912 and further expanded by him in a series of papers (see, e.g., ref. 15).

The basic assumption is that the observations are random variables drawn from an infinite population whose distribution is known except for a finite number of parameters. A simple example is drawing black and white balls from an urn; the distribution is known to be binomial, but the proportion of the black to white balls is not known and must be estimated from the samples (i.e., observations).

Another elementary example is measuring some physical quantity (such as the speed of light) with an apparatus known to introduce normally distributed zero mean errors. The measurements can be regarded as derived from a normally distributed population with an unknown mean which is the physical quantity under investigation. By estimating the mean of this population one obtains an estimate of the unknown quantity.

The principle of maximum likelihood estimation is to pick as estimates of the unknown parameters those values for which the set of observations would be "most likely" to occur, where "most likely" is defined to mean maximization of the so-called likelihood function.

The likelihood function is nothing more than the conditional density function of the observations given the unknown parameters and, as such, can be written as $p(y|\alpha)$, where y represents the observation and α the parameters to be estimated. In some of the literature the notation is $p(y;\alpha)$, which means simply the density function of the observations; the α is included in the argument to indicate the dependence of the distribution upon the unknown parameters. The effect is the same in either case, although the conditional distribution notation tends to imply that α is a random variable in its own right, and sometimes there are objections to this on philosophical grounds since one generally is dealing in such problems with only one specific (unknown) value of α. However, it scarcely matters in maximum likelihood whether α is or is not a random variable since its "distribution," if any, does not enter into the ML solution.

The ML solution consists in finding that value of α for which the probability $p(y|\alpha)dy$ is a maximum, where $p(y|\alpha)dy$ is the probability that the observations y will lie in the hypercube dy. Since only the density function and not dy is a function of α, we are concerned only with the maximum value of $p(y|\alpha)$, which with y given is a function only of α.

Examples of using the ML technique are given in many texts (e.g., see ref. 16 for some elementary cases, and ref. 17 for a nonlinear problem). Here, we will develop only one case, involving the multivariate normal distribution, which serves to illustrate the principles and which we will use later for comparison with other estimation methods.

First, assume observations are generated by the linear process

$$y_i = A_i x + e_i \tag{30}$$

where A_i is known, and the y_i and e_i are, respectively, observations and observation errors. It is further assumed that the e_i are normally distributed with zero mean. The covariances $E\left[e_i e_i^T\right]$ can be assumed either known partly or unknown. In the latter case these are to be estimated from the data along with estimating the unknown x, but in general, there cannot be a greater number of unknowns than there are observations. Note that A_i may be time-varying--that is, it may be different for each value of index i.

To put the problem in its simplest form, the set of observations is combined into a single vector variable y:

$$\begin{Bmatrix} y_i \\ \cdot \\ \cdot \\ \cdot \\ y_N \end{Bmatrix} = \begin{bmatrix} A_i \\ \cdot \\ \cdot \\ \cdot \\ A_N \end{bmatrix} x + \begin{Bmatrix} e_i \\ \cdot \\ \cdot \\ \cdot \\ e_N \end{Bmatrix} \tag{31}$$

or

$$y = Ax + e$$

Because of the normal distribution assumption, the likelihood function can be written (ref. 13, p. 416)

$$p(y|\alpha) = \frac{1}{[(2\pi)^N |Q|]^{1/2}} \exp \left[-\frac{1}{2} (y - Ax)^T Q^{-1} (y - Ax) \right] \tag{32}$$

where Ax is the mean of the distribution, and Q is the covariance matrix $E[ee^T]$. In the most general problem y represents the given data and α the set of elements of x and Q which are to be estimated.

To maximize p, one can just as well maximize $\log p$:

$$\log p = -\frac{N}{2} \log 2\pi - \frac{1}{2} \log |Q| - \frac{1}{2} \left[(y - Ax)^T Q^{-1} (y - Ax) \right] \tag{33}$$

It is seen immediately by comparing the above with the functional (27), which is minimized in weighted least squares, that the maximizing x is the solution of

$$A^T Q^{-1} Ax = A^T Q^{-1} y \tag{34}$$

which is precisely the normal equation of the minimum variance or weighted least-squares method. If Q is known, then the solution requires only the estimate of x:

$$\hat{x} = (A^T Q^{-1} A)^{-1} A^T Q^{-1} y \tag{35}$$

For the more general solution when some of the elements of Q are also assumed unknown, one needs to find the x and Q which simultaneously maximize $\log p$. Some of the problems involved in doing this and some methods of solution are given in reference 11. The results are only incidental to the main theme of this paper.

We have given above the formal statement and an example of the maximum likelihood method, which for the case considered is seen to give precisely the same result as minimum variance estimation in the linear normally distributed case. It remains to discuss some of the optimality aspects of the method.

Apparently, the statement of the ML principle involves only the intuitive idea that a "best" estimate should lie somewhere near the peak (or maximum point) of the conditional density function, $p(y|\alpha)$. Once this principle has been accepted, one then seeks to determine just how good is such an estimate. It has been shown (refs. 18, 6) that for a regular[2] estimation case of the continuous type, the mean square deviation of the estimate from the true value satisfies the Cramér-Rao inequality:

$$E(\hat{\alpha} - \alpha)(\hat{\alpha} - \alpha)^T - \left\{ E\left[\frac{\partial^2 p(y|\alpha)}{\partial \alpha_i \partial a_j} \right] \right\}^{-1} \geq 0 \tag{36}$$

[2] The condition of regularity is satisfied in most common applications and will not be dealt with here.

where the α_i, α_j are elements of the vector of unknowns, α. That is, the symmetric matrix on the left of equation (36) is positive semi-definite, and no estimates can have a covariance matrix smaller than the second term on the left. If for a particular estimate the sign of equality holds, then this is said to be an *efficient* estimate, and clearly no estimate can have a smaller covariance.

