NASA TECHNICAL NOTE

NASA TN D-5082

c. 1

NASA TN D-5082

A SUBOPTIMAL APPROXIMATION TO THE KALMAN FILTER

by Leon Bess

Electronics Research Center
Cambridge, Mass.

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION • WASHINGTON, D. C. • MARCH 1969

A SUBOPTIMAL APPROXIMATION TO THE KALMAN FILTER

By Leon Bess

Electronics Research Center
Cambridge, Mass.

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

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

ABSTRACT

A suboptimal estimation scheme similar to the Kalman filter is described which makes use of scalar weighting factors instead of matrix factors. It is shown that the accuracy degradation of the suboptimal estimator is not too great for most cases of practical interest. Moreover, it readily lends itself to physical interpretation.

A SUBOPTIMAL APPROXIMATION TO THE KALMAN FILTER

By Leon Bess
Electronics Research Center

SUMMARY

A suboptimal scheme similar to the Kalman filter is investigated here. The basic idea underlying the new device is that scalar weighting factors instead of matrix factors are used in constructing the estimate. It is quantitatively shown that the degradation in accuracy of the new estimator, in most cases, is not too great (typically around a factor of two). Aside from the possibility of simplifying the estimation procedure, the new device has the advantage of readily lending itself to physical interpretation. This, in turn, is shown to be useful in understanding the associated Kalman filter and allowing some significant a priori testing on it.

The treatment here is confined to the case where the dimensionality of the measurement vector is the same as that of the state vector, but it is suggested how the treatment could be extended to the general case.

I. - INTRODUCTION

The objective here is to investigate a certain type of estimator which is similar to the Kalman filter but is suboptimal in performance. The basic idea underlying the new device is that scalar weighting factors are used instead of matrix factors in construction of the estimate. It is shown that the degradation in accuracy of the suboptimal filter is, in most cases, not too great (typically being around a factor of two). An obvious advantage of the new filter is that, at least in certain cases, it can offer a significant simplification in the estimation procedure.

Another advantage is that the new filter readily lends itself to physical interpretation and this, in turn, may prove to be of value in understanding the associated Kalman filter. In particular, it will be shown that by calculating a certain factor, it should be possible to obtain a rough idea, a priori, of the absolute effectiveness of the Kalman estimator. The effectiveness criterion developed here is compared with the designation, "optimum", which heretofore has been the prevalent concept in estimator effectiveness. The a priori testing could be of practical value in constructing a computer program that would make the most effective use of the available facilities.

II. - PRELIMINARY MATHEMATICAL FORMULATION

This section is to be devoted to deriving those mathematical formulas which form the basis of the investigation to be described. The treatment here is to be limited to the case of discrete sampling intervals where the measurement data (in the form of the measurement vectors, z_{k-i}) are continuously supplied to the estimator at the discrete times, t_{k-i}, with a finite time interval, Δt_{k-i} $(= t_{k-i} - t_{k-i-1})$.

The Kalman-Bucy equations for this case are well known (refs. 1, 2, and 3). The canonical equations take the form:*

$$x(k+1) = \Phi(k+1,k)x(k) + \Gamma_{k+1,k}u(k) \tag{1a}$$

$$z(k) = H_k x(k) + v(k) \tag{1b}$$

The estimator equations are:

$$\hat{x}(k+1 \mid k) = \Phi(k+1,k)\hat{x}(k \mid k) \tag{2a}$$

$$\hat{x}(k \mid k) = (I - K_k H_k)\hat{x}(k \mid k-1) + K_k z(k) \tag{2b}$$

where

$$K_k \equiv \left[P(k \mid k-1)H_k^T\right]\left[H_k P(k \mid k-1)H_k^T + R_k\right]^{-1} \tag{2c}$$

The equation for the variance matrix, $P(k+1 \mid k)$ is:

$$P(k+1 \mid k) = \Phi(k+1 \mid k)P(k \mid k)\Phi^T(k+1,k) + \Gamma_{k+1,k}Q_k\Gamma^T_{k+1,k} \tag{3a}$$

$$P(k \mid k) = W_k P(k \mid k-1) \tag{3b}$$

---

*In regard to notation, the symbols used here are standard (such as employed in references 1 and 3) and are more or less self-explanatory in equation sets (1) through (3). In general, an upper case Greek or Latin letter will represent an (n×m) matrix. A lower case Latin letter will represent a (1×n) vector. A lower case Greek letter will represent a scalar.

where

$$W_k \equiv (I - K_k H_k)$$

It is understood that the matrices R_k and Q_k are defined by the following relations:

$$\operatorname{cov}[v(k),v(i)] = R_k \delta_{ik} \tag{3c}$$

$$\operatorname{cov}[u(k),u(i)] = Q_k \delta_{ik} \tag{3d}$$

$$\operatorname{cov}[u(k),v(i)] = 0 \tag{3e}$$

By making repeated use of the estimator equation set (2) for the time instants t_{k-i} (where $i = 0,1,2,3,\ldots,\infty$), it is possible to obtain an expression of the estimate which has the form:

$$\hat{x}(k \mid k) = \sum_{i=0}^{\infty} \Omega(k,k-i)\left[I - W_{k-i}\right]\left[H_{k-i}^{-1} z(k-i)\right] \tag{4}$$

where

$$\Omega(k,k-i) \equiv W_k \Phi(k,k-1) W_{k-1} \Phi(k-1,k-2) \ldots$$

$$\ldots W_{k+1-i} \Phi(k+1-i,k-i)$$

$$\text{for } i \geq 1$$

$$\Omega(k,k) \equiv I$$

It is, of course, apparent that the expression for $\hat{x}$ given in Eq. (4) is strictly valid only if the inverse H_{k-i}^{-1} exists. This assumption will be made here for the treatment which follows. Physically, this means that a complete zero bias estimate, $x_o(k \mid k)$, can be constructed only from the contemporary data, $z(k)$ and the dimensions of $z(k)$ must be the same as $x(k)$. Although in the interest of simplicity only the case just described is to be considered explicitly, it seems possible at this time that treatment developed here can be extended to the general case.*

In extending the i summation of Eq. (4) to infinity, it has been assumed that the estimator has reached a "steady state" where the initial measurement values, z(o), no longer influence the estimate, $\hat{x}$.

