√A

X-547-65-367

NASA TM X-55371

LINEAR ESTIMATION AND RELATED TOPICS

BY
RAYMOND V. BORCHERS

SEPTEMBER 1965

NASA

GODDARD SPACE FLIGHT CENTER
GREENBELT, MARYLAND

LINEAR ESTIMATION AND RELATED TOPICS

by

Raymond V. Borchers

Goddard Space Flight Center
Greenbelt, Maryland

CONTENTS

LIST OF SYMBOLS

$\Omega_{t_0}^{t}$	matrix relating residuals or small departures from some known orbit at times t_0 and t. A common name for this matrix is matrizant or state transition matrix.
x	true value of $p \times 1$ state vector
$\hat{x}$	estimate of x
y	$n \times 1$ vector of observations or measurements
$\hat{y}$	computed value of y vector
t	time
t_0	initial time
δy	$n \times 1$ vector of deviations in the observations ; $\delta y = y - \hat{y}$
δx	$p \times 1$ vector of deviations in the state variables from reference; $\delta x = x - \hat{x}$
e	$n \times 1$ noise vector associated with the observations
Q	covariance matrix of the observational error e
I	unit matrix
$H(t)$	matrix of partial derivatives of the observables with respect to the state variables relating δy to δx.
K	weighting matrix in optimal Kalman filter
m	message, $H \delta x$
P	covariance matrix of δx
R	covariance matrix of u
F	matrix of coefficients for linear differential perturbation equations
$E[\ ]$	expected value of $[\ ]$

List of Symbols (Continued)

$( )^T$	transpose of matrix ( )
$( )^{-1}$	inverse of matrix ( )
σ	standard deviation
s	quadratic form
tr( )	trace of a matrix
$(\hat{\ })$	estimate of ( )
θ	vector of parameters
$u_n(t)$	white noise
Z	augmented state vector

LINEAR ESTIMATION AND RELATED TOPICS

INTRODUCTION

The purpose of this paper is to present a collection of theorems, definitions, and other facts about the multivariate Gaussian distributions, to give desirable properties of linear estimators, and to examine some statistical techniques of linear estimation which are now used in the differential correction of orbits of near earth satellites. The basic concepts of classical orbit correction will be presented in the first part of the paper to show where the need arises for estimating such quantities as orbital parameters, geodetic parameters, etc. The motivation for writing this paper has arisen from the necessity of determining satellite orbits from tracking station data which is known to be contaminated with noise.

The different estimation techniques to be discussed are (1) unweighted least squares, (2) conventional weighted least squares, (3) maximum-likelihood, (4) Bayes, and (5) Kalman-Schmidt filter. All of these techniques except Kalman-Schmidt utilize all the observations simultaneously in arriving at an estimate for the orbital parameters. The Kalman-Schmidt filter permits a complete optimum estimate of the orbital parameters from each single observation and hence operates sequentially on the observations one at a time in the order of their occurrence. It has been shown reference [15] that the Kalman-Schmidt technique is equivalent to a linear least squares method and that the covariance matrices of the orbit parameters are the same. A comparative study of the different techniques based on similar notations will aid the astrodynamicist in selecting among them.

Conventional Differential Orbit Correction

Every observed or measured quantity contains errors of unknown magnitude due to a variety of causes, and hence a measurement is never exact. The resultant error in a given quantity is the difference between the measurement and its true value. For a single quantity which has been determined by observation, neither the resultant error nor any of its individual parts can ever be determined exactly, but can be fixed within certain probable limits.

One may describe a differential correction procedure as a systematic method for using the residuals to improve the values of a set of orbital parameters until the "best" set has been obtained. The differences between the actual measurements or observations and the computed values of the quantities observed as a function of the state vector $\hat{x}(t_0)$ are called the residuals. The basic concepts of a differential correction procedure can be illustrated as follows:

Let y_i be an observation or measurement of some quantity such as right ascension or declination at time $t = t_i$ and let e_i be the associated observational error. Then there exists a function $F_i\ (x\ (t_0);\ t_i)$ of the unknown state vector $x\ (t_0)$ such that

$$y_i = F_i(x(t_0)\ ;\ t_i) + e_i.$$

Let $\hat{y}_i\ (\hat{x}\ (t_0);\ t_i)$ be the computed value of the quantity observed at time $t = t_i$ based upon an estimated state vector $\hat{x}\ (t_0)$ at time $t = t_0$. If we assume that $\hat{x}\ (t_0)$ is sufficiently close to $x\ (t_0)$ so that squares and higher powers of $\delta x\ (t_0) = x(t_0) - \hat{x}\ (t_0)$ can be neglected, then $F_i\ (x(t_0);\ t_i)$ may be expanded in a Taylor series to a first order approximation about the estimated state vector $\hat{x}\ (t_0)$, i.e.

$$y_i \simeq F_i\ (\hat{x}\ (t_0)\ ; t_i) + \sum_j \left(\frac{\partial F_i}{\partial \hat{x}_j}\right)_{t=t_i} (x_j - \hat{x}_j) + e_i\ , \quad (i = 1,\ 2,\ \ldots,\ n). \qquad (1)$$

If we assume that the functions $F_i\ (\hat{x}(t_0);\ t_i)$ can be replaced by the known functions $\hat{y}_i\ (\hat{x}\ (t_0);\ t_i)$ equations (1) become

$$y_i \simeq \hat{y}_i\ (\hat{x}\ (t_0);\ t_i) + \sum_j \left(\frac{\partial \hat{y}_i}{\partial \hat{x}_j\ (t_0)}\right)_{t=t_i} (x_j - \hat{x}_j) + e_i,\ (i = 1,\ 2,\ \ldots,\ n). \qquad (2)$$

In matrix form the n linear equations (2) become

$$\delta y(t) = H(t)\ \delta x(t_0) + e \tag{3}$$

where

$$\delta y(t) = \begin{bmatrix} \delta y_1(t_1) \\ \delta y_2(t_2) \\ \cdots \\ \delta y_{n(tn)} \end{bmatrix} = \begin{bmatrix} y_1(t_1) - \hat{y}_1(\hat{x}(t_0)\ ;\ t_1) \\ y_2(t_2) - \hat{y}_2(\hat{x}(t_0)\ ;\ t_2) \\ \cdots\cdots\cdots\cdots \\ y_n(t_n) - \hat{y}_n(\hat{x}(t_0)\ ;\ t_n) \end{bmatrix}_{n\times 1}, \quad e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}_{n\times 1}, \tag{4}$$

$$H(t) = \begin{bmatrix} \left(\frac{\partial \hat{y}_1}{\partial \hat{x}_1}\right)_{t_1} & \left(\frac{\partial \hat{y}_1}{\partial \hat{x}_2}\right)_{t_1} & \cdots \left(\frac{\partial \hat{y}_1}{\partial \hat{x}_p}\right)_{t_1} \\ \left(\frac{\partial \hat{y}_2}{\partial \hat{x}_1}\right)_{t_2} & \left(\frac{\partial \hat{y}_2}{\partial \hat{x}_2}\right)_{t_2} & \cdots \left(\frac{\partial \hat{y}_2}{\partial \hat{x}_p}\right)_{t_2} \\ \cdots & \cdots & \cdots \\ \left(\frac{\partial \hat{y}_n}{\partial \hat{x}_1}\right)_{t_n} & \left(\frac{\partial \hat{y}_n}{\partial \hat{x}_2}\right)_{t_n} & \cdots \left(\frac{\partial \hat{y}_n}{\partial \hat{x}_p}\right)_{t_n} \end{bmatrix}_{n\times p}, \quad \delta x(t_0) = x(t_0) - \hat{x}(t_0) = \begin{bmatrix} x_1 - \hat{x}_1 \\ x_2 - \hat{x}_2 \\ \cdots \\ x_p - \hat{x}_p \end{bmatrix}_{p\times 1}.$$

The unknown state vector x (t_0) is the variable which is to be found by solving equation (3). Assume for the moment that δy (t) is not affected by observational errors but only by errors in the initial state vector $\hat{x}$ (t_0). It is known from the theory of linear equations for the case where n = p, that a necessary and sufficient condition that equation (3) has a unique solution is that the determinant of the matrix H (t) does not vanish. If the observations were not affected by errors an exact solution to equation (3) would be obtained. The presence of accidental errors in δy (t), however, prevents the true value of x (t_0) from being determined and hence only approximate values can be found. By increasing the number of observations the effect of the observational errors can be diminished.

If $n > p$, i.e., the number of observations is greater than the number of unknowns, then the resulting linear system of equations is overdetermined and a least square criterion is applied to fit the trajectory to the observations. In practice a weighting matrix or diagonal matrix W is applied to equation (3). If the observation errors are included in the matrix of residuals equation (3) becomes

$$\begin{aligned} WH\,\delta x &= W\,\delta y \\ (WH)^T(WH)\,\delta x &= (WH)^T W\,\delta y \qquad (5) \\ \delta x &= ((WH)^T(WH))^{-1}(WH)^T W\,\delta y \\ \delta x &= (H^T(W^TW)H)^{-1}H^T(W^TW)\,\delta y \\ \delta x &= (H^TW^2H)^{-1}H^TW^2\,\delta y. \qquad (6) \end{aligned}$$

The solution of the normal equation (5) provides $\delta x(t_0)$ given by equation (6), and the original estimated state vector $\hat{x}(t_0)$ is improved to give a new estimate of x

$$\hat{x}_{new}(t_0) = \hat{x}_{old}(t_0) + \delta x(t_0).$$

This solution is a least squares estimate of the correct solution based on n observations. In practice an iterative procedure is employed on the new state $x_{new}(t_0)$ to obtain a further improved state vector using as a test for convergence some statistical measure of the magnitude of the residuals. For example, at the Smithsonian Institution Astrophysical Observatory, the iteration is assumed to converge if the change of the standard deviation σ on two successive iterations differs by less than 1 part in 100, or

$$\frac{\sigma_i - \sigma_{i-1}}{\sigma_i} < 0.01 .$$

Multivariate Gaussian Distributions

The scientific investigator often knows or is willing to assume that the population from which he takes observations is of a certain functional form.