The argument in favor of maximum likelihood is an indirect one to the effect that *if an efficient estimate exists*--that is, an estimate which achieves the smallest possible covariance matrix--then it can always be formed by the method of maximum likelihood. The difficulty is that efficient estimates exist only under fairly restrictive circumstances. (The necessary and sufficient conditions are given in ref. 18.) When there is no efficient estimate, then no method of estimation can achieve the lower bound and there is no guarantee that some estimate other than ML might not be better. In fact, generally in such cases there *is* a better estimate than ML. Even so, the ML method is less complicated to apply in many circumstances and may be used for this reason even when it is known not to be the best.

Some remarks regarding the basic hypotheses of ML estimation are necessary for a proper perspective of the method. It will be noted that a fundamental premise of the ML theory is that the distribution form is known with only a finite number of its parameters to be determined. In view of the earlier discussion on probability, it is seen that this implies a certain amount of "perfect" prior knowledge. In fact, an explicit mathematical expression for the likelihood function is required before we can proceed at all.

Another basic premise of ML estimation is that there is no prior knowledge regarding the values of the unknown parameters, α. The validity of this assumption, of course, depends on the particular problem and involves philosophical considerations which have already been dealt with at some length.

Despite possible shortcomings, ML is clearly a more versatile method than the least squares and minimum variance techniques described in the previous sections because it permits the use of a complete description of the distribution of the observation errors.

Decision Theory and Bayes' Estimation

The development given in this section is based on a decision theoretic point of view which will be seen to be general enough to encompass nearly all scientific inference problems, including the special class arising in control. When applied to estimation, this approach is often termed Bayes' estimation after the famous Bayes' Theorem which plays a central role in the development.

The decision problem.--Let x represent the set of parameters required for making a decision, δ. Suppose further that x may take on a range of different values with probabilities describable by means of the density function, $p(x)$. Observations dependent upon x and subject to errors e are to be made:

$$y = h(x,e) \tag{37}$$

It is assumed that the probabilities of the errors are known for all possible values of x, so that we have the conditional density function (or likelihood function) $p(y|x)$.

Now, to make intelligent decisions, it is necessary to know which decisions are bad and which are good for us. That is, all possible decisions under all possible circumstances must be ranked and, further, assigned specific numerical values. The result is a function, l, dependent upon the true (unknown) state x and upon the decision, δ:

$$l = l(x,\delta)$$

This is called a cost or payoff or loss function, and its meaning is that if a value $x = x_a$ obtains, and the decision $\delta = \delta_a$ is made, then the cost or loss is $l(x_a,\delta_a)$.

An optimum decision strategy may now be defined as that strategy which on the average costs the least. That is, one wishes to minimize the average loss, $E[l(x,\delta)]$, over all possible x, and the optimum decision is

$$\delta_{opt} = \arg \min_{\delta} E_x[l(x,\delta)] \tag{38}$$

Of course, the decision will be based on the observations, y - that is, $\delta = \delta(y)$.

To compute the expectation in equation (38), one needs the probability distribution of x given the observations, which may be represented (for a continuous distribution) by $p(x|y)$. Then equation (38) becomes

$$\delta_{opt} = \arg \min_{\delta} \int_{\Omega_x} l(x,\delta)p(x|y)\,dx \tag{39}$$

In the usual problem $l(x,\delta)$ is given, but $p(x|y)$ must be computed as follows. Presumably the functional relationship $y = h(x,e)$ (where e is the observation error) and the joint density function $p(x,e)$ are given. These imply the knowledge of the joint distribution of x and y as given by $p(x,y)$, and also the distribution of y itself which is merely the marginal distribution of (x,y). Then $p(x|y)$ may be written as

$$p(x|y) = \frac{p(x,y)}{p(y)} \tag{40}$$

This is as far as the solution can be carried in very general terms. Any further result requires specific loss and density functions. If such are given in numerical or tabular or curve form, the calculation would have to proceed by numerical methods.

If we are so fortunate as to have an l and a $p(x|y)$ of certain forms, further general results can be obtained. Consider first the case where the functions $h(x,e)$, $p_1(x,e)$,[3] and $p(x)$ are of mathematically tractable form. If the inverse function $e = h^*(x,y)$ can be obtained, then $p_2(x,y)$ and $p(y)$ are given by (ref. 19):

$$\left.\begin{aligned} p_2(x,y) &= p_1[x,h^*(x,y)]\det\left[\frac{\partial h}{\partial e}\right] \\ p(y) &= \int_{\Omega_x} p_2(x,y)\,dx \end{aligned}\right\} \tag{41}$$

[3] Subscripts are used here on p to avoid ambiguity when different functional forms are involved.

These can then be used in equation (40). Thus, in certain cases an explicit mathematical expression can be obtained for $p(x|y)$ for use in equation (39).