In a manner similar to that used in Eq. (4), the repeated application of the equation set (3) can bring the variance matrix to the form:

$$P(k \mid k) = \sum_{i=1}^{\infty} \Omega(k,k-i)\left[D_{k-i}\right]\Omega^{T}(k,k-i)$$

$$+ \sum_{i=0}^{\infty} \Omega(k,k-i)\left[I - W_{k-i}\right]\left[E_{k-i}\right]$$

$$\cdot \left[I - W_{k-i}^{T}\right]\Omega^{T}(k,k-i) \qquad (5)$$

where

$$D_{k-i} \equiv \Gamma_{k-i}^{*} Q_{k-i} \Gamma_{k-i}^{*T}$$

$$\Gamma_{k-i}^{*} \equiv \Phi^{-1}(k+1-i,k-i)\Gamma_{k+1-i,k-i}$$

$$E_{k-i} \equiv H_{k-i}^{-1} R_{k-i} H_{k-i}^{T-1} = \left[H_{k-i}^{T} R_{k-i}^{-1} H_{k-i}\right]^{-1}$$

In studying Eq. (4) for the optimum estimate, $\hat{x}$, it is possible to regard the matrix factors, W_{k-i}, as weighting factors. It becomes very suggestive to replace these matrix factors with scalar factors, and construct a new kind of estimate, $\hat{x}_A$.

It can be shown that a zero bias estimate can indeed be constructed to have the form:

$$\hat{x}_A(k \mid k) = \sum_{i=0}^{\infty} \gamma_{k,i}\Phi(k,k-i)\left[H_{k-i}^{-1} z(k-i)\right] \qquad (6)$$

*For details on how this might be done see Appendix C.

However, it is necessary that the scalar factors, $\gamma_{k,i}$, satisfy the following relation:

$$\sum_{i=0}^{\infty} \gamma_{k,i} = 1 \tag{7}$$

Moreover, the covariance matrix,

$$P_A(k \mid k) \left(= \text{cov} \left\{ \left[x(k) - \hat{x}_A(k \mid k) \ , \ x(k) - \hat{x}_A(k \mid k) \right] \right\} \right)$$

corresponding to $\hat{x}_A$ is given by:

$$P_A(k \mid k) = \sum_{i=1}^{\infty} \beta_{k,i}^2 \Phi(k,k-i) \left[D_{k-i} \right] \Phi^T(k,k-i)$$

$$+ \sum_{i=0}^{\infty} \gamma_{k,i}^2 \Phi(k,k-i) \left[E_{k-i} \right] \Phi^T(k,k-i) \tag{8a}$$

where

$$\beta_{k,i} \equiv \sum_{j=i}^{\infty} \gamma_{k,j} \tag{8b}$$

The scalar factors $\beta_{k,i}$ and $\gamma_{k,i}$ can be made to take a more suggestive form if they are defined in terms of a new scalar factor, α_{k-i}, such that:

$$\beta_{k,i} \equiv \alpha_k \alpha_{k-1} \alpha_{k-2}, \ldots, \alpha_{k+1-i}$$

$$\text{for } i \geq 1 \tag{9a}$$

$$\beta_{k,o} \equiv 1$$

Thus, it can be shown that if:

$$\gamma_{k,i} = \beta_{k,i} - \beta_{k,i-1} = \beta_{k,i}(1-\alpha_{k-i}) \tag{9b}$$

both Eqs. (7) and (8b), are satisfied. Moreover, Eqs. (4) and (6) for the estimates $\hat{x}$ and $\hat{x}_A$ have a direct correspondence since Eq. (6) results from Eq. (4) if W_{k-i} is replaced by α_{k-i}. [This is also true for the variance Eqs. (3a) and (5)].

At this point, a few speculative remarks about the scalar weighting factor, α_{k-i}, might be of interest.

Intuitively, it seems reasonable to assume that when the α's are adjusted to make $\hat{x}_A$ an optimum (i.e., when $tr[P_A(k \mid k)]$ is a minimum), that α_{k-i} will bear some close relationship to W_{k-i} such as having a value* near $1/n\ tr(W_{k-i})$, the mean characteristic root of W_{k-i}. This assumption will later be shown to be reasonably valid at least for the case of greatest practical interest. It follows from this that α_{k-i} should always have a value between zero and unity. This result is also a consequence of the consideration that the infinite series expression for $\hat{x}$ in Eq. (4) must converge so that a "steady state" is reached (i.e., a condition is reached where the estimate is no longer influenced by the initial $z(k - i)$ data).

III. - QUANTITATIVE COMPARISON OF ESTIMATORS

In this section, the suboptimal filter being studied here will be compared with its corresponding Kalman filter by evaluating the two performance indices $tr[P(k \mid k)]$ and $tr[P_A(k \mid k)]$. This will actually be done only for the two extreme cases of high "predictability" and low "predictability". The general case is very difficult to treat and moreover, the high "predictability" case is usually the one of greatest practical interest.

Low Predictability Case

It can be shown** that for the case of low "predictability" (or more explicitly where $\|D_{k-i}\| >> \|E_{k-i}\|$ for all i)*** that:

$$W_k \cong \left[H_k^{-1} R_k H_k^{T-1}\right] \left[P_k^*\right]^{-1} \left\{I - \left[P_k^* (H_k^T R_k^{-1} H_k)\right]^{-1}\right\} \tag{10}$$

*Where $tr(A) \equiv$ Trace A and n is the dimensionality of the square matrix.