Usually one knows very little about the distribution of errors in the data and the assumption is generally made that the errors are normal.

The smoothing or estimation problem can be described as follows: One has a set of data which is known to be the sum of desired information and some random errors. The problem is to extract the desired information. A concrete form of this is given by equation (3) where the state vector $x(t)$ is to be determined. In the statistical literature this is referred to as "regression analysis" or "linear analysis."

The question of choosing a criterion for smoothing and estimation has a long history. In 1799 La Place encountered the question and proposed the use of the L_∞ norm. Since one could not perform any computation with this norm, Legendre proposed the L_2 or least squares norm in 1805. Since that time there has been a considerable amount of controversy about this question. The principle of least squares states that the best estimate $\hat{x}$ of x is that number which minimizes the sum of the squares of the deviations of the measurements from their estimate. The principle of least squares is normally defended on the basis of the assumption that the errors are normally distributed. It is undoubtedly true that the L_2 norm is efficient in such a situation, probably the most efficient possible.

Since the normal distribution is used very frequently, it is felt that some background information in the way of theorms and definitions may be of value in discussing the different estimation techniques.

Let $Y = \{y_1, y_2, \ldots, y_p\}$ be a p-dimensional random vector with mean vector $\mu = EY = \{\mu_1, \mu_2, \ldots, \mu_p\}$.

Definition 1. The random vector Y is normally distributed in p dimensions if the joint density of $y_1, y_2, \cdots, y_p$ is

$$f(Y) = f(y_1, y_2, \cdots, y_p) = \frac{|R|^{1/2}}{(2\pi)^{p/2}} e^{-1/2\,(Y-\mu)^T R\,(Y-\mu)} \qquad (1)$$

$$(-\infty < y_i < \infty,\ i = 1, 2, \cdots, p)$$

where R is a positive definite symmetric matrix whose elements r_{ij} are constants, μ is a $p \times 1$ vector whose elements μ_i are constants. We observe that for the special case $p = 1$ that $R = r_{11}$ which must be positive. If we set $r_{11} = 1/\sigma^2$, it is clear that the frequency or density function $f(Y)$ is the normal distribution in a single variable. The quantity $S = (Y - \mu)^T R (Y - \mu)$ is called the quadratic form of the p-variate normal. It is a quadratic form in the elements $y_i - \mu_i$. It can be written as

$$S = \sum_{j=1}^{p} \sum_{i=1}^{p} (y_i - \mu_i)(y_j - \mu_j) r_{ij}. \qquad (2)$$

In order for f(Y) to qualify as a density function it is essential that f(Y) $\geq$ 0. This is clear from the fact that the determinant of a positive definite matrix is positive. It is also necessary that the integral of f(Y) be equal to 1. A consequence of this fact is

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-1/2 (Y-\mu)^T R(Y-\mu)} dy_1 \, dy_2 \cdots dy_p = \frac{(2\pi)^{p/2}}{|R|^{1/2}} .$$

Definition 2. The expected value of a matrix or vector A which is written E(A) is defined as the expected value of each element of A. For example, if

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

$$E(A) = \begin{pmatrix} E(a_{11}) & E(a_{12}) \\ E(a_{21}) & E(a_{22}) \end{pmatrix}.$$

Definition 3. Let the p × 1 random vector Y be distributed as the p-variate normal; then $E(Y) = \mu$.

Definition 4. The variance of the random variable $y_i = E[y_i - E(y_i)]^2 = E[y_i - \mu_i]^2$, and the covariance of the two random variables y_i and y_j is $E(y_i - \mu_i)(y_j - \mu_j)$, $i \neq j$. In a p-variate normal there will be p variances one for each random variable y_i, and $p(p - 1)/2$ covariances. A p × p matrix has the covariance of y_i and y_j as its ij th element if $i \neq j$, and the variance of y_i as its i th diagonal element. Hence

$$V = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{pp} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix}$$

where $\sigma_{ij} = \sigma_{ji} = E(y_i - \mu_i)(y_j - \mu_j)$ or in matrix notation $V = E(Y - \mu)(Y - \mu)^T$. V is called the covariance matrix of the vector Y.

Theorem 5. In the p-variate normal the matrix R is the inverse of the covariance matrix V, i.e., $V^{-1} = R$ or also $R^{-1} = V$. As a consequence of this the p-variate normal can be written as

$$\frac{1}{|V|^{1/2}(2\pi)^{p/2}} e^{-1/2(Y-\mu)^T V^{-1}(Y-\mu)} .$$

If the covariance matrix V is a diagonal matrix then the two random variables y_i and y_j are independent. Clearly the non diagonal elements are 0 if and only if the correlation between the two random variables y_i and y_j is 0, since the correlation is defined as

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\ \sigma_{jj}}}$$

Theorem 6. Let the vector Y have a p-variate normal density with mean μ and covariance matrix V, and let $\alpha_1, \alpha_2, \ldots, \alpha_p$ be a set of constants. Then $Z = \sum_{i=1} \alpha_i\ y_i$ is distributed as the univariate normal with mean $\Sigma\, \alpha_i\ \mu_i$ and variance

$$\sum \alpha_i^2\, \sigma_i^2 + \sum_{i} \sum_{\substack{j \\ i \neq j}} \alpha_i\, \alpha_j\, \sigma_{ij} .$$

Example. Let the 3×1 vector Y have a p-variate normal density, where

$$\mu = \begin{pmatrix} 3 \\ -1 \\ 0 \end{pmatrix}, \quad R = \begin{pmatrix} 2 & 0 & 3 \\ 0 & 1 & 0 \\ 3 & 0 & 5 \end{pmatrix}$$

The covariance matrix is

$$R^{-1} = V = \begin{pmatrix} 5 & 0 & -3 \\ 0 & 1 & 0 \\ -3 & 0 & 2 \end{pmatrix} .$$

Here $E(y_1) = \mu_1 = 3$, $E(y_2) = \mu_2 = -1$, $E(y_3) = \mu_3 = 0$. $\sigma_{11} = 5$, $\sigma_{12} = \sigma_{21} = 0$, $\sigma_{13} = \sigma_{31} = -3$, $\sigma_{22} = 1$, $\sigma_{23} = \sigma_{32} = 0$, $\sigma_{33} = 2$. We say that the mean of y_2 is -1 and the variance of y_2 is 1. If $Z = y_1 - 3y_2 + 2y_3$ with $\alpha_1 = 1$, $\alpha_2 = -3$, $\alpha_3 = 2$, then

$$E(Z) = 1, E(y_1) - 3E(y_2) + 2E(y_3) = 1, 3 - (3)(-1) + 2(0) = 6$$

$$\text{var}(Z) = \sum_{i=1}^{3} \alpha_i^2 \sigma_i^2 + \sum_{i=1}^{3} \sum_{i=1}^{3} \alpha_i \alpha_j \sigma_{ij}$$

$$= \alpha_1^2 \sigma_1^2 + \alpha_2^2 \sigma_2^2 + \alpha_3^2 \sigma_3^2 + 2 [\alpha_1 \alpha_2 \sigma_{12} + \alpha_1 \alpha_3 \sigma_{13} + \alpha_2 \alpha_3 \sigma_{23}]$$

$$= (1)^2 (5) + (-3)^2 (1) + (2)^2 (2) + 2 [(1)(-3)(0) + (1)(2)(-3) + (-3)(2)(0)]$$

$$\text{var}(Z) = 5 + 9 + 8 + 2 [0 - 6 + 0] = 22 - 12 = 10$$

<u>Theorem 7</u>. If a vector Y is distributed with mean 0 and covariance matrix $\sigma^2 I$, the expected value of the quadratic form $Y^T AY$ is equal to σ^2 tr (A).

DESIRABLE PROPERTIES OF LINEAR ESTIMATORS

It is perhaps appropriate to review some important desirable properties of estimators: (1) sufficiency, (2) unbiasedness, (3) consistency, (4) efficiency, (5) minimum variance, (6) completeness, and (7) invariance under transformation.

<u>Sufficiency</u>. Let $f(y_1, y_2, \ldots, y_n; \theta_1, \theta_2, \ldots, \theta_K)$ be a joint frequency function involving K parameters. The statistics $\hat{\theta}_1 = h_1(y_1, y_2, \ldots, y_n)$, $\hat{\theta}_2 = h_2(y_1, y_2, \ldots, y_n), \ldots, \hat{\theta}_m = h_m(y_1, y_2, \ldots, y_n)$ are a set of sufficient

statistics if $g(y_1, y_2, \ldots, y_n \mid \hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m)$ is the conditional density function of the observations given the statistics. The original observations $y_1, y_2, \ldots, y_n$ always form a set of sufficient statistics but generally what is wanted is a minimal sufficient set.

Unbiasedness. An estimate $\hat{\theta}$ is said to be an unbiased estimator of θ if $E(\hat{\theta}) = \theta$.

Consistency. An estimator $\hat{\theta}_n$ (A sequence of estimators $\{\hat{\theta}_n\}$) is said to be consistent for θ if the limit of the probability that $\hat{\theta}_n - \theta = 0$ is 1 as $n \to \infty$. $\hat{\theta}_n$ is the estimate of θ based on a sample of size n.

