

# Numerical Aspects of Different Kalman Filter Implementations

MICHEL VERHAEGEN AND PAUL VAN DOOREN, MEMBER, IEEE

**Abstract**—A theoretical analysis is made of the error propagation due to numerical roundoff for four different Kalman filter implementations: the conventional Kalman filter, the square root covariance filter, the square root information filter, and the Chandrasekhar square root filter. An experimental analysis is performed to validate the new insights gained by the theoretical analysis.

## I. INTRODUCTION

SINCE the appearance of Kalman's 1960 paper [1], the so-called Kalman filter (KF) has been applied successfully to many practical problems, especially in aeronautical and aerospace applications. As applications became more numerous, some pitfalls of the KF were discovered such as the problem of divergence due to the lack of reliability of the numerical algorithm or to inaccurate modeling of the system under consideration [2].

Therefore, several modified implementations of the KF were presented in an effort to avoid these numerical problems. Many of these modifications were based on heuristics (as in the stabilized KF [3], or the conventional KF with lower bounding [4]) which often require much experience in order to implement them effectively. Later, more reliable KF implementations were described such as the square root filter (SRF) proposed by Potter in 1963 [5]. For this filter the reliability of the filter estimates is expected to be better because of the use of numerically stable orthogonal transformations for each recursion step. On the other hand, the SRF implementation required more computations than the conventional KF [6]. This problem of cost efficiency gave rise to the development of modified versions of the SRF such as the UDU-algorithms [7] and the Chandrasekhar form [9]. These implementations can be made as efficient as the conventional KF, or for the Chandrasekhar SRF even more efficient for some special experimental conditions.

In this paper we reconsider the numerical robustness of existing KF's and derive some results giving new and/or better insights into their numerical performance. Here we investigate four "basic" KF implementations: the conventional Kalman filter (CKF), the square root covariance filter (SRCF), the Chandrasekhar square root filter (CSRf), and the square root information filter (SRIF). (The implementations chosen come from [12]; these differ substantially from the forms described in [7] with the same names!) This certainly does not cover all possible implementations encountered in practice, but insights gained for these general cases are very useful in judging variants such as the efficient KF algorithms based on the sequential processing technique [7] or the "condensed form" versions [10], [11]. After a brief description of the above filters in Section II, we perform in Section III a detailed first-order perturbation study of the error propagation due

to roundoff for the above four KF implementations. In Section IV a realistic simulation study is performed in order to validate the results of the theoretical analysis. Section V then outlines a comparison between the different filter implementations using the results of the theoretical error analysis and the simulation study. We end with some concluding remarks in Section VI.

## II. NOTATION AND PRELIMINARIES

In this section we introduce our notation and list the different Kalman filter types that are discussed in the paper. We consider the discrete time-varying linear system

$$x_{k+1} = A_k x_k + B_k w_k + D_k u_k \quad (1)$$

and the linear observation process

$$y_k = C_k x_k + v_k \quad (2)$$

where  $x_k$ ,  $u_k$ , and  $y_k$  are, respectively, the state vector to be estimated ( $\epsilon R^n$ ), the deterministic input vector ( $\epsilon R^r$ ), and the measurement vector ( $\epsilon R^p$ ), where  $w_k$  and  $v_k$  are the process noise ( $\epsilon R^m$ ) and the measurement noise ( $\epsilon R^p$ ) of the system, and, finally, where  $A_k$ ,  $B_k$ ,  $C_k$ , and  $D_k$  are known matrices of appropriate dimensions. The process noise and measurement noise sequences are assumed zero mean and uncorrelated

$$E\{w_k\} = 0, \quad E\{v_k\} = 0, \quad E\{w_k v_j'\} = 0 \quad (3)$$

with known covariances

$$E\{w_j w_k'\} = Q_k \delta_{jk}, \quad E\{v_j v_k'\} = R_k \delta_{jk} \quad (4)$$

where  $E\{\cdot\}$  denotes the mathematical expectation and  $Q_k$  and  $R_k$  are positive definite matrices.

The assumption that  $Q_k$  is nonsingular does not restrict the generality of the system description, since for the case of singular  $Q_k$ , the linearly dependent components in  $w_k$  can always be removed first [12]. On the other hand, the regularity of  $R_k$  rules out the possibility of including perfect measurements not corrupted by noise. In the particular case of perfect measurements, special adaptations are required for some of the KF implementations, such as the use of the Moore-Penrose inverse for the CKF [12]. Such special implementations are not considered here except for a few comments in the concluding remarks.