Another potentially useful analytical result has been shown by Sherman (ref. 9) to apply for certain classes of loss and density functions. If l can be expressed as a function of a single argument ν such that (1) $0 \leq l(\nu) = l(-\nu)$ (i.e., l symmetric in ν) and (2) $0 \leq \nu_1 \leq \nu_2 \rightarrow l(\nu_1) \leq l(\nu_2)$ (i.e., l wide-sense monotonic increasing in the distance from the origin[4]), and if $p(x|y)$ for the given y is unimodal (i.e., has only one local maximum) and symmetric about the mean, μ, then the optimum δ is obtained by choosing δ to minimize $l(x,\delta)$. For instance, suppose $l = |x - d^*(\delta)|$, where d^* is the inverse of a function d (i.e., $d[d^*(\delta)] = \delta$). Then its minimum value occurs at $\delta = d(x)$, and the optimum decision is $\delta_{opt} = d(\mu)$. The essence of this result is that under the proper conditions, the mean of the posterior distribution furnishes all the information needed to obtain the optimum decision.

A geometrical picture is helpful to illustrate this interesting result. Figure 4 shows a loss function of the type considered. For the illustration, a problem which is single-dimensional in both x and δ is employed; that is, a scalar state and scalar decision are assumed. The loss, being a function of both x and δ is represented by a surface whose height above the x, δ plane is l. The particular form of l used here is a function symmetric in the argument $[x - d^*(\delta)]$, and monotone increasing as the magnitude of the argument increases. Thus, each cross section of l parallel to the x,l plane is the same except for a translation in the x direction by an amount $d^*(\delta)$. The appearance

Figure 4 - Example loss function

[4] Any non-negative real-valued convex function of the argument can serve as the "distance" in this application. (See ref. 6, p. 35.)

is that of a river valley whose floor is level and traces the curve $\delta = d(x)$ in the x,δ plane.

Figure 5 shows a typical symmetrical unimodal density function with a maximum at $x = \mu$, plotted on the same x, δ base grid. Since it is not a function of δ every cross section is the same. The comparison with a uniform mountain ridge is irresistible.

Now if the product lp (the integrand of 39) is plotted, its cross sections have the appearance shown in Figure 6 for various δ. The area under each curve is the integral $\int lp\, dx$ for a specific δ. The minimum value of this integral occurs for that value of δ for which the peak of p (i.e., $x = \mu$) and the valley of l (i.e., $\delta = d(x)$) occur at the same value of x. Hence, this is the optimum δ, that is, $\delta_{opt} = d(\mu)$.

This, of course, is just an illustration. For further details see reference 9.

The principal thought in the special case just illustrated is that sometimes the optimum decision is a known function of the conditional mean (or median or mode in this case, since the three are coincident for symmetric unimodal distributions). This means that in such a case, given y, all one needs to do is compute the conditional mean of x. Although the special conditions required for this result cannot be expected to occur commonly, it may be noted that often the conditions may be a respectable approximation to the true situation, particularly since the specification of distributions and loss functions is somewhat arbitrary anyway.

It is readily seen that the special case of a quadratic loss function where the loss is quadratic in both x and δ,

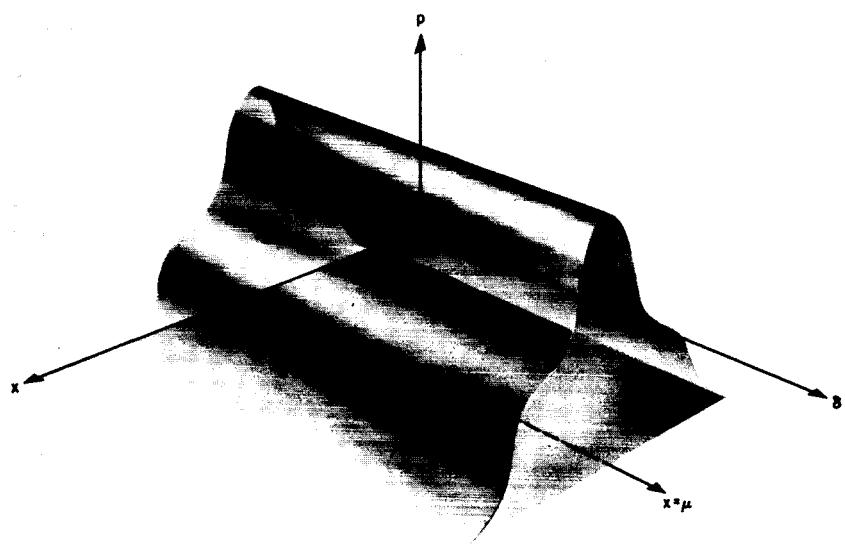

Figure 5 - Example symmetrical unimodal loss function

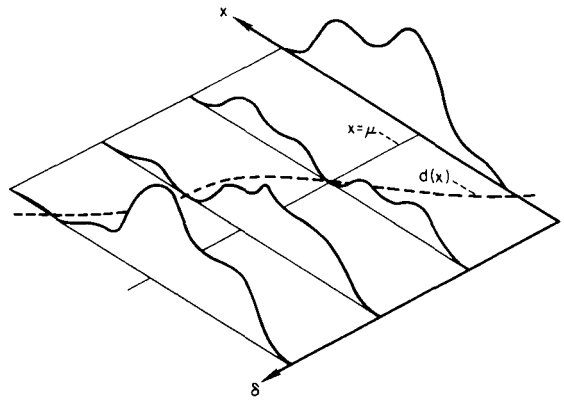

Figure 6 - Cross sections of example lp function

$$l = [Bx - \delta]^T C[Bx - \delta] \tag{42}$$

belongs to the class of loss functions described above. In this case it can be seen that by substituting equation (42) into equation (38) and carrying out the indicated minimization the optimum decision is given by

$$\left.\begin{aligned} \delta_{opt} &= B \int_{\Omega_x} x p(x|y)dx \\ &= BE_x(x|y) \end{aligned}\right\} \tag{43}$$

Thus, when the loss function is quadratic the optimum decision is a linear function of the conditional expectation, regardless of the form of the conditional distribution. This, of course, reduces to the same result as the previous case when $p(x|y)$ is symmetric and unimodal, for then the mean is the same as the median and mode.