Efficiency. An estimator $\hat{\theta}_n$ (A sequence of estimators $\{\hat{\theta}_n\}$) is said to be efficient when the following two conditions are satisfied: (1) $\sqrt{n}\,(\hat{\theta}_n - \theta)$ is asymptotically normally distributed with mean 0 and variance σ^2 where n is the sample size, (2) the variance σ^2 is less than the variance of any other estimator θ_n^* that satisfies condition 1.

Minimum variance. An estimator $\hat{\theta}$ is said to be a minimum variance estimator of θ if

$$E[\hat{\theta} - E(\hat{\theta})]^2 \leq E[\theta^* - E(\theta^*)]^2$$

where θ^* is any other estimator for θ. If an estimator is efficient, then it is consistent and unbiased in the limit but need not be unbiased for finite sample sizes. An unbiased estimator is not necessarily consistent.

Completeness. An estimator $\hat{\theta}$ is complete if there exists no unbiased estimator of zero in the frequency function $f(\hat{\theta}; \theta)$ (except zero itself).

Invariance. An estimator $\hat{\theta}$ of θ is said to be an invariant estimator for a certain class of transformations g if the estimator is $g(\hat{\theta})$ when the transformation changes the parameter to $g(\theta)$.

The mean-squared error can be written as

$$E[(\hat{\theta} - \theta)^2] = E[\{\hat{\theta} - E(\hat{\theta})\} - \{\theta - E(\hat{\theta})\}]^2 = \mathrm{var}(\hat{\theta}) + [\theta - E(\hat{\theta})]^2.$$

The term $\theta - E(\hat{\theta})$ is called the bias of the estimator and can be either positive, negative, or zero. If an estimator can be found with bias close to zero and such that var $(\hat{\theta})$ is small, the mean squared error will be small. Thus it seems that it may be desirable to have an estimator whose bias is 0. If we consider only estimators which are unbiased, then $E(\hat{\theta}) = \theta$, and the mean-squared error of an unbiased estimator is equal to the variance of the estimator. Thus the mean squared error becomes $E[(\hat{\theta} - \theta)^2] = \mathrm{var}(\hat{\theta})$ if $\hat{\theta}$ is unbiased.

Definition 8 – Minimum-variance unbiased estimator. Let $y_1, y_2, \cdots, y_n$ be a random sample from $F(y; \theta)$. Let $\hat{\theta} = d(y_1, y_2, \cdots, y_n)$ be an estimator of θ such that (a) $E(\hat{\theta}) = \theta$; that is, $\hat{\theta}$ is unbiased (b) var $(\hat{\theta})$ is less than the variance of any other unbiased estimator. Then $\hat{\theta}$ is the minimum-variance unbiased estimator of θ.

Definition 9 – Maximum Likelihood Estimators. The likelihood function of n random variables $y_1, y_2, \cdots, y_n$ is the joint density of the n random variables $L(y_1, y_2, \cdots, y_n; \theta_1, \theta_2, \cdots \theta_K)$ which is considered to be a function of the k parameters. In particular, if $y_1, y_2, \cdots, y_n$ is a random sample from the density $f(y_1, y_2, \cdots, y_n; \theta_1, \theta_2, \cdots, \theta_K)$ then the likelihood function is

$$L(\theta_1, \theta_2, \cdots, \theta_K) = \prod_{i=1}^{n} f(y_i; \theta_1, \theta_2, \cdots, \theta_K).$$

Definition 10 – Maximum Likelihood Estimator. If the likelihood function contains k parameters, i.e.,

$$L(\theta_1, \theta_2, \cdots, \theta_K) = \prod_{i=1}^{n} f(y_i; \theta_1, \theta_2, \cdots, \theta_K)$$

then the maximum likelihood estimators of the parameters $\theta_1, \theta_2, \cdots, \theta_K$ are the random variables $\hat{\theta}_1 = d_1(y_1, y_2, \cdots, y_n)$; $\hat{\theta}_2 = d_2(y_1, y_2, \cdots, y_n)$; $\cdots$, $\hat{\theta}_K = d_K(y_1, y_2, \cdots, y_n)$, where $\hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_K$ are the values in the parameter space which maximize $L(\theta_1, \theta_2, \cdots, \theta_K)$. If certain regularity conditions are satisfied, the point where the likelihood is a maximum is a solution of the equations

$$\frac{\partial L}{\partial \theta_1}(\theta_1, \theta_2, \cdots, \theta_K) = 0$$

$$\frac{\partial L}{\partial \theta_2}(\theta_1, \theta_2, \cdots, \theta_K) = 0$$

$$\cdots\cdots\cdots\cdots$$

$$\frac{\partial L}{\partial \theta_K}(\theta_1, \theta_2, \cdots, \theta_K) = 0.$$

Theorem 11. The maximum likelihood estimators $\hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_K$ for the parameters of a density $f(y_1, y_2, \cdots, y_n; \theta_1, \theta_2, \cdots, \theta_K)$ from samples of size n are, for large samples, approximately distributed by the multivariate normal distribution with means $\theta_1, \theta_2, \cdots, \theta_K$ and with matrix nR in the quadratic form, where

$$r_{ij} = -E\left[\frac{\partial^2}{\partial \theta_i \theta_j} \log f(y_1, y_2, \cdots y_n; \theta_1, \theta_2, \cdots, \theta_K)\right].$$

The variances and covariances of the estimators are $(1/n)V$, where $V = R^{-1}$.

Definition 12. The model $Y = \Phi\theta + e$ where Y is a random observed vector, e is a random vector, Φ is an $n \times p$ matrix of known fixed quantities, θ is a $p \times 1$ vector of unknown parameters is called a general linear model of full rank, if the rank of Φ is equal to p where $p \leq n$.

Theorem 13. If the general linear hypothesis model of full rank $Y = \Phi\theta + e$ is such that the conditions on the random vector are

$$E(e) = 0$$

$$E(e\,e^T) \,\square\, \sigma^2 I$$

the best (minimum variance) linear unbiased estimate of the vector θ is given by least squares; i.e. $\hat{\theta} = (\Phi^T\Phi)^{-1}\ \Phi^T Y$ is the best linear unbiased estimate of θ. This estimate is the same as the maximum likelihood estimate under normal theory. This theorem is called the Gauss - Markhoff theorem.

Under certain quite general conditions the maximum likelihood method gives rise to estimators that are consistent, efficient, and sufficient.

If the method of maximum likelihood is to be used, the form of the frequency function must be known. The maximum likelihood estimators also possess a property which is called invariance, i.e., if $\hat{\theta} = d(y_1, y_2, \cdots, y_n)$ is the maximum likelihood estimator of θ in the density $f(y; \theta)$ and $u(\theta)$ is a function having a single valued inverse, then the maximum likelihood estimator of $u(\theta)$ is $u(\hat{\theta})$.

The maximum likelihood estimator is not always unbiased, but frequently it can be changed so that it becomes unbiased. For example

$$\hat{\sigma}^2 = \frac{1}{n} \sum (x_i - \bar{x})^2$$

is a biased estimator of σ^2 but $n/(n-1)\hat{\sigma}^2$ is unbiased.

LEAST SQUARES ESTIMATORS

In the method of least squares the form of the frequency function need not be known, but the method can be used if it is known. In many important cases, even if the form of the frequency function is unknown, the method of least squares gives rise to estimators that are unbiased and consistent and under certain conditions minimum variance unbiased. If the observations are uncorrelated and of equal weight the method is simply called unweighted least squares. If the assumption is made that the observations are uncorrelated and weighted unequally, the method will be referred to as weighted least squares. Generalized least squares is the treatment which assumes the observations to be correlated and to have the general multivariate normal distribution. In the case of weighted least squares the weighting matrix is a diagonal $n \times n$ matrix W whose elements are the weights W_i.

First we will derive the estimates for the case of weighted least squares. The estimates for unweighted least squares are then easily obtained by taking the weighting matrix to be the identity matrix. Consider

$$\delta y(t) = H(t)\ \delta x(t_0) + e$$

and assume $E[e] = 0$, $E[e\,e^T] = \sigma^2\ W^{-1}$. Let the least squares value of δx, i.e., the value of $\delta x(t_0)$ as obtained from the normal equations, be denoted by $\hat{\delta} x\ (t_0)$. Let $\sqrt{W}$ be a diagonal matrix of order n the elements of the diagonal being the square root of the weights W_i. In the case of equal weighting, $\sqrt{W}$ is the unit or identity matrix. Let the column matrix $\sqrt{W}\ e$ of residuals be defined thus:

$$\sqrt{W}\ e = \sqrt{W}\ (\delta y - H\ \delta x)$$

The sum of the squares of the residuals, $e^T We$, may be written

$$e^T\ We = (\delta y - H\hat{\delta} x)^T\ W\ (\delta y - H\ \hat{\delta} x)$$

where

$$W = \sqrt{W^T}\,\sqrt{W} = \sqrt{W}\,\sqrt{W}\,.$$

Hence

$$e^T We = [(\delta y)^T - (\hat{\delta} x)^T H^T]\, W\, (\delta y - H \hat{\delta} x)$$

$$= (\delta y)^T W (\delta y) - (\delta y)^T WH(\hat{\delta} x) - (\hat{\delta} x)^T H^T W \delta y + (\hat{\delta} x)^T H^T WH (\hat{\delta} x)\,.$$

By the method of least squares we shall find the value $\hat{\delta} x$ such that the sum of weighted residuals $e^T We$ is a minimum. The value of δx that minimizes $e^T We$ is given by the solution to

$$\frac{\partial}{\partial \delta x}(e^T We) = \frac{\partial}{\partial \delta x}[(\delta y)^T W \delta y - (\delta y)^T WH(\hat{\delta} x) - (\hat{\delta} x)^T H^T W (\delta y) +$$

$$(\hat{\delta} x)^T (H^T WH)\, (\hat{\delta} x)] = 0$$

$$= 0 - H^T W \delta y - H^T W \delta y + 2\, H^T WH \delta x = 0$$

$$= 2\, H^T WH \delta x - 2\, H^T W \delta y = 0$$

The weighted least squares estimate of $\delta x\, (t_0)$ is therefore,

$$\delta x\, (t_0) = (H^T WH)^{-1}\, H^T W \delta y$$

which is the same as the maximum-likelihood estimate under normal theory. To examine $\hat{\delta} x\, (t_0)$ for unbiasedness, we proceed as follows:

$$E\,[\hat{\delta} x] = E\,[(H^T WH)^{-1} H^T W \delta y] = (H^T WH)^{-1} H^T WE\,[\delta y]$$

$$= (H^T WH)^{-1} H^T WE\,[H \delta x + e] = (H^T WH)^{-1}\,(H^T WH)\,\delta x = I\,\delta x = \delta x\,.$$

So $\hat{\delta} x$ is an unbiased estimate of δx.