**See Appendix A for details of the calculation.

***Where $\|A\| \equiv [tr(A^T A)]^{1/2}$.

where

$$P_k^* \equiv \left[H_k^{-1} R_k H_k^{T-1} + \Gamma_{k,k-1} Q_{k-1} \Gamma_{k,k-1}^T\right]$$

From Eq. (10), and its approximate relationship to W_k stated in Section II, it would follow that for this case $\alpha_{k-i} \ll 1$

If Eq. (10) is used in evaluating $P(k \mid k)$ from Eq. (3b) along with the approximation for $P(k \mid k - 1)$ developed in Eq. (5a) in Appendix A, the result is:

$$P(k \mid k) \cong \left[E_k\right] \left\{I - \left[E_k^{-1} P_k^*\right]^{-1}\right\} \tag{11}$$

In studying Eq. (11), it can be seen that the Kalman variance matrix, $P(k \mid k)$, (which is also the system error matrix) is determined almost entirely by the measurement error matrix, R_k, corresponding to the contemporary measurement vector, $z(k)$, (since the second matrix in the bracket is small compared to I). This result can be understood on purely physical grounds since as R_k grows smaller (and $\|E_k\|$ becomes small compared to $\|D_k\|$), the optimum estimate for the present instant of time, t_k, should be based more on $z(k)$ and less on any of the $z(k - i)(i \geq 1)$. The uncertainty in "updating" will start to produce errors large compared to E_k and when this happens, the $z(k - i)(i \geq 1)$ data can have little value in determining the optimum estimate and therefore must be excluded. Thus, for the case of low "predictability", the system error should be nearly the same as the measurement error and the Kalman filter provides little improvement over an estimator using only the current measurement data, $z(k)$. Since, as seen above $\alpha_{k-i} \ll 1$ for this case, it can readily be shown that $P_A(k \mid k) \cong E_k$ and the same statement can be made for the suboptimal filter. It is then apparent that the two performance indices $tr[P(k \mid k)]$ and $tr[P_A(k \mid k)]$ are nearly equal for low "predictability" cases.

High-Predictability Case

The case of high "predictability" (i.e., where $\|D_{k-i}\| \ll \|E_{k-i}\|$) to be considered in this section is much more difficult to treat generally. Some initial assumptions are now to be made which should not seriously reduce the generality of the treatment but will greatly simplify the calculations. The assumptions are that for $i = 0,1,2,\ldots,\infty$, the following relations are true:

$$t_{k-i} - t_{k-i-1} = \Delta t \;;\; D_{k-i} = D \;;\; E_{k-i} = E \;;$$

$$\Phi(k-i,k-i-1) = e^{F\Delta t} \;;\; W_{k-i} = I - \Delta W \;;\; \alpha_{k-i} = \alpha$$

The matrices, D, E, F, and ΔW are constants. All of the foregoing assumptions are reasonable if the estimator has reached its "steady state", where the various parameters show little variation in a time interval, $N_e \Delta t$, [where N_e is the "effective" number of contributions to the estimate from the various $z(k-i)$].

Using the foregoing assumptions in Eqs. (5) and (8a) for the two variance matrices, the result will be:

$$P(k \mid k) = \sum_{i=1}^{\infty} \left[(I - \Delta W)e^{F\Delta t}\right]^{i} [D] \left[e^{F^T\Delta t}(I - \Delta W^T)\right]^{i}$$

$$+ \sum_{i=0}^{\infty} \left[(I - \Delta W)e^{F\Delta t}\right]^{i} \left[(\Delta W)E(\Delta W^T)\right] \left[e^{F^T\Delta t}(I - \Delta W^T)\right]^{i} \tag{12}$$

$$P_A(k \mid k) = \sum_{i=1}^{\infty} (\alpha)^{2i} \left[e^{iF\Delta t} D e^{iF^T\Delta t}\right]$$

$$+ \sum_{i=0}^{\infty} (\alpha)^{2i} (1-\alpha)^2 \left[e^{iF\Delta t} E e^{iF^T\Delta t}\right] \tag{13}$$

The following relationship* will prove to be of value in the developments of this section:

$$(I - W_k) = P(k \mid k) E_k^{-1} \tag{14}$$

---

*This relationship can readily be derived from Eq. (VIII) of ref. 3.

As will be seen later, as $D_k \to 0$; $P(k \mid k) \to 0$. Thus, it follows that as $[\|D\|/\|E\|] \to 0$; $\Delta W \to 0$ and moreover $(1 - \alpha) \to 0$.

It can be shown that for the "high-predictability" case, the infinite sums of Eqs. (12) and (13) can be approximated by infinite integrals with the resulting errors being only of the order of $\|\Delta W\|$ and $(1 - \alpha)$. The actual integral for $P(k \mid k)$ would be of the form:

$$P(k \mid k) = \int_0^\infty dy\ e^{(\Delta F - \Delta W)y} \left[D + (\Delta W) E (\Delta W^T)\right] e^{(\Delta F^T - \Delta W^T)y} \qquad (15)$$

where

$$\Delta F \equiv F \Delta t$$

To obtain Eq. (15) from Eq. (12), the D integration lower limit was extended from 1 to 0, and the quantity $(I - \Delta W)$ was approximated by $e^{-\Delta W}$. Both of these operations would introduce errors only of the order of $\|\Delta W\|$. Similarly, the integral form for $P_A(k \mid k)$ can be shown to be:

$$P_A(k \mid k) = \int_0^\infty dy\ e^{(\Delta F - \Delta\alpha)y} \left[D + (\Delta\alpha)^2 E\right] e^{(\Delta F^T - \Delta\alpha)y} \qquad (16)$$

where

$$\Delta\alpha \equiv -\log \alpha \cong (1 - \alpha)$$

Again, the D integration lower limit was extended from 1 to 0, and error here is of the order of $\Delta\alpha$. The evaluation of integrals of the form given in Eqs. (15) and (16) has been treated extensively*, and the results can be expressed in the form of the following sets of algebraic matrix equations:

$$(\Delta W - \Delta F)P + P(\Delta W^T - \Delta F^T) = D + (\Delta W) E (\Delta W^T) \qquad (17)$$

---

*See, for example, "Introduction to Matrix Analysis" by R. Bellman, page 231; McGraw-Hill Book Company, Inc., 1960.

$$(\Delta\alpha - \Delta F)P_A + P_A(\Delta\alpha - \Delta F^T) = D + (\Delta\alpha)^2 E \tag{18}$$