Bayes' estimation.--The general decision theory solution given above is easily specialized to one of pure estimation by defining the "decision" as one of simply determining, as accurately as possible, the true state x. The δ then becomes $\hat{x}$ (i.e., estimate of x), and the optimum estimate is given by

$$\hat{x}_{opt} = \arg \min_{\hat{x}} \int_{\Omega_x} l(x,\hat{x}) p(x|y)dx \tag{44}$$

To get some specific results for comparison with previously described estimates, we may first apply Sherman's result so that for a symmetric nondecreasing l whose argument is $(x - \hat{x})$, and a symmetric unimodal $p(x|y)$, we obtain

$$\hat{x}_{opt} = E_x[x|y] \tag{45}$$

That is, the optimum estimate is simply the mean (or median or mode) of the conditional (or posterior) distribution.

Three interesting points can be made regarding this result: First, under the special conditions for which equations (43) and (45) apply, the general decision problem separates quite naturally into independent estimation and decision parts. This is seen by comparing equations (43) and (45). Under the proper conditions on l and p, one may first estimate the true state of affairs without regard to what is to be done with the estimate, and then use the estimate to take some course of action just as though the estimate were actually the true state

$$\left.\begin{aligned} \hat{x}_{opt} &= \underset{x}{E}[x \mid y] \\ \delta_{opt} &= B\hat{x}_{opt} \end{aligned}\right\} \tag{46}$$

This fact is of considerable value in a number of areas, particularly control applications, because the separation of a complicated problem into simpler parts can result in a system less difficult to implement.

In control theory the legitimacy of separation has been proved for linear systems with quadratic optimality criteria and normally distributed errors (see refs. 20 and 21). However, the separation property should apply to all decision problems (as is indicated in ref. 5, pp. 180-1), if one can define an "imputed estimation loss" function which properly reflects the loss incurred in making a decision based on an incorrect estimate. In other words, the use of the estimate determines the proper estimation loss function. As a limited example of this fact consider a simple control problem in which the state x of a system is to be driven to zero by means of a control δ which has an effect on the state $\Delta x = d^*(\delta)$. The "miss" resulting from applying the control δ is $x' = x - \Delta x = x - d^*(\delta)$, and clearly the optimum control is $\delta = d(x)$. However, when x is unknown, zero miss cannot be assured, and one must pick a δ which is best in the sense of minimizing the average value of a suitable loss function. Suppose that the appropriate loss function is $l_s[x'] = l_s[x - d^*(\delta)]$, where l_s is a function of the Sherman class. If the distribution of x is symmetric and unimodal, then the optimum control is $\delta = d(\mu)$, where μ is the mode of x. Now, if one were to consider the pure estimation problem of determining x with an estimation loss $l_s[\tilde{x}] = l_s[x - \hat{x}]$, it is clear that the optimum estimate would be $\hat{x}_{opt} = \mu$. Thus, we see that this control problem can be decomposed into an estimation problem, followed by deterministic control $\delta_{opt} = d(\hat{x}_{opt})$. It may be noted that in this case the estimation loss function need not even be of the same form as the decision loss function so long as it is of the Sherman class. This is because of the assumption of a symmetric and unimodal $p(x \mid y)$. In the case of a more general distribution, such freedom should not be expected.

The second point to be made regarding the result (45) is that it is reminiscent of the basic concept of maximum likelihood estimation; to wit, the choice of an estimate to maximize the likelihood function $p(y \mid x)$. In equation (45) the estimate maximizes the posterior density function $p(x \mid y)$, which is more or less different from $p(y \mid x)$ depending upon the "sharpness" or concentration of the distribution $p(x)$. Because of this similarity, the estimate which maximizes $p(x \mid y)$ is sometimes called a maximum likelihood estimate in the Bayesian sense; it is also called a most probable estimate for obvious semantic reasons (see ref. 22).

The third observation which can be made regarding the result (45) is that for some distribution forms equation (45) can be written as an explicit formula for the estimate. Of particular interest is the case of a normal distribution.