The covariance matrix of $\hat{\delta}x$ is

$$\text{cov}\,(\hat{\delta}x) = E\,[(\hat{\delta}x - \delta x)\,(\hat{\delta}x - \delta x)^T]$$

$$= E\ [(H^TWH)^{-1}\,H^TW\,\delta y - \delta x)\,((H^TWH)^{-1}\,H^TW\,\delta y - \delta x)^T]\ .$$

If we substitute $H\,\delta x + e$ for δy we get

$$\text{cov}\,(\hat{\delta}x) = E\ [\{(H^TWH)^{-1}\,H^TW\,(H\,\delta x + e) - \delta x\}$$

$$\{(H^TWH)^{-1}\,H^TW\,(H\,\delta x + e) - \delta x\}^T]$$

$$= E\ [\{(H^TWH)^{-1}\,(H^TWH)\,\delta x + (H^TWH)^{-1}\,H^TWe - \delta x\}$$

$$\{(H^TWH)^{-1}\,(H^TWH)\,\delta x + (H^TWH)^{-1}\,H^TWe - \delta x\}]$$

$$= E\ [\{(H^TWH)^{-1}\,H^TWe\}\,\{(H^TWH)^{-1}\,H^TWe\}]^T$$

$$= E\ [\,(H^TWH)^{-1}\,H^TWe\,e^T\,WH\,(H^TWH)^{-1}]$$

$$= (H^TWH)^{-1}\,H^TW\,E\,[e\,e^T]\,WH\,(H^TWH)^{-1},\ E\,[e\,e^T] = \sigma^2\,W^{-1}$$

$$= (H^TWH)^{-1}\,H^TW\,\sigma^2\,(W^{-1}W)\,H\,(H^TWH)^{-1}$$

$$\text{cov}\,\left(\hat{\delta}x\,(t_0)\right) = \sigma^2\,(H^TWH)^{-1}.$$

In order to get an unbiased estimate $\hat{\sigma}^2$ of σ^2 we proceed with

$$e^T\,We = (\hat{\delta}x)^T\,[H^TWH\,\hat{\delta}x - H^TW\,\delta y] - (H^TW\,\delta y)^T\,\hat{\delta}x + (\delta y)^T\,W\,\delta y$$

$$= (\delta y)^T\,W\,\delta y - (H^T\,W\,\delta y)^T\,\hat{\delta}x$$

Since, by virtue of the normal equations, $(H^TWH)\,\hat{\delta}x - H^TW\,\delta y = 0$. Hence replacing $\hat{\delta}x$ by $(H^TWH)^{-1}\,H^T\,W\,\delta y$ gives

$$
\begin{aligned}
e^T W e &= (\delta y)^T W \delta y - (H^T W \delta y)^T (H^T W H)^{-1} H^T W \delta y \\
&= (\delta y)^T W \delta y - (\delta y)^T W H (H^T W H)^{-1} H^T W \delta y \\
&= (\delta y)^T [I - W H (H^T W H)^{-1} H^T] W \delta y .
\end{aligned}
$$

Now let $E[\delta y]$ be denoted by η. If the δy's are all free of error, i.e., if $\delta y = \eta = E[\delta y]$, it is clear from

$$
0 = E[e] = E[\delta y - H \hat{\delta} x] = E[\delta y] - H E[\hat{\delta} x] = \delta y - H \delta x = e
$$

that $e = 0$. Consequently

$$
e^T W e = \eta^T [I - W H (H^T W H)^{-1} H^T] W \eta = 0 .
$$

Hence

$$
E[e^T W e] = E[(\delta y)^T \{I - W H (H^T W H)^{-1} H^T\} W \delta y + \eta^T \{I - W H (H^T W H)^{-1} H^T\} W \eta]
$$

$$
= E[(\delta y - \eta)^T (I - W H (H^T W H)^{-1} H^T) W (\delta y - \eta)] .
$$

We shall now invoke a theorem on the expected value of a quadratic form. Let

$$
Y = \delta y - \eta, \quad E[\delta y - \eta] \square 0, \quad E[Y Y^T] = \sigma^2 W^{-1} .
$$

Let

$$
(b_{ij}) = I - W H (H^T W H)^{-1} H^T, \quad W = \mathrm{diag}\,(W_1, W_2, \cdots, W_n) .
$$

Let

$$
A = [I - W H (H^T W H)^{-1} H^T] W .
$$

then

$$E[Y^T AY] = E\left[\sum_{i=1}^{n} b_{ii} W_i y_i^2\right] = b_{11} W_1 E[y_1^2] + b_{22} W_2 E[y_2^2] + \cdots + b_n W_n E[y_n^2].$$

$$= b_{11} (W_1 \sigma_1^2) + b_{22} (W_2 \sigma_2^2) + \cdots + b_n (W_n \sigma_n^2)$$

$$= \sigma^2 \sum_{i=1}^{n} b_{ii} \quad (W_i = \sigma^2/\sigma_i^2; \; i = 1, 2, \cdots, n)$$

$$= \sigma^2 \operatorname{tr} [I - WH(H^T WH)^{-1} H^T].$$

Therefore,

$$E[e^T We] = \sigma^2 \operatorname{tr} [I_n - WH(H^T WH)^{-1} H^T]$$

$$= \sigma^2 [\operatorname{tr} I_n - \operatorname{tr} WH(H^T WH)^{-1} H^T] \quad (\operatorname{tr} A+B = \operatorname{tr} A + \operatorname{tr} B)$$

$$= \sigma^2 \left[n - \operatorname{tr} (H^T WH) (H^T WH)^{-1}_{p x p}\right] \quad (\operatorname{tr} ABC = \operatorname{tr} CAB)$$

$$= \sigma^2 [n - \operatorname{tr} I_p]$$

$$= \sigma^2 (n - p)$$

since I_n and I_p are unit matrices of orders n and p respectively. In a weighted least squares fit to the data an unbiased estimate $\hat{\sigma}^2$ of σ^2 is given by

$$\hat{\sigma}^2 = \frac{e^T We}{n - p}$$

since

$$E[\hat{\sigma}^2] = \frac{1}{n-p} E[e^T We] = \frac{1}{n-p} \sigma^2 (n - p) = \sigma^2.$$

Given the linear relation between observations and the required orbit information

$$\delta y\,(t) = H(t)\;\delta x\,(t_0) + e\,,$$

the following estimates obtain (a) unweighted least squares

$$\hat{\delta} x\,(t_0) = (H^T\,H)^{-1}\,H^T\,\delta y$$

(b) weighted least squares

$$\hat{\delta} x\,(t_0) = (H^T W H)^{-1}\,H^T\,W\,\delta y$$

(c) generalized least squares (correlated observations)

$$\hat{\delta} x(t_0) = (H^T\,Q^{-1}\,H)^{-1}\,H^T\,Q^{-1}\,\delta y$$

In the case of correlated observations it is necessary to calculate the inverse of the covariance matrix Q and this is especially difficult when the Q matrix is ill conditioned.

BAYES ESTIMATORS

Bayes methods is a name given to statistical methods that introduce distributions of parameters at some stage of their development. Bayes procedures are based on information, experiences, and hunches of the individual. Their motivation is that, although the state of nature may not be known, it is unusual when one does not have some information about the state of nature, which could and should be used to assist in a decision. The Bayes principle assigns to each possible action a measure of the consequences of that action, the Bayes solution then being to take the action with the minimum Bayes measure. This measure is taken to be the average of the losses for a given action with respect to the prior distribution on the state.

A decision rule or strategy that is to lead to a decision based on the observations y must take into account that y can have many values. A rule for choosing an action is not then complete until it prescribes an action a for each conceivable value of y. Such a rule is called a statistical decision function: $a = d\,(y)$.