As indicated by Eqs. (17) and (18), the operation of main interest here is obtaining explicit forms of tr(P) and tr(P_A) and comparing them. To do this, it is expedient to calculate the quantity, $\Delta P (\equiv P_A - P)$ from these equations. Thus, if Eq. (17) is subtracted from Eq. (18) and using the relation of Eq. (14) [$P = (\Delta W)E$], it is possible to obtain the following relation:

$$(\Delta\alpha - \Delta F)\Delta P + \Delta P(\Delta\alpha - \Delta F^T) + 1/2(\Delta\alpha - \Delta W)(P - \Delta\alpha E)$$

$$+ 1/2(P - \Delta\alpha E)(\Delta\alpha - \Delta W^T) = 0 \tag{19}$$

Since $P = E(\Delta W^T)$, it follows from Eq. (19) that:

$$\mathrm{tr}\left[(\Delta\alpha - \Delta F)\Delta P\right] = 1/2\,\mathrm{tr}\left[E(\Delta W^T - \Delta\alpha)^2\right] \tag{20}$$

Eq. (20) can be changed to a more suggestive form by again using the relation of Eq. (14) so that it becomes:

$$\mathrm{tr}\left[(I - B)\Delta P\right] = \left(\frac{1}{2\Delta\alpha}\right)\left[\mathrm{tr}(PE^{-1}P) - (2\Delta\alpha)\mathrm{tr}(P) + (\Delta\alpha)^2\mathrm{tr}(E)\right] \tag{21}$$

where

$$B \equiv \left(\frac{1}{2\Delta\alpha}\right)(\Delta F + \Delta F^T)$$

The l.h.s. of Eq. (21) results from the fact that ΔP is a symmetric matrix.

As can be seen from Eq. (21), tr(ΔP) is a function of the scalar, $\Delta\alpha$; and it will assume a minimum value for a certain choice of $\Delta\alpha$. In general, the variation of the matrix factor, $(I - B)$, will not greatly affect $(\Delta\alpha)_m$, the value of $\Delta\alpha$ for which tr(ΔP) is a minimum; so for simplicity, $(\Delta\alpha)_m$ is to be determined by minimizing the r.h.s. of Eq. (21). This can be done as a straightforward differentiation problem, but it is more suggestive to do it by making the following substitution:

$$\Delta\alpha = \delta_0 \sqrt{\eta} \tag{22}$$

where

$$\delta_0 \equiv \left[\mathrm{tr}(PE^{-1}P)/\mathrm{tr}(E)\right]^{1/2}$$

When Eq. (22) is substituted in Eq. (21) and both sides are divided by tr(P), the result is:

$$\frac{\mathrm{tr}[(I - B)\Delta P]}{\mathrm{tr}(P)} = [1/2(\sqrt{\eta} + 1/\sqrt{\eta})\sqrt{\xi} - 1] \tag{23}$$

where

$$\xi \equiv \frac{\mathrm{tr}(E) \cdot \mathrm{tr}(PE^{-1}P)}{\mathrm{tr}^2(P)}$$

It is evident that the r.h.s. of Eq. (23) (which represents a rough measure of the percentage deviation of P and P_A) is a minimum when the scalar parameter, $\eta = 1$, so that $(\Delta\alpha)_m = \delta_0$. Moreover, the minimum is fairly broad since η can vary from 1/2 to 2 and the r.h.s. of Eq. (23) will stay within six percent of its minimum value.

It is to be noted that both symmetric matrices ΔP and $(I - B)$ must be positive definite*. This is true for ΔP because P_A must always exceed P, the possible minimum. It is true for $(I - B)$, since otherwise the integral in Eq. (16) will become

---

*It should be noted at this point that the tentative assumption made in Section II about the magnitude of α_{k-i} can now be verified. From Eq. (22) it follows that $(1 - \alpha) = [\mathrm{tr}(\Delta W^T E \Delta W)/\mathrm{tr}(E)]^{1/2}$ whereas at the beginning of Section III it was, in effect, assumed that $(1 - \alpha) = (1/n)\mathrm{tr}(\Delta W)$. It can be shown that, in general, the two values will roughly agree (to within a factor of two).

infinite. It can therefore be shown that*:

$$\lambda_M \mathrm{tr}(\Delta P) \geq \mathrm{tr}[(I - B)\Delta P] \geq \lambda_m \mathrm{tr}(\Delta P) \tag{24a}$$

where

$$\lambda_M \geq \lambda_m \geq 0$$

λ_M is the maximum characteristic root of the matrix (I - B) and λ_m is the minimum characteristic root. Hence by evaluating λ_M and λ_m, the range of variation of tr $(\Delta P)/\mathrm{tr}(P)$ can be determined from Eqs. (23) and 24a). However, since in what follows only a rough estimate of $\mathrm{tr}(\Delta P)/t(P)$ is desired, Eq. (24a) suggests that the effects of the matrix factor, (I - B), can be approximated with sufficient accuracy (at least in most cases) by means of the following relation:

$$\mathrm{tr}[(I - B)\Delta P] \cong \lambda_B \mathrm{tr}(\Delta P) \tag{24b}$$

where

$$\lambda_B \equiv (1/n)\mathrm{tr}(I - B) = (1/n)\sum_{i=1}^{n} \lambda_i$$

The λ_i are the characteristic roots of (I - B).

It is apparent in studying Eq. (23) that the parameter, ξ, is of central importance in determining the value of $\mathrm{tr}(\Delta P)$. It is first to be noted that $\xi \geq 1$, since otherwise $\mathrm{tr}(\Delta P)$ might be negative. Moreover, (assuming that $\eta = 1$ so the factor in front of $\sqrt{\xi}$ is unity), it is seen that the quantity $[\mathrm{tr}(\Delta P)/\mathrm{tr}(P)]$ which is the measure of the accuracy degradation has a value roughly equal to $[(\sqrt{\xi} - 1)/\lambda_B]$. The value of ξ in any particular case, of course, depends on the nature of the estimation problem (i.e., on the exact form of the matrices D, E, and F); and it can assume values ranging from those near unity to others much greater than unity. It is felt that it might be of value to estimate the parameter, ξ, for a couple of selected examples which should illustrate the behavior of this parameter. This is to be undertaken in what follows, but first it is to be noted that in order to evaluate ξ, it is necessary to calculate P.

---

*See Appendix B for details of the derivation.

The operation can be accomplished by solving the system of algebraic equations derived from Eqs. (14) and (17) and represented in matrix form as:

$$D + P(\Delta F^T) + (\Delta F)P - PE^{-1}P = 0 \tag{25}$$

In obtaining a unique form of P from Eq. (25), the supplementary condition that as $D \to 0$, $P \to 0$ is to be used.