The SRF algorithm uses the Choleski factors of the covariance matrices or their inverse in order to solve the optimal filtering problem. Since the process noise covariance matrix  $Q_k$  and the measurement noise covariance matrix  $R_k$  are assumed to be positive definite, the following Choleski factorizations<sup>2</sup> exist:

$$Q_k = Q_k^{1/2} [Q_k^{1/2}]', \quad R_k = R_k^{1/2} [R_k^{1/2}]' \quad (5)$$

<sup>1</sup> with nonsingularity required for the SRIF

<sup>2</sup> Notice that historically  $Q_k^{1/2}$  and  $R_k^{1/2}$  have erroneously been called "square roots" instead of "Choleski factors." However, we will maintain the adjective "square root" as far as the names of the filters are concerned because of the familiarity that they have acquired.

where the factors  $Q_k^{1/2}$  and  $R_k^{1/2}$  may be chosen *upper* or *lower* triangular. This freedom of choice is exploited in the development of the fast KF implementations presented in Section II. The problem is now to compute the *minimum variance estimate* of the stochastic variable  $x_k$ , provided  $y_1$  up to  $y_j$  have been measured

$$\hat{x}_{k|j} = \hat{x}_{k|y_1, \dots, y_j} \quad (6)$$

When  $j = k$  this estimate is called the *filtered estimate* and for  $j = k - 1$  it is referred to as the one-step predicted or, shortly, the *predicted estimate*. The above problem is restricted here to these two types of estimates except for a few comments in the concluding remarks. Kalman filtering is a recursive method to solve this problem. This is done by computing the variances  $P_{k|k}$  and/or  $P_{k|k-1}$  and the estimates  $\hat{x}_{k|k}$  and/or  $\hat{x}_{k|k-1}$  from their previous values, this for  $k = 1, 2, \dots$ . Thereby one assumes  $P_{0|-1}$  (i.e., the covariance matrix of the initial state  $x_0$ ) and  $\hat{x}_{0|-1}$  (i.e., the mean of the initial state  $x_0$ ) to be *given*.

### A. The Conventional Kalman Filter (CKF)

The above recursive solution can be computed by the CKF equations, summarized in the following "covariance form" [12]:

$$R_k^e = R_k + C_k P_{k|k-1} C_k' \quad (7)$$

$$K_k = A_k P_{k|k-1} C_k' [R_k^e]^{-1} \quad (8)$$

$$P_{k+1|k} = A_k [I - P_{k|k-1} C_k' [R_k^e]^{-1} C_k] P_{k|k-1} A_k' + B_k Q_k B_k' \quad (9)$$

$$\hat{x}_{k+1|k} = A_k \hat{x}_{k|k-1} - K_k [C_k \hat{x}_{k|k-1} - y_k] + D_k u_k \quad (10)$$

This set of equations has been implemented in various forms; see [12]. An efficient implementation that exploits the symmetry of the different matrices in (7)–(10) requires per step  $3n^3/2 + n^2(3p + m/2) + n(3p^2/2 + m^2) + p^3/6$  "flops" (where 1 flop = 1 multiplication + 1 addition). By not exploiting the symmetry of the matrices in (7)–(10) one requires  $(n^3/2 + n^2m/2 + np^2/2)$  more flops. In the error analysis, it is this "costly" implementation that is initially denoted as the CKF for reasons that are explained there. In Section II-E we also give some other variants that lead to further improvements in the number of operations.

### B. The Square Root Covariance Filter (SRCF)

Square root covariance filters propagate the Choleski factors of the error covariance matrix  $P_{k|k-1}$

$$P_{k|k-1} = S_k \cdot S_k' \quad (11)$$

where  $S_k$  is chosen to be lower triangular. The computational method is summarized by the following scheme [12]:

$$\underbrace{\begin{pmatrix} R_k^{1/2} & C_k S_k & 0 \\ 0 & A_k S_k & B_k Q_k^{1/2} \end{pmatrix}}_{\text{(prearray)}} \cdot U_1 = \underbrace{\begin{pmatrix} R_k^{e/2} & 0 & 0 \\ G_k & S_{k+1} & 0 \end{pmatrix}}_{\text{(postarray)}} \quad (12)$$

$$\hat{x}_{k+1|k} = A_k \hat{x}_{k|k-1} - G_k R_k^{e-1/2} (C_k \hat{x}_{k|k-1} - y_k) + D_k u_k \quad (13)$$