This is a mapping from the space of possible observation points to the space of actions. The action taken is random, since the data to which the decision rule is applied are random. Hence, the loss is random: $\ell(\theta, d(y))$. It is customary to base the analysis on the expected value of this loss:

$$R(\theta, d) = E_\theta\ [\ell(\theta, d(y)] .$$

The average is computed with respect to the distribution of the data y. This distribution of y depends on θ, so the dependence of the function $R(\theta, d)$ on θ enters through the θ in $\ell(\theta, a)$ and also through the θ in the distribution of y. The expected loss $R(\theta, d)$ is called the risk function. In many situations, θ, is or may be regarded as a random variable and we may have an a priori knowledge of its probability density function $g(\theta)$. Assuming continuous distributions, we find the density of the posterior distribution $h(\theta|y)$ as follows:

$$h(\theta|y) = \frac{g(\theta)\ f(y|\theta)}{f(y)}$$

where $g(\theta)$ is the prior density, $f(y|\theta)$ is the conditional probability density of the data y given that the parameters have the value θ, and $f(y)$ is the absolute density of y defined by

$$f(y) = \int f(y|\theta)\ g(\theta)\ d\theta = E_g\ [f(y|\theta)] .$$

The posterior distribution can be used as a tool in computing Bayes strategies. The method is to compute the posterior expected loss:

$$E_n\ [\ell(\theta, a)] = \int \ell(\theta, a)\ h(\theta|y)\ d\theta$$

and then to determine a that minimizes this quantity. The dependence of the chosen a on the given y provides a decision function $a = d^*(y)$, which is precisely the Bayes decision function for the given prior distribution. To establish this claim, we manipulate the expression for the Bayes risk corresponding to a decision function d(y):

$$B(d) = E_g\ [R(\theta, d)] = \int R(\theta, d)\ g(\theta)\ d\theta$$

$$= \iint \ell\big(\theta, d(y)\big)\ \ f(y|\theta)\ g(\theta)\ dy\ d\theta$$

or, if the order of integration is interchanged, as

$$B(d) = \int [\int \ell\big(\theta, d(y)\big) \; h(\theta \mid y) \, d\theta] \; f(y) \, dy$$

$$= \int E_n \; [\ell(\theta, d(y)] \; f(y) \, dy .$$

Now the inner integral is the second moment about the point $a = d(y)$. Since the second moment of a variable is a minimum when it is taken about the mean of the variable, it follows that this integral is minimized for each value of y if d is chosen as the mean of the conditional distribution of θ for y fixed. From the last line of B(d) we see that the integral is minimized by a function d(y) whose value for each y is selected to minimize the expected posterior loss, $E_n[\ell(\theta, a)]$.

In summary, the Bayes estimator of θ is a function $d(y) = E[\theta \mid y]$. The problem of finding a Bayes solution is now seen to reduce to the problem of finding the conditional expected value of the parameter θ when the variable y is held fixed. This solution is quite general because y may be chosen to be a statistic, such as a maximum likelihood estimate, from which an estimate $a = d(y)$ is to be constructed, or it may be a vector variable. Consequently, in estimating the mean of a normal distribution, y might well represent a vector variable.

Like the maximum likelihood estimators, it can be shown under quite general conditions that the Bayes estimator is consistent, efficient, and sufficient. In addition it can be shown that the Bayes estimator differs from the maximum likelihood estimator by an amount which is small compared with $1/\sqrt{n}$, where n is the sample size.

STATISTICAL FILTER THEORY ESTIMATORS (KALMAN-SCHMIDT)

In reference [2] R. E. Kalman presented a new approach to linear filtering and prediction problems. He introduced an alternative approach to a linear dynamic system (Wiener filter) which accomplishes the prediction of a random signal. In this paper is found the formulation and solution of the Wiener problem using state and state transition concepts from control theory.

R. E. Kalman developed optimal estimates using the idea of orthogonal projection in a multidimensional vector space (linear manifold). A careful analysis, reference [9], reveals that the Bayes estimation and the filter theory approach are basically the same. The two approaches differ in (1) the manner in which the linearization of a basically non-linear process is performed and in assuming Gaussian distributions, and (2) in the type of equations used in estimation. In Bayes estimation the equations are expressed in a manner whereby the estimate is obtained by operating on the entire set of observations at a time. In the filter theory approach the procedure for estimation involves taking the observations one at a time in the order of occurrence. It has been adequately shown in reference [9] that for the case of uncorrelated observations the methods are equivalent and the Bayes estimate is identical to the estimate obtained by the filter theory approach.

There are some fundamental differences between the classical least-squares method and the Kalman-Schmidt technique. First the least squares estimate is based on the minimization of the sum of squares of weighted residuals and the Kalman-Schmidt technique minimizes the expected value of a certain risk function. This will be explained in detail in the derivation of the Kalman-Schmidt technique. Secondly, the data is processed serially instead of in parallel as in least squares.

The tracking data is processed serially and the nominal orbit is updated as follows: A priori estimates of the state vector and its associated error exist prior to the actual processing of tracking data and before an improved estimate of the orbit is obtained. Both a priori estimates are updated to an instant of time when tracking data becomes available. Revised estimates of the nominal trajectory and the associated covariance matrix are obtained based upon the tracking data available for this time. These revised estimates remain until additional data becomes available and then the estimates are again updated to the time of the new observation. This process continues until all observations have been used.

There are computational advantages which exist when the Kalman-Schmidt technique is employed but do not exist in the method of least squares.

The Kalman-Schmidt method requires a fewer number of iterations than the method of least squares. This is so because the linear assumptions required for the updating theory are violated to a lesser degree than in the method of least squares where the estimate of the state vector is based on an entire sequence of observational residuals spread over an extended time arc. Another advantage of the Kalman-Schmidt method is that it avoids the inversion of large matrices. In least squares the order of the matrix to be inverted is equal to the number of quantities to be estimated while in the Kalman-Schmidt method the order of the matrix to be inverted is equal to the number of different types of tracking data available at a particular instant of time. The Kalman-Schmidt technique processes the tracking data sequentially and updates the orbit continuously. In the least squares method the data is processed in batches and when additional data is received the normal equations are again solved to obtain a new estimate of the state vector.

DERIVATION OF THE KALMAN-SCHMIDT METHOD

Let the nonlinear equations of motion in rectangular coordinates be given by

$$\begin{aligned}\ddot{x} &= g_1\,(x,\ y,\ z,\ t)\\ \ddot{y} &= g_2\,(x,\ y,\ z,\ t)\\ \ddot{z} &= g_3\,(x,\ y,\ z,\ t).\end{aligned} \tag{1}$$

It is desirable to obtain linear differential equations that represent perturbations of the actual trajectory from the reference. The usual assumption is made that the perturbations to the coordinates are small and that equations (1) can be expanded in a Taylor series retaining only terms of first order. For example, the equation for $\ddot{x}$, when expanded about the reference position x_R, y_R, z_R, becomes

$$\ddot{x} \simeq g_1\,(x_R,\ y_R,\ z_R) + \left(\frac{\partial g_1}{\partial x_R}\right)_{t=t_R} (x-x_R)$$

$$+ \left(\frac{\partial g_1}{\partial y_R}\right)_{t=t_R} (y-y_R) + \left(\frac{\partial g_1}{\partial z_R}\right)_{t=t_R} (z-z_R). \tag{2}$$

The equations for $\ddot{y}$ and $\ddot{z}$ are written similarly. The partial derivatives are evaluated at the reference position and it is to be understood that the reference quantities x_R, y_R, z_R, are varying functions of time obtained from a reference trajectory. If the reference is the estimated trajectory, then the partial derivatives are evaluated each time a new estimate of the state vector is obtained. Define a deviation state vector $\delta x(t)$ to be the vector of deviations of position and velocity from reference

$$\delta x(t) = \begin{bmatrix} \delta x_1(t) \\ \delta x_2(t) \\ \delta x_3(t) \\ \delta x_4(t) \\ \delta x_5(t) \\ \delta x_6(t) \end{bmatrix} = \begin{bmatrix} x(t) - x_R(t) \\ y(t) - y_R(t) \\ z(t) - z_R(t) \\ \dot{x}(t) - \dot{x}_R(t) = \delta\dot{x}_1(t) \\ \dot{y}(t) - \dot{y}_R(t) = \delta\dot{x}_2(t) \\ \dot{z}(t) - \dot{z}_R(t) = \delta\dot{x}_3(t) \end{bmatrix} . \tag{3}$$

With the above definition equations (2) take the form

$$\begin{aligned} \delta\ddot{x}_1 &= \left(\frac{\partial g_1}{\partial x_R}\right)_{t=t_R} \delta x_1 + \left(\frac{\partial g_1}{\partial y_R}\right)_{t=t_R} \delta x_2 + \left(\frac{\partial g_1}{\partial z_R}\right)_{t=t_R} \delta x_3 \\ \delta\ddot{x}_2 &= \left(\frac{\partial g_2}{\partial x_R}\right)_{t=t_R} \delta x_1 + \left(\frac{\partial g_2}{\partial y_R}\right)_{t=t_R} \delta x_2 + \left(\frac{\partial g_2}{\partial z_R}\right)_{t=t_R} \delta x_3 \\ \delta\ddot{x}_3 &= \left(\frac{\partial g_3}{\partial x_R}\right)_{t=t_R} \delta x_1 + \left(\frac{\partial g_3}{\partial y_R}\right)_{t=t_R} \delta x_2 + \left(\frac{\partial g_3}{\partial z_R}\right)_{t=t_R} \delta x_3 . \end{aligned} \tag{4}$$