The matrix Eq. (25) is seen to closely resemble Eq. Set (3), the Kalman variance equations [it becomes identical if 0 on the r.h.s. is replaced by $\Delta t(dP/dt)$]. The main difference, of course, is that it is a set of algebraic rather than differential equations. The details of obtaining a practical solution to Eq. (25) is to be left for the following section. In this section, it is simply assumed that P can be made available from Eq. (25).

Having p_{ij} (the components of P), the parameter ξ can be calculated and can be shown to take the explicit form:

$$\xi = \frac{\left[\sum_i e_{ii}\right]\left[\sum_{ij} p^2_{ij}/e_{jj}\right]}{\left[\sum_{ij} p_{ii}p_{jj}\right]} \tag{26}$$

where $i, j = 1, 2, 3, \ldots, n$.

In obtaining Eq. (26), the simplifying assumption has been made that the matrix E is diagonal. This assumption is actually not too restrictive because it covers many, if not most, cases of practical interest. It is to be noted that $p^2{}_{ij} = c_{ij}p_{ii}p_{jj}$, where $-1 \le c_{ij} \le 1$ which follows from the definition of $P(k|k)$ ($\equiv$ P).

The illustrative case to be considered now is where $c_{ij} = 1$ and all the components of E, e_{ii}, are nearly equal. It is readily seen from Eq. (26) that $\xi = n$ here, and the degradation factor is $[(\sqrt{n} - 1)/\lambda_B]$. Thus, unless the matrix dimensionality is very large, it is seen that the accuracy degradation for the suboptimal estimator need not be too great for this case.

The next case to be considered is where one particular component of E, (say e_{11}), is much larger (at least a factor of 2n) than the others and, moreover, this will cause its corresponding P component, p_{11}, to be much larger (by at least a factor of

$2n[e_{11}/e_{ii}]^{1/2}$) than the others. For these conditions, it can be seen that $\xi \cong 1$ and the accuracy degradation factor is much smaller than unity. For this case, the suboptimal estimator is very nearly as good as the Kalman filter.

The two examples treated above were chosen because the first case represents an easily realizable situation where the suboptimal estimator would be performing at nearly its worst compared to a Kalman filter, and the second case represents a situation where the suboptimal estimator is performing nearly the best that is possible. In general, the situation might be roughly summarized by saying that the accuracy degradation factor will have a value of around 2. This is to say that the agreement between $P_A(k \mid k)$ and $P(k \mid k)$ is usually fairly close.

IV. - ESTIMATOR EFFECTIVENESS

One of the benefits to be derived from the viewpoint developed in the previous section is that it allows an absolute criterion in judging the effectiveness of an estimator. Heretofore, the designation, "optimum", has been prevalent in the question of estimator effectiveness. As will be seen, this is actually a vague criterion which is much more meaningful to a mathematician than to a physicist or engineer. It is proposed here that the standard against which any estimator should be compared should be the estimator which uses only contemporary data [where the estimate is $H_k^{-1}z(K)$]. In this light, the practical limitations of the designation, "optimum", become apparent when it is noticed that it is perfectly possible that one estimator be only about 1.1 times as accurate as its corresponding contemporary data estimator, while another estimator be 100 times as accurate and both could be optimum estimators.

It is also proposed here that the parameter which is the measure of estimator effectiveness be N_e; the "effective" number of contributions to the optimum estimate. From the central limit theorem of probability, it would follow that the error in the optimum estimate is roughly equal to the error in any one contribution divided by N_e. This would seem to fit in with the proposal to use the contemporary data estimator as a standard, since N_e can therefore be defined by the relation $tr[P(k \mid k)] = 1/N_e tr(E_k)$.

It would then follow from Eqs. (22) and (23) that

$$N_e \equiv \frac{tr(E)}{tr(P)} = \left(\sqrt{\xi}/\delta_0\right) \qquad (27)$$

Thus, it is seen that N_e can indeed be the measure of the effectiveness of an estimator, since $tr(E_k)$ is the measure of the overall error of a contemporary data estimator just as $tr(P)$ is the overall measure of its corresponding Kalman filter. The case of the Kalman filter is given by Eq. (27); and from this it would follow that the error reduction can be very large, since (as will be seen), δ_0 can be made quite small. The actual calculation of N_e requires the knowledge of the covariance matrix, $P[= P(k \mid k)]$. The evaluation of P is to be the main concern of the following section, and as will be seen, involves an iteration solution to the matrix algebraic equation, Eq. (25). Although there would be only one evaluation, the actual performance could be quite involved and generally requires the services of a digital computer. If only a rough a-priori evaluation of N_e be sufficient, the calculation could become quite simple. Drawing from the results of the following section, one approach is to approximate P by $\bar{P}$ of Eq. (34) (in Section V) and use this in the definition of N_e given in Eq. (27). N_e then becomes the quantity $(1/\sigma_0)$ [defined in Eq. (34)] which can easily be calculated. It is of interest to note that in certain cases [i.e., where $[\Delta t\ tr(FE)]^2 \gg tr(E) \cdot tr(D)$] $(1/\sigma_0)$ will simplify to $[tr(D)/tr(E)]^{1/2}$.

Using this rough estimate of N_e, it is apparent that as the "predictability" increases (i.e., as $\|D\|/\|E\| \to 0$), N_e also increases and eventually approaches infinity.

V. - SUBOPTIMAL ESTIMATOR

As will be shown in the following development, obtaining the optimum value of the scalar weighting factor, α, in the suboptimal scheme proposed here requires only the knowledge of the scalar factor $tr(FP)$ not the whole covariance matrix, P, as in the Kalman filter. Thus, only one scalar quantity is needed instead of n^2 quantities. Moreover, as can be seen from Eq. (23), the $tr(FP)$ need not, in most cases, be evaluated too accurately (since it was shown that the parameter, η, could vary a factor of 2 without greatly affecting the results). All of this would suggest that it should be possible to effect a significant simplification in the calculational procedure (especially for an x of a large number of dimensions) in instituting the suboptimal scheme rather than the Kalman filter. In what follows, one proposal will be suggested as to how this simplification can be realized. There probably exist other approaches which could perhaps even be more advantageous.