Now, for $p(x|y)$ to be normal, both $p(x)$ and $p(y|x)$ must be normal, since by Bayes' rule

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)} \tag{47}$$

If we now consider as before that the observations y are generated by the linear relationship

$$y = Ax + e \tag{48}$$

then the posterior (or conditional) density function takes the form

$$p(x|y) = D \exp\left\{-\frac{1}{2}\left[(y - Ax)^T Q^{-1} (y - Ax) + x^T P^{-1} x - y^T (Q + APA^T)^{-1} y\right]\right\} \tag{49}$$

where Q and P are the covariance matrices of e and x, respectively, and D is the multiplicative constant

$$D = \frac{|Q + APA^T|^{1/2}}{[(2\pi)^p |P| \ |Q|]^{1/2}} \tag{50}$$

(Here, p is the number of elements of the vector x.) With a bit of manipulation equation (49) can be rewritten as

$$p(x|y) = D \exp\left[-\frac{1}{2}(x - \Lambda A^T Q^{-1} y)^T \Lambda^{-1} (x - \Lambda A^T Q^{-1} y)\right] \tag{51}$$

where

$$\Lambda = (P^{-1} + A^T Q^{-1} A)^{-1}$$

Obviously, the mean (or mode or median) of this distribution is $\mu = \Lambda A^T Q^{-1} y$. Thus, in this case the optimum estimate can be written as the explicit form

$$\hat{x}_{opt} = (P^{-1} + A^T Q^{-1} A)^{-1} A^T Q^{-1} y \tag{52}$$

and the variance of the estimate is given by

$$\Lambda = (P^{-1} + A^T Q^{-1} A)^{-1} \tag{53}$$

Comparison of this solution with that obtained by the minimum variance and ML methods under the same assumptions of normal distributions and a linear system, namely,

$$\hat{x} = (A^T Q^{-1} A)^{-1} A^T Q^{-1} y \tag{54}$$

reveals immediately that the only difference is the addition of P^{-1} to the $A^T Q^{-1} A$ matrix. Since P is the prior covariance matrix of x, we see that the difference between the minimum variance and Bayes' estimation methods is that the latter makes use of prior

information whereas the former does not. More to the point, the minimum variance (or ML method) pointedly omits any reference to prior information, while Bayes' estimation includes it in the problem statement.

An interesting point is that extremely poor prior information implies a large P matrix, so that P^{-1} is very small compared to $A^T Q^{-1} A$, and for practical purposes could be disregarded in the estimation equation (52). In such a case, the Bayes' estimate differs only inconsequentially from that obtained by the minimum variance method. Thus, the two results are consistent, provided the same assumptions regarding system linearity, normality, prior information, and optimality criterion are used in the two cases.

It should be noted that the special Bayes' estimate (52) requires a loss function of a particular class, to which the quadratic form belongs; that the minimum variance method by its very name implies a quadratic loss function; and that the maximum likelihood method is optimum only if an efficient estimate exists, which also implies a quadratic criterion.

Obviously Bayesian estimation is the most general of all the methods in that it (1) permits consideration of the broadest class of criteria, (2) makes use of a complete description of the prior distributions of both e and x, and (3) yields a complete description of the posterior distribution. All other estimation methods can be regarded as special cases.

Smoothing, Filtering, and Prediction

Suppose there exists a set of unknown parameters, x, having the value $x(0)$ at time t_0 and changing over the course of time (i.e., x is a stochastic sequence). Suppose further that observations relative to x are made at times $t_1, \ldots, t_N$, at which times x has values $x(1), x(2), \ldots, x(N)$.

In this situation there are at least three different types of estimates of x which may be formed. First is the "smoothing" estimate, which is the set of estimates of all the $x(i)$, $i = 1, \ldots, N$, based on the entire set of observations $y(i)$, $i = 1, \ldots, N$. Second is the "filtering" estimate, which is the estimate of any $x(k)$ based only upon the observations up to that time. Third is the "prediction" estimate, which is the estimate of any $x(k)$ based on a sequence of observations preceding t_k (i.e., $y(i)$, $i = 1, \ldots, k-n$), $0 < n < k$.

Smoothing.--The smoothing estimate may be constructed by collecting all the $x(i)$ into a vector X:

$$X = \begin{Bmatrix} x(1) \\ x(2) \\ \cdot \\ \cdot \\ \cdot \\ x(N) \end{Bmatrix}$$

and the observations into a vector Y:

$$Y = \begin{Bmatrix} y(1) \\ \cdot \\ \cdot \\ \cdot \\ y(N) \end{Bmatrix}$$

The solution then proceeds precisely by the method given in the previous section. One, of course, requires the distributions of X and of $(Y|X)$ and a loss function. The estimate is then

$$\hat{X}_0 = \arg\min_{\hat{X}} \int_{\Omega_x} l(X,\hat{X}) p(X|Y) dx \tag{55}$$

The simplicity of equation (55) is, of course, deceptive inasmuch as l and p are generally very complex due to the multidimensionality of the problem. There are nevertheless many cases where the problem can be rendered more tractable. One such is the case in which $\{x\}$ can be regarded as a Markov sequence--that is,

$$x(k+1) = f[x(k), u(k)] \tag{56}$$

where $x(k+1)$ depends only upon the previous $x(k)$ and a random "disturbance," $u(k)$, whose distribution is known. This is the situation, for instance, for a space vehicle. The vehicle's "state" is comprised of the components of its position and velocity,[5] which at any time may be given in terms of a previous state and any disturbances which may perturb the trajectory in the intervening period.

Under these conditions one can use the relatively simple distributions of $x(0)$ and $u(0), u(1), \ldots, u(N-1)$ to construct $p(X)$ by the law of derived distributions. This still may not be very easy to do, but at least it is a *modus operandi* which in many cases will be successful. (See ref. 22.)