It is desirable to reduce the set of equations (4) to a set of first order equations

$$\begin{aligned} \delta\dot{x}_1 &= \delta x_4 \\ \delta\dot{x}_2 &= \delta x_5 \\ \delta\dot{x}_3 &= \delta x_6 \\ \delta\dot{x}_4 &= \frac{\partial g_1}{\partial x_R} \delta x_1 + \frac{\partial g_1}{\partial y_R} \delta x_2 + \frac{\partial g_1}{\partial z_R} \delta x_3 \end{aligned}$$

$$\delta \dot{x}_5 = \frac{\partial g_2}{\partial x_R} \delta x_1 + \frac{\partial g_2}{\partial y_R} \delta x_2 + \frac{\partial g_2}{\partial z_R} \delta x_3$$

$$\delta \dot{x}_6 = \frac{\partial g_3}{\partial x_R} \delta x_1 + \frac{\partial g_3}{\partial y_R} \delta x_2 + \frac{\partial g_3}{\partial z_R} \delta x_3 . \tag{5}$$

It is convenient to consider the linear system of perturbation equations (5) in matrix notation as

$$\delta \dot{x} = F(t) \ \delta x(t) \tag{6}$$

where δx (t) is a 6 vector and F(t) is a continuous matrix function of the argument t in some interval (a, b)

$$F(t) = \left[\begin{array}{ccc|ccc} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ \hline \frac{\partial g_1}{\partial x_R} & \frac{\partial g_1}{\partial y_R} & \frac{\partial g_1}{\partial z_R} & 0 & 0 & 0 \\ \frac{\partial g_2}{\partial x_R} & \frac{\partial g_2}{\partial y_R} & \frac{\partial g_2}{\partial z_R} & 0 & 0 & 0 \\ \frac{\partial g_3}{\partial x_R} & \frac{\partial g_3}{\partial y_R} & \frac{\partial g_3}{\partial z_R} & 0 & 0 & 0 \end{array}\right]_{6\times 6} = \begin{bmatrix} \underset{3\times 3}{0} & \underset{3\times 3}{I} \\ \underset{3\times 3}{G(t)} & \underset{3\times 3}{0} \end{bmatrix}, \tag{7}$$

where

$$G(t) = \begin{bmatrix} \frac{\partial g_1}{\partial x_R} & \frac{\partial g_1}{\partial y_R} & \frac{\partial g_1}{\partial z_R} \\ \frac{\partial g_2}{\partial x_R} & \frac{\partial g_2}{\partial y_R} & \frac{\partial g_2}{\partial z_R} \\ \frac{\partial g_3}{\partial x_R} & \frac{\partial g_3}{\partial y_R} & \frac{\partial g_3}{\partial z_R} \end{bmatrix}_{3\times 3} . \tag{8}$$

We consider the system of differential equations (6)

$$\delta \dot{x} = \delta \frac{dx}{dt} = \frac{d}{dt} (\delta x) = F(t)\ \delta x(t) \tag{9}$$

where F (t) is a continuous matrix function of the argument t in some interval (a, b). An integral matrix of the system (9) shall be defined as a square matrix $[U(t)]_{6\times 6}$ whose columns are 6 linearly independent solutions of the system. Since every column of U(t) satisfies (9), the integral matrix U (t) satisfies the equation

$$\frac{dU}{dt} = F(t)\ U. \tag{10}$$

From the theorem on the existence and uniqueness of the solution of a system of differential equations it follows that the integral matrix U (t) is uniquely determined when the value of the matrix for some initial value $t = t_0$ is known, $U(t_0) = U_0$. We can take an arbitrary nonsingular square matrix of order n for the matrix U_0. In particular if $U(t_0) = I$, the integral matrix U (t) will be called normalized. Let us define Ω^{t}_{t0} (F) or simply Ω^{t}_{t0} as the normalized solution of (10) and call this solution the matricant or state transition matrix.

Ω^{t}_{t0} (F) can be obtained one column at a time by solving (9) 6 times each with a different member of $\delta x(t_0)$ set equal to unity and all the other elements set equal to zero. After Ω^{t}_{t0} (F) has been obtained the solution of (9) for $\delta x(t)$ is given by

$$\delta x(t) = \Omega^{t}_{t0} (F)\ \delta x(t_0) \tag{11}$$

where $\delta x(t_0)$ is the given set of initial conditions. The transition in the states of the system of equations from time t_0 to time t is given by (11) and hence relates deviations from the reference trajectory at time t to the initial deviations at time t_0.

The equations that relate the observables to the state variables are also linearized. Let y be the n vector of observations. Let $\hat{y}$ ($\hat{x}(t)$; t) be the estimate of the measured data based on the current estimate of the trajectory. The first order Taylor expansion about a reference trajectory then gives

$$\delta y(t) = y(t) - \hat{y}(t) = H(t)\ \delta x(t) + e \tag{12}$$

where H(t) is the matrix given by equation (4) in the section called "conventional differential orbit correction."

It is required to find the optimum estimate $\hat{\delta} x(t)$ from equation (12). It will be assumed that there exists a loss function which measures the loss incurred if incorrect estimates are made. Clearly the loss is a positive and non-decreasing function of the estimation error. It will also be assumed that the optimal estimate is a linear estimate, i.e., a linear function of the observations. Results obtainable by linear estimation can be improved by non-linear estimation only when the random processes are non-Gaussian and only then by considering at least third order probability distribution functions, reference [2].

The assumptions that were made about the optimal estimate permit one to regard this optimal estimator as a linear filter where the input signal consists of an observation corrupted by Gaussian noise and the output is the estimate of the state at present time. The optimal properties of the filter depend upon a certain weighting matrix K(t). The observation errors, e (t), are represented as the output of a linear dynamic system excited by white noise, a standard engineering trick, valid when only second-order statistics are concerned. The injection conditions are regarded as a vector valued random variable with zero mean. The matrix P is defined as being the covariance matrix of deviations between the actual and reference trajectories. For example at injection, the matrix P is the covariance matrix of injection errors and is assumed to be known. The covariance matrix P at any time t is defined as

$$P(t) = E\,[\delta x(t)\ \delta x^T(t)] \tag{13}$$

where $\delta x\ (t)$ is the deviation between the actual and reference trajectories. If $\delta x\ \ (t_0)$ is known then

$$\delta x(t_1) = \Omega_{t_0}^{t_1}\ \delta x(t_0). \tag{14}$$

By substitution of (14) into (13)

$$\begin{aligned} P(t_1) &= E\left[\delta x(t_1)\ \delta x^T(t_1)\right] = E\left[\Omega_{t_0}^{t_1}\ \ \delta x(t_0)\ \delta x^T(t_0)\ \left(\Omega_{t_0}^{t_1}\right)^T\right] \\ &= \Omega_{t_0}^{t_1}\ E\left[\delta x(t_0)\ \delta x^T(t_0)\right]\left(\Omega_{t_0}^{t_1}\right)^T \\ &= \Omega_{t_0}^{t_1}\ P(t_0)\left(\Omega_{t_0}^{t_1}\right)^T. \end{aligned}$$

In general

$$P(t_k) = \Omega_{t_{k-1}}^{t_k} \; P(t_{k-1}) \left(\Omega_{t_{k-1}}^{t_k}\right)^T . \qquad (15)$$

The covariance matrix of the observations as derived from (12) is

$$\begin{aligned} Y(t) &= E\,[\delta y(t)\;\delta y^T(t)] = E\,[(H(t)\;\delta x(t) + e)\;(H(t)\;\delta x(t) + e)^T] \\ &= E\,[(H(t)\;\delta x(t) + e)\;(\delta x^T(t)\;H^T(t) + e^T)] \\ &= H(t)\;E\,[\delta x(t)\;\delta x^T(t)]\;H^T(t) + E\,[e\;e^T] \\ &= H(t)\;\Omega_{t_0}^{t}\;P(t)\;H^T(t) + Q. \end{aligned} \qquad (16)$$

If the stochastic processes employed are assumed to be Gaussian then orthogonal projection as mentioned earlier is actually identical with conditional expectation. The approach to be taken here is one of applying concepts from statistical decision theory as explained earlier in the section called "Bayes estimators." Assuming continuous distributions, the density of the posterior distribution $h(\delta x \mid \delta y)$ is given by

$$h(\delta x \mid \delta y) = \frac{f(\delta y \mid \delta x)\;g(\delta x)}{f(\delta y)} \qquad (17)$$

where $g(\delta x)$ is the prior density, $f(\delta y \mid \delta x)$ is the conditional distribution of δy given a certain value of δx, and $f(\delta y)$ is the absolute density of δy defined by

$$f(\delta y) = \int f(\delta y \mid \delta x)\;g(\delta x)\;d(\delta x) = E_g\,[f(\delta y) \mid \delta x] .$$

The method is to compute the expected posterior loss

$$E_h\,[\ell(\delta x,\;\hat{\delta} x)] = \int_{-\infty}^{\infty} \ell(\delta x,\;\hat{\delta} x)\;h(\delta x \mid \delta y)\;d(\delta x) \qquad (18)$$

and then to determine $\hat{\delta} x$ which minimizes this quantity. Assume the loss function to be

$$\ell(\delta x,\;\hat{\delta} x) = \hat{\delta} x^T\;\hat{\delta} x . \qquad (19)$$

Then,

$$E\,[\ell(\delta x,\ \hat{\delta} x)] = \int_{-\infty}^{\infty} \hat{\delta} x^T\ \hat{\delta} x\ h(\delta x \,|\, \delta y)\ d(\delta x) \tag{20}$$

Since the second moment of a variable is a minimum when it is taken about the mean of the variable, it follows that this integral is minimized for each value of δy, if $\hat{\delta} x$ is chosen as the mean of the conditional distribution of δx for δy fixed. Hence the conditional mean is the optimal estimate, or

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

Let $Z = \binom{x}{e}$ be an augmented state vector where e is the noise vector in $\delta y(t) = H\,(t)\ \delta x(t) + e$. Let it be assumed that the state vector Z belongs to a normal distribution in order to simplify the formulas for updating the best estimate $\hat{\delta} x$. This assumption is not at all necessary since the results to be obtained hold under much more general conditions.