The implementation of the suboptimal estimator being proposed here is more or less the same as that of the Kalman filter, except that instead of Eq. (26), the corresponding estimator equation is:

$$\hat{x}_A(k \mid k) = \left[\hat{\alpha}_{k-\ell}\right]\hat{x}_A(k \mid k-1) + \left[1 - \hat{\alpha}_{k-\ell}\right]\left[H_k^{-1} z(k)\right] \tag{28}$$

where

$$\hat{\alpha}_{k-\ell} \equiv \left[1 - \delta_0\right]$$

$\hat{\alpha}_{k-\ell}$ in Eq. (28) is the optimum value of the weighting factor, α. It is to be calculated at the instant $t_{k-\ell}$. $\hat{\alpha}_{k-\ell}$ is determined by Eq. (22) which in turn requires the knowledge of $tr(PE^{-1}P)$. From Eq. (25) it follows that:

$$\hat{\alpha}_{k-\ell} = 1 - \left\{\frac{tr\left(D_{k-\ell}\right) + (2\Delta t)\, tr\left[F_{k-\ell} P(k-\ell \mid k-\ell)\right]}{tr\left(E_{k-\ell}\right)}\right\}^{1/2} \tag{29}$$

Since it has been assumed in the previous development that the basic estimator matrices D, E, and F are only slowly varying, staying practically constant in a time interval, $N_e\Delta t$, the weighting factor, $\hat{\alpha}_{k-\ell}$ need not be calculated more frequently than at intervals of $N_e\Delta t$ apart. Hence, it follows that the same $\hat{\alpha}_{k-\ell}$ be used over the whole time interval $t_{k-\ell}$ to t_k (where $\ell = N_e\Delta t$).

As indicated before, the main advantage of the suboptimal scheme resides in the simplification in the required evaluation of the covariance matrix, P. The proposed evaluation procedure is now to be presented in detail and its advantages pointed out.

The evaluation is accomplished by the approximate solution of the set of algebraic equations, Eq. (25), and it is to be noted that this equation set is valid for any given instant of time. Thus, if the matrices, D, E, and F, all correspond to the instant, $t_{k-\ell}$; then $P = P(k-\ell \mid k-\ell)$.

The method to be used in solving for P in Eq. (25) is an iteration procedure where the $(i-1)^{th}$ iteration of P is obtained from $(i)^{th}$ iteration by the relation:

$$\overset{(i+1)}{P} = \overset{(i)}{P} + \Delta P_i \tag{30}$$

By substituting Eq. (30) into Eq. (25), it can be shown that ΔP_i is in turn determined by the relation:

$$\Delta P_i \left[E^{-1}\left(\overset{(i)}{P} + \frac{1}{2}\,\Delta P_i\right) - \Delta F^T\right]$$

$$+ \left[\left(\overset{(i)}{P} + \frac{1}{2}\,\Delta P_i\right)E^{-1} - \Delta F\right]\Delta P_i = \Delta \overset{(i)}{V} \tag{31}$$

where

$$\Delta \overset{(i)}{V} \equiv D + \Delta F \overset{(i)}{P} + \overset{(i)}{P}\,\Delta F^T - \overset{(i)}{P}\,E^{-1}\,\overset{(i)}{P}$$

If $\overline{\Delta P_i}$ is defined so as to satisfy the relation:

$$\left[\left(\overset{(i)}{P} + \frac{1}{2}\,\Delta P_i\right)E^{-1} - \Delta F\right]\overline{\Delta P_i} = \frac{1}{2}\,\Delta \overset{(i)}{V} \tag{32a}$$

and thus

$$\overline{\Delta P_i} \cong \left[\left(\overset{(i)}{P} + \frac{1}{2}\,\Delta P_{i-1}\right)E^{-1} - \Delta F\right]^{-1}\left[\frac{1}{2}\,\Delta \overset{(i)}{V}\right] \tag{32b}$$

It can be seen that $\overline{\Delta P_i}$ is actually a solution to Eq. (31), but it is not symmetric (the matrix ΔV_i however is symmetric). It is to be expected that an approximate symmetric solution to Eq. (31) is of the form:

$$\Delta P_i = \frac{1}{2}\left(\overline{\Delta P_i} + \overline{\Delta P_i}^T\right) \tag{33}$$

Eqs. (30), (32b), and (33) form the working relations for the iteration cycle. There is now left only the task of finding the initial trial function, P_o. This is to be determined from the set of relations:

$$\overset{(o)}{P} \equiv \sigma_0 E \tag{34}$$

and

$$\Delta P_{-1} \equiv 0$$

where

$$\sigma_0 = \left[\frac{(\Delta t)\,|\,tr(FE)\,|}{tr(E)}\right]\left\{\left[\frac{tr(E)\quad tr(D)}{(\Delta t)^2 tr^2(FE)} + 1\right]^{1/2} - 1\right\}$$

The scalar factor, σ_o, has been chosen so that $tr(\Delta V_o) = 0$, assuming that $tr(FE) \leq 0$.

It is of interest to note here that Eq. (25) can be solved by exactly the same procedure used to solve the Riccati difference equation of the Kalman filter. This would be an iteration procedure similar to the one described above except that instead of using the relations of Eqs. (32b) and (33), the relation $\Delta P_i = \Delta V_i$ would be used. P_o would have some arbitrary value which would usually be far removed from the final assymtotic one. As a rough estimate, it can be shown that the number of interations that would be necessary for a reasonably accurate estimate of P would be of the order of $(1/\delta_o)$. For a high-"predictability" type of estimator, the quantity $(1/\delta_o)$ could very easily have values in the range from 10 to 100, and this would also be the number of iterations necessary when using standard Kalman procedure in solving for P.