It may be noted that it is frequently desired to form smoothing estimates in a multistage fashion. That is, at time t_k one may want an estimate of $[x(0), x(1), \ldots, x(k)]$ based on observations $[y(1), \ldots, y(k)]$, then at t_{k+n} a new estimate of $[x(0), \ldots, x(k+n)]$ based on $[y(1), \ldots, y(k+n)]$, etc. Basically, this requires a repetition of the computation procedure outlined above at each stage, although it will be recognized that the work required to construct $p(X)$, $p(Y|X)$, and $l(X,\hat{X})$ will involve only the addition of the new states and observations encountered since the previous estimate was computed.

Filtering.--Filtering is simpler than smoothing in that only one of the elements of $\{x(i)\}$ is to be estimated at a time rather than the entire set. One proceeds in the same manner, but needs less information--namely, $p[x(k) \mid y(1), \ldots, y(k)]$ and $l[x(k), \hat{x}(k)]$--whence the estimate is obtained by

$$\hat{x}_0(k) = \arg\min_{\hat{x}(k)} \int l\left[x(k), \hat{x}(k)\right] p\left[x(k) \mid y(1), \ldots, y(k)\right] dx(k) \tag{57}$$

[5] Any six elements which completely characterize the trajectory (such as orbital elements) may be used to represent the state.

Now, it will be recognized that the l and p used here are contained in $l(X,\hat{X})$ and $p(X|Y)$ which were used for the smoothing estimate; and, of course, if these are available, they can be used. The problem is simpler than smoothing, then, simply because the minimization in equation (57) needs to be done with respect to one (vector) variable, $\hat{x}(k)$, rather than the entire set $[\hat{x}(i)]$, $i = 1, \ldots, k$.

As in smoothing, filtering frequently involves a multistage or sequential operation. In fact, the usual connotation of the term "filtering" is that a continuing operation is involved, where one wishes to form successively the estimates of $x(i)$ based on observations $y(1), \ldots, y(i)$ for $i = 1, \ldots, k$. From equation (57), it is clear that for sequential estimation of this type one needs to determine successively the posterior (or conditional) distribution as represented by $p[x(i)|y(1), \ldots, y(i)]$ for $i = 1, 2, \ldots, k$. Presuming that the prior $p[x(0)]$ is given, that the law by which $x(0), \ldots, x(i-1)$ and $u(0), \ldots, u(i-1)$ map into $y(i)$ is known, and that the relations $p[y(i)|x(i)]$ are given, the procedure at each stage requires (1) the application of the law of derived distributions to obtain $p[x(i)|y(1), \ldots, y(i-1)]$, and (2) the application of Bayes' rule to obtain

$$p[x(i)|y(1), \ldots, y(i)] = \frac{p[x(i)|y(1), \ldots, y(i-1)]p[y(i)|x(i)]}{p[y(i)]} \tag{58}$$

This conditional distribution is the general estimate of $x(i)$; a specific estimate for a particular loss function may be obtained from it by application of equation (57).

Prediction.--The prediction estimate is so named because a future state is to be estimated on the basis of only those observations made prior to the time the state of interest will occur. In other words, the problem is one of extrapolating an estimate already at hand.

Presuming that one knows the distribution of the current state and the law by which this state will propagate into the future state space, the distribution of the future state can be computed by the theory of derived distributions. For instance, suppose the propagation law is

$$x(k+1) = f[x(k), u(k)] \tag{59}$$

If one has the inverse function $x(k) = f^*[x(k+1), u(k)]$ and the joint density function of $x(k)$ and $u(k)$, $p_1[x(k), u(k)]$, then the joint density function of $x(k+1)$ and $u(k)$ is given by

$$p_2[x(k+1), u(k)] = p_1\{f^*[x(k+1), u(k)], u(k)\} \det\left[\frac{\partial f}{\partial x(k)}\right] \tag{60}$$

Then $p[x(k+1)]$ is the marginal density of equation (60), or

$$p[x(k+1)] = \int p_2[x(k+1), u(k)]du(k) \tag{61}$$

with $p[x(k+1)]$ and a suitable loss function, the prediction estimate is obtained by application of equation (57).

The same remarks made previously for the smoothing and filtering estimates with respect to sequential estimation apply equally to the prediction estimate. That is, frequently one is interested in repeated redetermination of a predicted state as more observations are obtained. An example is the case of space vehicle guidance, where one

needs to predict end-point conditions (in order to make course corrections) repeatedly as new information regarding the trajectory is acquired.

The Kalman filter.--Under certain conditions the sequential estimation procedure outlined above may be reduced to explicit formulas. To show this it is instructive to apply the conditions one at a time so that intermediate results may be obtained.

First, if $\{y(k)\}$ is a Markov sequence, that is,

$$\left.\begin{aligned} y(k) &= h[x(k), e(k)] \\ x(k) &= f[x(k-1), u(k-1)] \end{aligned}\right\} \tag{62}$$

then the updating of the conditional distribution $p[x(k)|y(1), \ldots, y(k)]$ is somewhat simplified since at each stage only the distributions $p[e(k)]$, $p[u(k)]$, and $p[x(k-1)|y(1), \ldots, y(k-1)]$ are required. The latter is, of course, the conditional distribution obtained at the previous stage and can be presumed given. The starting point in the sequence is simply the prior distribution $p[x(0)]$.

Second, if the loss function is of a special class, then the estimate is obtained directly by finding the mean of the conditional distribution. This simplifies the problem by eliminating the minimization required in equation (57).

Third, if the conditional distribution is symmetric and unimodal, the conditional mean, mode, and median are synonymous; and one of the latter may be easier to compute than the mean, thereby simplifying the problem solution.