where  $U_1$  is an orthogonal transformation that triangularizes the prearray. Such a triangularization can, e.g., be obtained using Householder transformations [13]. This recursion is now initiated with  $\hat{x}_{0|-1}$  and the Choleski factor  $S_0$  of  $P_{0|-1}$  as defined in (11). The number of flops needed for (12) and (13) is  $7n^3/6 + n^2(5p/2 + m) + n(p^2 + m^2/2)$ . In order to reduce the amount of work, we only compute here the diagonal elements of the covariance

matrix  $P_{k+1|k}$ , since usually  $\text{diag}\{P_{k+1|k}\}$  carries enough information about the estimate  $\hat{x}_{k+1|k}$  (namely the variance of the individual components). For this reason our operation counts differ, e.g., from those of [6]. In Section II-E we shortly discuss some other variants that lead to further improvements in the number of operations.

### C. The Chandrasekhar Square Root Filter (CSRF)

If the system model (1), (2) is *time-invariant*, the SRCF described in Section II-B may be simplified to the Chandrasekhar square root filter, described in [14], [9]. Here one formulates recursions for the *increment* of the covariance matrix, defined as

$$\text{inc } P_k = P_{k+1|k} - P_{k|k-1} \quad (14)$$

In general, this matrix can be factored as

$$\text{inc } P_k = L_k \cdot \underbrace{\begin{pmatrix} I_{n_1} & 0 \\ 0 & -I_{n_2} \end{pmatrix}}_{\Sigma} \cdot L_k' \quad (15)$$

where the rank of  $\text{inc } P_k$  is  $n_1 + n_2$  and  $\Sigma$  is called its signature matrix. The CSRF propagates recursions for  $L_k$  and  $\hat{x}_{k+1|k}$  using [14]:

$$\underbrace{\begin{pmatrix} R_{k-1}^{e/2} & CL_{k-1} \\ G_{k-1} & AL_{k-1} \end{pmatrix}}_{\text{(prearray)}} \cdot U_2 = \underbrace{\begin{pmatrix} R_k^{e/2} & 0 \\ G_k & L_k \end{pmatrix}}_{\text{(postarray)}}, \quad (16)$$

$$\hat{x}_{k+1|k} = A_k \hat{x}_{k|k-1} - G_k R_k^{e-1/2} (C_k \hat{x}_{k|k-1} - y_k) + D_k u_k \quad (17)$$

with  $L_0 \Sigma L_0' = P_{1|0} - P_{0|-1}$ . Here  $U_2$  is a  $\Sigma_p$ -unitary transformation, i.e.,  $U_2 \Sigma_p U_2' = \Sigma_p$  with

$$\Sigma_p = \begin{pmatrix} I_p & 0 \\ 0 & \Sigma \end{pmatrix} \quad (18)$$

Such transformations are easily constructed using "skew Householder" transformations (using an indefinite  $\Sigma_p$ -norm) and require as many operations as the classical Householder transformations [14]. (Later, it is noted that numerically they are not always well behaved.) For this implementation the operation count is  $(n_1 + n_2)(n^2 + 3np + p^2)$  flops.

### D. The Square Root Information Filter (SRIF)

The information filter accentuates the recursive least-squares nature of filtering [7], [12]. The SRIF propagates the Choleski factor of  $P_{k|k}^{-1}$  using the Choleski factor of the inverse of the process- and measurement-noise covariance matrices

$$P_{k|k}^{-1} = T_k' \cdot T_k \quad (19)$$

$$Q_k^{-1} = [Q_k^{-1/2}]' \cdot Q_k^{-1/2} \quad (20)$$

$$R_k^{-1} = [R_k^{-1/2}]' \cdot R_k^{-1/2} \quad (21)$$

where the right factors are all chosen upper triangular. We now present the Dyer and McReynolds formulation of the SRIF (except for the fact that the time and measurement updates are combined here as in [12]) which differs from the one presented by Bierman (see [7] for details). One recursion of the SRIF algorithm

is given by [12]:

$$\begin{aligned}
 U_3 \cdot \underbrace{\begin{pmatrix} Q_k^{-1/2} & 0 & 0 \\ T_k A_k^{-1} B_k & T_k A_k^{-1} & T_k \hat{x}_{k|k} \\ 0 & R_{k+1}^{-1/2} C_{k+1} & R_{k+1}^{-1/2} y_{k+1} \end{pmatrix}}_{\text{(prearray)}} \\
 = \underbrace{\begin{pmatrix} Q_{k+1}^{e-1/2} & * & * \\ 0 & T_{k+1} & \hat{x}_{k+1|k+1} \\ 0 & 0 & r_{k+1} \end{pmatrix}}_{\text{(postarray)}} \quad (22)
 \end{aligned}$$