For the proposed suboptimal iteration procedures [represented by Eqs. (30) to (33)], it is expected that not more than 3 or 4 iterations will be required (since, usually, it should be sufficient to calculated tr(FP) to within a factor of 2 of its true value). The decrease in the number of required iterations appears for the following reasons. The first is that P_o of Eq. (34) is much closer to the correct P [at least when it is used in tr(FP)] than the usual P_o used in the standard Kalman procedure. The second reason is that Eq. (32b) is used to calculate ΔP_i instead of the relation $\Delta P_i = \Delta V$. Thus, in each iteration ΔP_i advances toward its assymtatic value by a bigger step (a factor of the order of $\|[(P_i + \frac{1}{2}\Delta P_{i-1})E^{-1} - \Delta F]^{-1}\|$ larger) that it would have in the standard Kalman procedure.

Although the significant simplification in solving for P in the suboptimal scheme would seem to be established theoretically, it would be of great value to verify it experimentally. This would mean initiating a computer program solving various types of simulated estimator problems. The implementation of this program, however, must be left for a future investigation due to the large effort required.

VI. - CONCLUDING REMARKS

In summarizing the foregoing developments, it can be seen that probably the most important result derived is proof that the suboptimal estimator using scalar weighting factors can produce an estimate which, in most cases of interest, is nearly as good as the optimal Kalman estimate. In view of this result, the suboptimal scheme, being easier to understand physically, can become useful not only as a simplified estimator but also as a means of gaining additional insight into the associated Kalman filter. One of the results of this new insight is the possibility of an a priori evaluation of the effectiveness of a Kalman filter, which was shown to be more meaningful in practice than the designation "optimum".

Another by-product of the new insight is one that might appeal to the more practical physicists and engineers. This is the ability of being able to provide a rough qualitative description in physical terms of how a Kalman filter operates. The description is now to be presented in what follows, but first it is necessary to give a physical interpretation of the suboptimal estimation scheme. This is best done by studying Eq. (6) which, in essence, describes how the estimate, $\hat{x}_A(k \mid k)$, is constructed.

It has already been established in Section III that $\gamma_{k,i}$ is a scalar weighting factor whose value is given by Eq. (9b). The second factor, $[\Phi(k,k-i)H_{k-1}^{-1}z(k-i)]$, on the r.h.s. of Eq. (6) can be interpreted as the measurement taken at the time, t_{k-i}, converted to an estimate (by H_{k-i}^{-1}) and "updated" to the time, t_k, [by $\Phi(k,k-i)$]. It now becomes possible to give a qualitative explanation of how the estimator functions. It has already been shown in Section III that when the "predictability" is low (i.e., $\|D_{k-i}\| \gg \|E_{k-i}\|$) α_{k-i} becomes very small. Thus, the series in Eq. (6) converges rapidly with the "effective" number, N_e, of terms being small. This means that only a few of the contributions of the earlier measurements, $z(k-i)$ (where $i > 1$), are used in the estimate. Most of the contributions are being rejected because of the unreliability in the "updating" caused by the random forcing function $u(k-i)$ in the interval, $(t_k - t_i)$. On the other hand, if the "predictability" is high (i.e., if $\|D_{k-i}\| \gg \|E_{k-i}\|$), α_{k-i} can be shown to approach unity. The series in Eq. (6) now is slowly converging and the "effective" number of terms, N_e, is large. This, of course, implies that contributions from many of the earlier measurements are being used in the estimate.

Since (as has been shown in Section III) there is a fairly close agreement between $\hat{x}_A(k \mid k)$ and $\hat{x}(k \mid k)$ [or actually between $tr(P_A)$ and $tr(P)$], it can be inferred that the main processes taking place in the $\hat{x}_A(k \mid k)$ estimate should take place in the

$\hat{x}(k \mid k)$ estimate. Thus, as a result of studying the suboptimal filter, it does become possible to give the rough qualitative physical description mentioned above. One of the main functions of a Kalman filter is to essentially "update" the measurement data of the previous instants of time to a set of values corresponding to the present instant. These "updated" measurements are used along with the current measurements to form an optimum estimate of the state vector. It is to be noted that because of the random disturbance term, $u(k-i)$, in the "Canonical" equations, there is an uncertainty in "updating" the measurements of the previous instants. In fact, the earlier the instant, the greater will be the uncertainty. The Kalman filter scheme takes account of this fact by essentially assigning weighting factors to the contributions of the previous instants so that the weighting factor becomes smaller as the corresponding instant goes back in time. The fact that the Kalman filter uses matrix weighting factors so that each component in the state vector can be weighted individually probably accounts in part for its superiority in performance to the suboptimal scheme proposed here (using only scalar weighting factors).

As seen in Section IV, the more precise the knowledge of the underlying natural processes (i.e., the greater the "predictability") the greater will be the "effective" number, N_e, of the measurement vectors used to construct the optimum estimate and also the more accurate will be that estimate. Thus, it follows that the Kalman filter is a better estimator than one using only contemporary measurement data only because it has available extra information in the form of partial knowledge of the natural processes generating the measurement data (i.e., it has the "Canonical" equations), and it uses this information to supply itself with extra data derived from the measurements of the previous instants.

National Aeronautics and Space Administration
Electronics Research Center
Cambridge, Massachusetts, October 1968
127-49-10-08-25

REFERENCES

1. Kalman, R. E., and Bucy, R. S.: J. Basic Eng. Trans. ASME, vol. 83D, 1961, pp. 95-108.

2. Kalman, R. E.: Rept. TR61-1. Research Inst. for Advanced Studies (RIAS), Baltimore, Md., November 1960.

3. Sorenson, H. W.: Kalman Filtering Techniques. Advances in Control Systems, Academic Press, 1966, pp. 219-292.

APPENDIX A

If M and S are square matrices, such that $\|SM^{-1}\| < 1$, then it is possible to verify that:

$$[M + S]^{-1} \cong M^{-1} - M^{-1}SM^{-1} + M^{-1}SM^{-1}SM^{-1} + \ldots \tag{1A}$$

Thus, in evaluating the matrix W_k of the test, it is to be noted that

$$W_k = \left[H_k^{-1}R_k\right]\left[H_kP_kH_k^T + R_k\right]^{-1}\left[H_k\right]$$

where

$$P_k \equiv P(k \mid k-1) \tag{2A}$$

If $\|R_k(H_kP_kH_k^T)^{-1}\| < 1$ using the approximation developed in Eq. (1A), W_k becomes

$$W_k \cong \left[H_k^{-1}R_k\right]\left[H_KP_KH_K^T\right]^{-1}\left[I - R_k(H\ P\ H_k^T)^{-1}\right]H_k \tag{3A}$$

where only the first two terms in Eq. (1A) have been included. With the use of matrix algebra, Eq. (3A) can be brought to the form:

$$W_k \cong \left[H_k^{-1}R_kH_k^{T-1}\right]P_k^{-1}\left[I - \left(H_k^{-1}R_kH_k^{T-1}\right)P_k^{-1}\right] \tag{4A}$$