Finally, if the process $\{x(i)\}$ is generated by or can be represented as the output of a linear system with normally distributed (Gaussian) inputs, the filtering problem can be reduced to explicit formulas. The linear system assumption means that x and y obey the equations

$$\left.\begin{aligned} x(k) &= Fx(k-1) + Gu(k-1) \\ y(k) &= Ax(k) + e(k) \end{aligned}\right\} \tag{63}$$

The Gaussian assumption means that $x(0)$, $u(i)$, and $e(i)$, $i = 1, 2, \ldots, k$, are all normally distributed independent random variables. (Zero mean is also assumed for this example, but this does not limit the generality of the results since a transformation of all variables to their mean values is always possible.)

Now, with the Guassian assumption it is apparent that all $x(i)$, $i = 0, \ldots, k$, will be Gaussian, as will the observations and all conditional distributions. In this case all distributions involved in the sequential estimation problem can be fully characterized by their means and covariance matrices. The computation of the conditional distribution can thereby be reduced to the computation of its mean (which is the same as its mode and median) and its covariance matrix.

Consider the situation which obtains at the time of any one of the sequence of observations, which we shall call $y = Ax + e$. Define $p(x)$ as the density function which summarizes all the knowledge about x except for that contributed by y, and define its

mean (or mode or median) as μ and its covariance matrix as P. Now we have for the various density functions:

$$\left.\begin{aligned} p(x) &= D_1 \exp\left[-\tfrac{1}{2}(x-\mu)^T P^{-1}(x-\mu)\right] \\ p(y|x) &= D_2 \exp\left[-\tfrac{1}{2}(y-Ax)^T Q^{-1}(y-Ax)\right] \\ p(y) &= D_3 \exp\left[-\tfrac{1}{2}(y-A\mu)^T (APA^T+Q)^{-1}(y-A\mu)\right] \end{aligned}\right\} \quad (64)$$

Combining these by means of Bayes' rule and manipulating the result to obtain the simplest form, we obtain

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

$$= \frac{D_1 D_2}{D_3} \exp\left\{-\frac{1}{2}\left[x-\mu-\Lambda A^T Q^{-1}(y-A\mu)\right]^T \Lambda^{-1}\left[x-\mu-\Lambda A^T Q^{-1}(y-A\mu)\right]\right\} \quad (65)$$

where $\Lambda = (P^{-1} + A^T Q^{-1} A)^{-1}$. From this expression it is seen that the mean (or mode or median), hence the optimal estimate of x, is given by

$$\hat{x} = \mu + (P^{-1} + A^T Q^{-1} A)^{-1} A^T Q^{-1}(y - A\mu) \quad (66)$$

and the covariance matrix of the distribution is

$$\Lambda = (P^{-1} + A^T Q^{-1} A)^{-1} \quad (67)$$

Now we note the following. Since we are assuming a sequential estimation, it is clear that the mean μ of the prior distribution is the optimal estimate of x based on all prior observations. We may denote this fact by writing $\mu = \hat{x}(k|k-1)$. Similarly, P is the prior covariance, which we may write as $P(k|k-1)$, and Λ is the posterior covariance, $\Lambda = P(k|k)$. Furthermore, it is convenient to use a well-known matrix inversion formula (ref. 23) to write

$$\Lambda = (P^{-1} + A^T Q^{-1} A)^{-1} = P - PA^T(APA^T + Q)^{-1} AP \quad (68)$$

Using also the matrix identity (ref. 1)

$$(P^{-1} + A^T Q^{-1} A)^{-1} A^T Q^{-1} = PA^T(APA^T + Q)^{-1} \quad (69)$$

and the altered notation, we may write equations (66) and (67) in the forms

$$\hat{x}(k|k) = \hat{x}(k|k-1) + K\left[y(k) - Ax(k|k-1)\right] \quad (70)$$

$$P(k|k) = [I - KA]P(k|k-1) \quad (71)$$

where

$$K = P(k|k-1)A^T[AP(k|k-1)A^T + Q]^{-1}$$

To complete the sequential estimation problem, we need further formulas for updating from one observation to the next. Since it is assumed that $x(k + 1) = Fx(k) + Gu(k)$, where $u(k)$ has zero mean and is assumed normally distributed and independent of $x(k)$, then the distribution of x in propagating from k to $k + 1$ remains Gaussian (by the law of derived distributions), and its mean and covariance are given by

$$\hat{x}(k + 1|k) = F\hat{x}(k|k) \tag{72}$$

$$P(k + 1|k) = FP(k|k)F^T + GRG^T \tag{73}$$

where R is the covariance of u, $R = E[uu^T]$.

Equations (70) - (71) together with (72) - (73) are the complete recursive relations required to compute, sequentially, the optimal estimate of x at each stage. It is seen that these are precisely the equations of the Kalman filter (discrete version), which were originally developed (ref. 24) from a somewhat different point of view.

REMARKS ON THE UTILITY OF STATISTICAL INFERENCE METHODS

The principal conclusion of the foregoing development is that the differences between the various estimation methods can be characterized fundamentally by (1) the degree of generality in the optimality criterion and (2) the completeness of the probabilistic statements made about e and x. Bayesian estimation allows the use of any type of criterion which can be expressed as a function of the state, while maximum likelihood utilizes only the intuitive notion that the optimal estimate should be near the mode of the likelihood function, minimum variance is restricted to criteria of the mean-square error type, and least squares uses a pragmatic nonprobabilistic criterion. In regard to probabilistic statements, Bayesian estimation utilizes complete descriptions of the prior distributions of x and e, maximum likelihood omits the distribution of x, minimum variance is further restricted to consideration only of the second order statistics of e, and least squares makes no use of probabilities at all.