The problem of updating the estimate $\hat{\delta} x$ is divided into two parts: the first is to update $\hat{\delta} x$ when at time t new data $y\,(t)$ becomes available. The second is to update $\hat{\delta} x$ from t to $t + \Delta t$.

The covariance matrix of $Z - \hat{Z}$ is assumed to be of the form,

$$E\,[(Z-\hat{Z})\ (Z-\hat{Z})^T] = E\left[\begin{pmatrix} x-\hat{x} \\ e-\hat{e} \end{pmatrix} \begin{pmatrix} x-\hat{x} \\ e-\hat{e} \end{pmatrix}^T\right] = E\left[\begin{pmatrix} x-\hat{x} \\ e-\hat{e} \end{pmatrix}\ (x-\hat{x})^T\ (e-\hat{e})^T\right]$$

$$= E\begin{bmatrix} (x-\hat{x})\ (x-\hat{x})^T, & (x-\hat{x})\ (e-\hat{e})^T \\ (e-\hat{e})\ (x-\hat{x})^T & (e-\hat{e})\ (e-\hat{e})^T \end{bmatrix} = \begin{bmatrix} P & 0 \\ 0 & Q \end{bmatrix} = P_Z$$

and since e is assumed to be independent of x, P is the covariance matrix of δx, and Q is the covariance matrix of e.

We can write

$$y = M\,Z = (H_1^1\ I) \begin{pmatrix} x \\ e \end{pmatrix} = Hx + e\,.$$

It is assumed that $E[e] = 0$. Hence

$$E[Z] = \begin{pmatrix} E[x] \\ E[e] \end{pmatrix} = \begin{pmatrix} \hat{x} \\ 0 \end{pmatrix}.$$

We wish to obtain an improved value of the current estimate $\hat{x}$ when new information y is available.

The probability density function $P(Z)$ is

$$P(Z) = \frac{1}{|P_Z|^{1/2} (2\pi)^{n/2}} e^{-1/2 (Z-\hat{Z})^T P_Z^{-1} (Z-\hat{Z})}$$

where n is the number of components of Z. Since

$$y - \hat{y} = MZ - M\hat{Z} = M(Z-\hat{Z}),$$

$$E[(y-\hat{y})(y-\hat{y})^T] = E\left[M(Z-\hat{Z})\left(M(Z-\hat{Z})\right)^T\right] = E[M(Z-\hat{Z})(Z-\hat{Z})^T M^T]$$

$$= M E[(Z-\hat{Z})(Z-\hat{Z})^T] M^T = M P_Z M^T$$

so,

$$P(y) = \frac{1}{|M P_Z M^T|^{1/2} (2\pi)^{m/2}} e^{-1/2 (y-\hat{y})^T (M P_Z M^T)^{-1} (y-\hat{y})}$$

$$= \frac{1}{|M P_Z M^T|^{1/2} (2\pi)^{m/2}} e^{-1/2 (Z-\hat{Z})^T M^T (M P_Z M^T)^{-1} M (Z-\hat{Z})}$$

where m is the number of components of y.

Suppose we have two random vectors y and Z that have a multivariate normal distribution with means μ_1 and μ_2 respectively. Suppose we want to estimate the expected value of Z for a given y. This suggests the use of the conditional distribution of the multivariate normal. To obtain an analytical expression for P(Z|y), Bayes theorem is employed:

$$P(Z|y) = \frac{P(y|Z)\,P(Z)}{P(y)}$$

$$= \frac{(2\pi)^{\frac{m-n}{2}}}{|P_z|^{1/2}} |MP_zM^T|^{1/2}\, e^{-1/2} \left[(Z-\hat{Z})^T P_z^{-1}(Z-\hat{Z}) - (Z-\hat{Z})^T M^T (MP_zM^T)^{-1} M(Z-\hat{Z})\right]$$

The optimal estimate is the mean value of this density function, i.e., E (Z|y). The mean value is obtained when the exponent is minimized. Hence,

$$\frac{\partial}{\partial Z}\left[(Z-\hat{Z})^T P_z^{-1}(Z-\hat{Z}) - (Z-\hat{Z})^T M^T (MP_zM^T)^{-1} M(Z-\hat{Z})\right] = 0$$

or

$$2P_z^{-1}(Z-\hat{Z}) - 2M^T(MP_zM^T)^{-1}M(Z-\hat{Z}) = 0$$

$$P_z\left[P_z^{-1} - M^T(MP_zM^T)^{-1}M\right](Z-\hat{Z}) = 0$$

$$I(Z-\hat{Z}) - P_zM^T(MP_zM^T)^{-1}M(Z-\hat{Z}) = 0$$

$$Z-\hat{Z} = P_zM^T(MP_zM^T)^{-1}(y-\hat{y})$$

$$Z = \hat{Z} + P_zM^T(MP_zM^T)^{-1}(y-\hat{y}) = \hat{Z} + K(t)\,(y-\hat{y}).$$

Since

$$y = MZ \text{ and } \hat{y} = M\hat{Z}$$

$$P_Z M^T = \begin{pmatrix} P & 0 \\ 0 & Q \end{pmatrix} (H | I)^T = \begin{pmatrix} P & 0 \\ 0 & Q \end{pmatrix} \begin{pmatrix} H^T \\ I \end{pmatrix} = \begin{pmatrix} P H^T \\ Q \end{pmatrix},$$

$$M P_Z M^T = (H | I) \begin{pmatrix} P H^T \\ Q \end{pmatrix} = H P H^T + Q,$$

the new estimate Z_n is given by

$$\hat{Z}_n = \hat{Z}_p + \begin{pmatrix} P H^T \\ Q \end{pmatrix} (H P H^T + Q)^{-1} (y - \hat{y})$$

where $\hat{Z}_p$ represents the estimate prior to the time new data is used. Now

$$\hat{x}_n - \hat{x}_p = P H^T (H P H^T + Q)^{-1} (y - \hat{y}).$$

Premultiplication of this equation by H results in

$$H \Delta x = H P H^T (H P H^T + Q)^{-1} \Delta y$$

where

$$\Delta x = \hat{x}_n - \hat{x}_p, \quad \Delta y = y - \hat{y}.$$

This equation shows the weighting effect of Q. For example, if $Q = 0$ then $H \Delta x = (HPH^T)(HPH^T)^{-1} \Delta y = I \Delta y = \Delta y$.

Let the weighting matrix $K = P_Z M^T (MP_Z M^T)^{-1}$. Then $\hat{Z}_n = \hat{Z}_p + K(y - \hat{y})$ and the new covariance matrix is

$$E (Z_n - \hat{Z}_n)(Z_n - \hat{Z}_n)^T = E\left(Z_n - \hat{Z}_p - K(y - \hat{y})\right) \left(Z_n - \hat{Z}_p - K(y - \hat{y})\right)^T$$

$$= E\left((I - KM)(Z_n - \hat{Z}_p)\right) \left((I - KM)(Z_n - \hat{Z}_p)\right)^T$$

$$= (I-KM)\,E(Z_n-\hat{Z}_p)\,(Z_n-\hat{Z}_p)^T\,(I-KM)^T = (I-KM)\,P_Z\,(I-KM)^T$$

$$= (P_Z-KMP_Z)\,(I-(KM)^T)$$

$$= P_Z - KMP_Z - P_Z\,(KM)^T + KMP_Z\,(KM)^T$$

$$= P_Z - KMP_Z - P_Z M^T K^T + P_Z M^T (MP_Z M^T)^{-1}\,(MP_Z M^T)\,K^T$$

$$= P_Z - KMP_Z - P_Z M^T K^T + P_Z M^T K^T$$

$$= (I - KM)\,P_Z.$$

Hence the new estimate of P_Z is

$$P_{Z\,new} = (I - KM)\,P_{Z\,old}\,.$$

$$I - KM = I - P_Z\,M^T\,(MP_Z M^T)^{-1}\,M$$

$$= I - \begin{pmatrix} PH^T \\ Q \end{pmatrix} (MP_Z M^T)^{-1}\,(H|I).$$

If the y vector is of dimension one or the noise components are independent of one another and y can be treated one component at a time, then $HPH^T + Q = MP_Z M^T$ is a scalar.