Using the first term of the approximation of Eq. (4A) in Eqs. (3a) and (3b) of the text, the result can be:

$$P_{k+1} = \Phi(k+1,k)\left[H_k^{-1}R_kH_k^{T-1}\right]\Phi^T(k+1,k) + \Gamma_{k+1,k}Q_k\Gamma_{k+1,k}^T \tag{5A}$$

In view of the initial assumption, the first term in the r.h.s. of Eq. (5A) is small compared to the second. Thus, P_k can be approximately given by P_k^* which is defined as:

$$P_k^* \equiv H_k^{-1} R_k H_k^{T-1} + \Gamma_{k,k-1} Q_{k-1} \Gamma_{k,k-1}^T \qquad (6A)$$

Upon substituting P_A^* from Eq. (6A) in Eq. (4A), the result will be Eq. (10) of the text.

APPENDIX B

Assuming that M and S are positive definite symmetric square matrices, it follows from matrix theory that*:

$$M = U\Lambda_m U^T$$

and

$$S = T\Lambda_s T^T \tag{1B}$$

where U and T are unitary matrices and Λ_m and Λ_s are diagonal matrices. It also follows from matrix theory that*:

$$\mathrm{tr}(MS) = \mathrm{tr}(M^*\Lambda_s) \tag{2B}$$

where

$$M^* \equiv T^T M^T = (T^T U)\Lambda_m (U^T T)$$

It is to be noted that the product $(T^T U)$ is also a unitary matrix* and, thus, M^* is also a positive definite symmetric matrix. Eq. (2B) can be rewritten as:

$$\mathrm{tr}(MS) = \sum_i m_{ii}\sigma_i \tag{3B}$$

where m_{ii} are the diagonal elements of M^* and σ_i are the diagonal elements of Λ_s. Since $\sigma_i \geq 0$, it follows that:

$$(m)_M \sum_i \sigma_i \geq \mathrm{tr}(MS) \geq (m)_m \sum_i \sigma_i \tag{4B}$$

*See "Introduction to Matrix Analysis" by R. Bellman, pp. 38 and 95, McGraw-Hill Book Co., Inc., 1960.

where $(m)_M$ is the maximum of the m_{ii} elements and $(m)_m$ is the minimum. Moreover, it follows from Eq. (1B) that*:

$$m_{ii} = \sum_i \alpha_{ij}^2 \mu_j \tag{5B}$$

where

$$\sum_j \alpha_{ij}^2 = 1$$

The α_{ij} are the elements of $(T^T U)$ and μ_j are the diagonal elements of Λ_m. Hence, it is readily seen that:

$$(\mu)_M \geq m_{ii} \geq (\mu)_m$$

where $(\mu)_M$ is the maximum of the μ_j elements and $(\mu)_m$ is the minimum.

Using Eqs. (4B) and (6B) and identifying M with (I - B) and S with ΔP, it can easily be seen that Eq. (24) of the text will result.

---

*See "Introduction to Matrix Analysis" by R. Bellman, pp. 38 and 95, McGraw-Hill Book Co., Inc., 1960.

APPENDIX C

It is to be assumed here that systems to be considered are only those whose associated Kalman filters are completely controllable and completely observable. If this is true, it becomes apparent by considering the definition of observability* (at least in the case where the matrices of the Canonical equations are slowly varying) that the relations developed in the following section can be valid.

First, a new measurement vector $z^*(k)$ must be defined so that it has a dimensionality of n instead of the m dimensionality of $z(k)$ (where $m < n$).

Thus:

$$z^*(k) \equiv \begin{bmatrix} z(k) \\ \Phi(k,k-1)z(k-1) \\ \vdots \\ \Phi(k,k-s+1)\bar{z}(k-s+1) \end{bmatrix} \tag{1C}$$

The integer, s, is chosen in Eq. (1C) so that $n/m \leq s < n/m + 1$. $\bar{z}(k-s+1)$ is a vector using only the first p components of $z(k-s+1)$ where $p = n - (s-1)m$. It would follow that it is possible to construct a new $n \times m$ matrix H_k^* which has an inverse and is defined by:

$$z^*(k) = H_k^* x(k) + v^*(k) \tag{2C}$$

where

$$v^*(k) \equiv \begin{bmatrix} v(k) \\ \Phi(k,k-1)v(k-1) \\ \vdots \\ \Phi(k,k-s+1)\bar{v}(k-s+1) \end{bmatrix}$$

---

*See "Optimal Estimation Identification and Control" by Robert C.K. Lee, pp. 82-83, M.I.T. Press, Cambridge, Mass., 1964.

In terms of the n×m, matrix H_k, the new matrix H_k^* would have the explicit form:

$$H_k^* = \begin{bmatrix} H_k \\ \Phi(k,k-1)H_{k-1} \\ \vdots \\ \Phi(k,k-s+1)\bar{H}_{k-s+1} \end{bmatrix} \tag{3C}$$

where $\bar{H}_{k-s+1}$ is the reduced p×n matrix formed by using the first p rows of H_{k-s+1}.

By using $z^*(k-i)$, H_{k-i}^* and R_{k-i}^* instead of $z(k-i$, H_{k-i} and $R_{k-i})$ in the estimator systems, it is seen that the treatment developed in the text can be extended to the general case. However, it is apparent that by using this scheme estimates cannot be made at the end of every time t_{k-i} (where $i = 0,1,2,\ldots$), but the intervals between estimates would be $s\Delta t$ apart (i.e., estimates would occur at the times t_{k-i} where $i = 0,s,2s,3s,\ldots$). Moreover, since some of the $z(k-i)$ data is thrown away [to form $\bar{z}(k-s+1)$] the accuracy of the estimate would be reduced.

NASA-Langley, 1969 —— 19 C-55