The Bayesian decision procedure is the most universally applicable of the mathematical methods of statistical inference described herein. There are, however, practical considerations to be taken into account in utilizing the general theory. There is first the subjective matter of specifying in mathematical terms the probabilities and optimality criteria, and then there is the matter of computational complexity where problems do not have simple mathematical models.

The first of these difficulties can be partly overcome by diligent efforts toward objectivity, and the second can be alleviated by the development of efficient computational algorithms and methods of approximation. (Typical recent contributions along these lines are given in refs. 25 and 26.) Of course, all such efforts are costly and can be justified only by a substantial pay-off from the increased accuracy which such efforts engender in the decision-making process. In a wide variety of important problems, this pay-off is not sufficient, and the less complete methods of statistical inference (such as least squares) which can be implemented with less effort are quite appropriate approximate methods for their solution.

Ames Research Center
National Aeronautics and Space Administration
Moffett Field, Calif., Aug. 19, 1966

REFERENCES

1. Smith, Gerald L.; Schmidt, Stanley F.; and McGee, Leonard A.: Application of Statistical Filter Theory to the Optimal Estimation of Position and Velocity on Board a Circumlunar Vehicle. NASA TR R-135, 1962.

2. Battin, Richard H.: A Statistical Optimizing Navigation Procedure for Space Flight. MIT Instrumentation Lab. Rep. R-341, Sept. 1961 (Rev. May 1962).

3. Freund, John E.: Modern Elementary Statistics. Prentice-Hall, N.Y., 1952.

4. Jeffreys, Sir Harold: Scientific Inference. Second ed., Cambridge Univ. Press, 1957.

5. Raiffa, Howard; and Schlaifer, Robert: Applied Statistical Decision Theory. Harvard University, 1961.

6. Kalman, R. E.: New Methods and Results in Linear Prediction and Filtering Theory. RIAS Tech. Rep. 61-1, 1961.

7. Breakwell, J. V.; and Tung, F.: Minimum Effort Control of Several Terminal Components. SIAM Ser. A, Control 1964, vol. 2, no. 3, pp. 295-316.

8. Novosel'tsev, V. N.: Time-Optimal Control Systems in the Presence of Random Noise. Automation and Remote Control, vol. 23, no. 12, Dec. 1962, pp. 1519-1529.

9. Sherman, Seymour: Non-Mean-Square Error Criteria. IRE Trans. on Information Theory, vol. IT-4, no. 3, Sept. 1958, pp. 125-126.

10. Decell, Henry P., Jr.: An Application of Generalized Matrix Inversion to Sequential Least Squares Parameter Estimation. NASA TN D-2830, 1965.

11. Anderson, T. W.: An Introduction to Multivariate Statistical Analysis. John Wiley and Sons, Inc., 1958.

12. Todd, John, ed.: Survey of Numerical Analysis. McGraw-Hill Book Co., Inc., 1962.

13. Scheffé, Henry: The Analysis of Variance. John Wiley and Sons, Inc., 1959.

14. Magness, T. A.; and McGuire, J. B.: Comparison of Least Squares and Minimum Variance Estimates of Regression Parameters. Ann. Math. Stat. vol. 33, no. 2, June 1962, pp. 462-470.

15. Fisher, Sir R. A.: Contributions to Mathematical Statistics. John Wiley and Sons, 1950.

16. Mood, A. M.: Introduction to the Theory of Statistics. McGraw Hill Book Co., 1950.

17. Shapiro, I. I.: The Prediction of Ballistic Missile Trajectories From Radar Observations. McGraw-Hill Book Co., 1958.

18. Cramér, Harald: Mathematical Methods of Statistics. Princeton Univ. Press, 1946.

19. Davenport, W. B., Jr.; and Root, W. L.: An Introduction to the Theory of Random Signals and Noise. McGraw-Hill Book Co., 1958.

20. Joseph, Peter D.; and Tou, Julius T.: On Linear Control Theory. Trans. AIEE, pt. II (Applications and Industry), vol. 80, no. 56, Sept. 1961, pp. 193-196.

21. Gunckel, T. L., II; and Franklin, Gene F.: A General Solution for Linear, Sampled-Data Control. Trans. ASME, J. Basic Engrg., vol. 85-D, no. 2, June 1963, pp. 197-203.

22. Lee, Robert C. K.: Optimal Estimation, Identification, and Control. (Research Monograph 28), MIT Press, Cambridge, Mass., 1964.

23. Householder, Alston S.: Principles of Numerical Analysis. McGraw-Hill Book Co., 1953.

24. Kalman, R. E.: A New Approach to Linear Filtering and Prediction Problems. RIAS, Inc., Monograph 60-11, 1960.

25. Cox, Henry: Recursive Linear Filtering. Proc. of the National Electronics Conf., vol. XXI, Oct. 25-27, 1965, pp. 770-775.

26. Friedland, Bernard; Bernstein, Irwin; and Ellis, Jordan: Extensions to the Kalman Filter Theory. Final Rep. NASA Contract NAS 9-3505, General Precision, Inc., Nov. 1, 1965, NASA CR-65344.