In this case

$$I - \begin{pmatrix} PH^T \\ Q \end{pmatrix} (MP_Z M^T)^{-1}\,(H|I) = I - \frac{1}{HPH^T + Q} \begin{pmatrix} PH^T H & PH^T \\ QH & Q \end{pmatrix}$$

and

$$\hat{Z}_n = \hat{Z}_p + \frac{1}{HPH^T + Q} \begin{pmatrix} PH^T (y - \hat{y}) \\ Q\,(y - \hat{y}) \end{pmatrix}$$

$$\hat{y} = M\hat{Z} = H\hat{x} \text{ since } \hat{e} = 0$$

$$\begin{pmatrix} \hat{x}_n \\ \hat{e}_n \end{pmatrix} = \begin{pmatrix} \hat{x}_p \\ 0 \end{pmatrix} + \frac{1}{HPH^T + Q} \begin{pmatrix} PH^T y - PH^T H \hat{x}_p \\ Qy - QH\hat{x}_p \end{pmatrix} .$$

If the noise is time independent or the components are independent then when the P_Z matrix is updated n time only the P part remains. Hence, the only part of the P_Z matrix that need be kept is the lower right hand corner. For the case where y is a scalar

$$P_{Z\,new} = (I - KM)\, P_{Z\,old} = \left[I - \frac{1}{HPH^T + Q} \begin{pmatrix} PH^T H & PH^T \\ QH & Q \end{pmatrix} \right] \begin{bmatrix} P & 0 \\ 0 & Q \end{bmatrix}$$

$$\begin{bmatrix} P_{new} & 0 \\ 0 & Q_{new} \end{bmatrix} = \begin{bmatrix} P & 0 \\ 0 & Q \end{bmatrix} - \frac{1}{HPH^T + Q} \begin{pmatrix} PH^T HP & PH^T Q \\ QHP & Q^2 \end{pmatrix} .$$

Hence,

$$P_{new} = P - \frac{PH^T HP}{HPH^T + Q}$$

$$Q_{new} = Q - \frac{Q^2}{HPH^T + Q} .$$

The Matricant

We consider a system of differential equations

$$\frac{dx}{dt} = F(t)\ x \qquad (1)$$

where $F(t) = [F_{ik}]_{n\times n}$ is a continuous matrix function of the argument t in some interval (a, b). We use the method of successive approximations to determine a normalized solution of (1), i.e., a solution that for $t = t_0$ becomes the unit matrix (t_0 is a fixed number of the interval (a, b)). The successive approximations x_k $(k = 0, 1, 2, \ldots)$ are found from the recurrence relations

$$\frac{dx_k}{dt} = F(t)\, x_{k-1} \; (k = 1, 2, \ldots),$$

when x_0 is taken to be theunit matrix I.

Setting $x_k(t_0) = I$ $(k = 0, 1, 2, \ldots)$ we may represent x_k in the form

$$x_k = x_k(t_0) + \int_{t_0}^{t} F(\tau)\, x_{k-1}\, d\tau.$$

Thus,

$$x_0 = I,\; x_1 = I + \int_{t_0}^{t} F(\tau)\, d\tau,\; x_2 = I + \int_{t_0}^{t} F(\tau)\, x_1\, d\tau =$$

$$I + \int_{t_0}^{t} F(\tau) \left[I + \int_{t_0}^{\tau} F(\sigma)\, d\sigma \right] d\tau$$

or

$$x_2 = I + \int_{t_0}^{t} F(\tau)\, d\tau + \int_{t_0}^{t} F(\tau) \int_{t_0}^{\tau} F(\sigma)\, d\sigma\, d\tau, \ldots$$

i.e., x_k $(k = 0, 1, 2, \ldots)$ is the sum of the first $k + 1$ terms of the matrix series

$$I + \int_{t_0}^{t} F(\tau)\, d\tau + \int_{t_0}^{\tau} F(\tau) \int_{t_0}^{\tau} F(\sigma)\, d\sigma\, d\tau, \ldots \qquad (2)$$

The series (2) is absolutely and uniformly convergent in every closed subinterval of the interval (a, b) and determines the required solution of (1). This solution will be denoted by $\Omega^t_{t_0}$ (F) or simply $\Omega^t_{t_0}$ and is often called the matricant. The representation of the matricant in the form of such a series was first obtained by Peano in 1888.

Properties of the Matricant

1. $\Omega^t_{t_0} = \Omega^t_{t_1} \Omega^{t_1}_{t_0} \quad (t_0 , t_1, t \in (a, b))$

2. $\Omega^t_{t_0} (F + G) = \Omega^t_{t_0} (F) \Omega^t_{t_0} (G)$ with $G = [\Omega^t_{t_0} (F)]^{-1} G \Omega^t_{t_0} (F)$

3. $\ln |\Omega^t_{t_0} (F)| = \int_{t_0}^t \mathrm{tr}\, F \, d\tau$

We shall now show how to express by means of the matricant the general solution of a system of linear differential equations with right hand sides:

$$\frac{dx_i}{dt} = \sum_{k=1}^{n} F_{ik}(t)\, x_k + F_i(t) \; ;$$

F_{ik} (t) and F_i (t) (i, k = 1, 2, . . . , n) are continuous functions of t in some interval.

By introducing the column matrices $x = (x_1, x_2, \ldots, x_n)$ and $F = (f_1, f_2, \ldots, f_n)$ and the square matrix F, we write the system as follows:

$$\frac{dx}{dt} = F(t)\, x + f(t). \tag{3}$$

We shall look for a solution of this equation in the form

$$x = \Omega^t_{t_0} (F)\, z,$$

where z is an unknown column depending on t. We substitute this expression for x in (3) and obtain:

$$F\, \Omega^t_{t_0} (F)\, z + \Omega^t_{t_0} (F) \frac{dz}{dt} = F\, \Omega^t_{t_0} (F)\, z + f(t);$$

hence

$$\frac{dz}{dt} = [\Omega_{t_0}^{t} (F)]^{-1} f(t).$$

Integrating this, we find:

$$z = \int_{t_0}^{t} [\Omega_{t_0}^{\tau} (F)]^{-1} f(\tau)\, d\tau + c,$$

where c is an arbitrary constant vector. Substituting this expression in (3), we obtain:

$$x = \Omega_{t_0}^{t} (F) \int_{t_0}^{t} [\Omega_{t_0}^{\tau} (F)]^{-1} f(\tau)\, d\tau + \Omega_{t_0}^{t} (F)\, c \tag{4}$$

When we give to t the value t_0, we find: $x(t_0) = c$. Therefore (4) assumes the form

$$x = \Omega_{t_0}^{t} (F)\, x(t_0) + \int_{t_0}^{t} K(t, \tau)\, f(\tau)\, d\tau \tag{5}$$

where

$$K(t, \tau) = \Omega_{t_0}^{t} (F)\, [\Omega_{t_0}^{\tau} (F)]^{-1} \text{ (Cauchy matrix).}$$

Let us consider the matricant $\Omega_{t_0}^{t}$ (F). We divide the basic interval (t_0, t) into n parts by introducing intermediate points $t_1, t_2, \ldots, t_{n-1}$ and set $\Delta t_k = t_k - t_{k-1}$ $(k = 1, 2, \ldots, n;\ t_n = t)$. Then by property 1 of the matricant

$$\Omega_{t_0}^{t} = \Omega_{t_{n-1}}^{t} \ldots \Omega_{t_1}^{t_2}\, \Omega_{t_0}^{t_1}.$$

The multiplicative integral first introduced by Volterra in 1887 is given by

$$\Omega_{t_0}^{t} (F) = \int_{t_0}^{t} (I + F dt) = \lim_{\Delta t_k \to 0} [I + F(\tau_n)\Delta t_n] \ldots [I + F(\tau_1)\Delta t_1].$$

We now introduce the multiplicative derivative

$$D_t x = \frac{dx}{dt} \; x^{-1}, \text{ where } x = \Omega_{t_0}^{t} (F).$$

Updating the Estimate $\hat{x}(t)$ in Time, i.e., from t to $t + \Delta t$.

The linear perturbation equation appears in the form

$$\dot{x}(t) = F(t) \; x(t) + f(t)$$

where $x(t)$ is the state vector and $F(t)$ is the perturbation matrix. All solutions of this linear differential equation can be written in the form

$$x(t) = \Omega_{t_0}^{t} \; x(t_0)$$

where $x(t_0)$ is a vector of initial conditions at time t_0, $\Omega_{t_0}^{t}$ is the state transition matrix.

The solution of the differential equation is updated from t_0 to t by (5), i.e.,

$$x(t) = \Omega_{t_0}^{t} (F) \; x(t_0) + \int_{t_0}^{t} K(t, \tau) \; f(\tau) \, d\tau$$

where

$$K(t, \tau) = \Omega_{t_0}^{t} (F) \; [\Omega_{t_0}^{\tau} (F)]^{-1}.$$

REFERENCES

1. Duane Brown, A Matrix Treatment of the General Problem of Least Squares Considering Correlated Observations, BRL Report No. 937, May 1955

2. R. E. Kalman, A New Approach to Linear Filtering and Prediction Problems, Journal of Basic Engr. Vol. 82, No. 1, March 1960, pp. 35-50

3. T. A. Magness, J. B. McGuire, Comparison of Least Squares and Minimum Variant Estimates, Annals. of Mathematical Statistics, Vol. 33, June 1962

4. T. A. Magness, J. B. McGuire, Statistics of Orbit Determination Correlated Observations, Prepared for JPL Cal. Tech. by STL Inc., 1962

5. Alexander M. Mood, Franklin A. Graybill, Introduction to the Theory of Statistics, McGraw-Hill Book Co. 1963

6. Franklin A. Graybill, An Introduction to Linear Statistical Models, Vol. I, McGraw-Hill Book Co., 1961

7. H. Scheffe, The Analysis of Variance, John Wiley and Sons, 1959

8. Kenneth S. Miller, "Multidimensional Gaussian Distributions, SIAM, John Wiley and Sons, 1964

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

10. John Todd, Editor, Survey of Numerical Analysis, McGraw-Hill Book Co. 1962

11. Lionel Weiss, Statistical Decision Theory, McGraw-Hill Book Co. 1961

12. I. I. Shapiro, The Prediction of Ballistic Missile Trajectories from Radar Observations, McGraw-Hill Book Co., 1957

13. Gerald L. Smith, Multivariable Linear Filter Theory Applied to Space Vehicle Guidance, SIAM series A: Control, Vol. 2, No. 1, 1964

14. S. Pines, A Comparison of the Kalman and Bayes Minimum Variance Linear Filtering Methods with Time Correlated Noise, First Quarterly Report, Contract NAS 9-4036 April 1965

15. H. Engel, Comparison of Kalman-Schmidt and Linear Least Squares, Apollo Note No. 150, December 1963