and the filtered state estimate is computed by

$$\hat{x}_{k+1|k+1} = T_{k+1}^{-1}(\hat{x}_{k+1|k+1}) + D_k u_k. \quad (23)$$

An operation count of this filter is  $7n^3/6 + n^2(p + 7m/2) + n(p^2/2 + m^2)$  flops. Here we did not count the operations needed for the inversion and/or factorization of  $Q_k$ ,  $R_k$ , and  $A_k$  (for the time-invariant case, e.g., these are computed only once) and again (as for the SRCF) only the diagonal elements of the information matrix  $P_{k,k}^{-1}$  are computed at each step.

### E. Efficient Implementations

Variants of the above basic KF implementations have been developed which mainly exploit some particular structure of the given problem in order to reduce the amount of computations; e.g., when the measurement noise covariance matrix  $R_k$  is *diagonal*, it is possible to perform the *measurement update* in  $p$  scalar updates. This is the so-called *sequential processing* technique, a feature that is exploited by the *UDU'*-algorithm to operate for the multivariable output case. A similar processing technique for the *time update* can be formulated when the process noise covariance matrix  $Q_k$  is *diagonal*, which is then exploited in the SRIF algorithm. Notice that no such technique can be used for the CSRf. The *UDU'*-algorithm also saves operations by using unit triangular factors  $U$  and a diagonal matrix  $D$  in the updating formulas for which then special versions can be obtained [7]. By using *modified Givens rotations* [15] one could also obtain similar savings for the updating of the usual Choleski factors, but these variants are not reported in the sequel.

For the *time-invariant* case, the matrix multiplications and transformations that characterize the described KF implementations can be made more efficient when the system matrices  $\{A, B, C\}$  are first transformed by *unitary similarity transformations* to so-called *condensed form*, whereby these system matrices  $\{A_i, B_i, C_i\}$  contain many zeros. From the point of view of *reliability*, these forms are particularly interesting here because no loss of accuracy is incurred by these unitary similarity transformations [10]. The following types of condensed forms can be used to obtain considerable savings in computation time in the subsequent filter recursions [10]: the *Shur form*, where  $A_i$  is in upper or lower Schur form, the *observer-Hessenberg form*, where the compound matrix  $(A_i', C_i')$  is upper trapezoidal, and the *controller-Hessenberg form*, where the compound matrix  $(A_i, B_i)$  is upper trapezoidal. In [10], an application is considered where these efficient implementations are also valid for the *time-varying* case. Note that the use of condensed forms and "sequential processing" could very well be combined to yield even faster implementations.

The operation counts for particular mechanizations of these variants are all given in Table I and indicated by, respectively, the

TABLE I  
OPERATION COUNTS FOR THE DIFFERENT KF'S

| filter | type    | complexity   |
|--------|---------|--|
| CKF    | full    | $(3/2)n^3 + n^2(3p + m/2) + n(3p^2/2 + m^2) + p^3/6$   |
|        | seq.    | $(3/2)n^3 + n^2(3p + m/2) + n(p^2 + m^2)$              |
|        | Schur   | $(3/4)n^3 + n^2(5p/2 + m/2) + n(3p^2/2 + m^2) + p^3/6$ |
|        | o-Hess. | $(3/4)n^3 + n^2(7p/2 + m/2) + n(2p^2 + m^2) + p^3/6$   |
| SRCF   | full    | $(7/6)n^3 + n^2(5p/2 + m) + n(p^2 + m^2/2)$            |
|        | seq.    | $(7/6)n^3 + n^2(5p/2 + m) + n(m^2/2)$                  |
|        | Schur   | $(1/6)n^3 + n^2(5p/2 + m) + n(2p^2)$                   |
|        | o-Hess. | $(1/6)n^3 + n^2(3p/2 + m) + n(2p^2) + 2p^3/3$          |
| SRIF   | full    | $(7/6)n^3 + n^2(p + 7m/2) + n(p^2/2 + m^2)$            |
|        | seq.    | $(7/6)n^3 + n^2(p + 7m/2) + n(p^2/2)$                  |
|        | Schur   | $(1/6)n^3 + n^2(p + 5m/2) + n(2m^2)$                   |
|        | c-Hess. | $(1/6)n^3 + n^2(3m/2 + p) + n(m^2 + p^2/2)$            |
| CSRf   | full    | $(n_1 + n_2)(n^2 + 3np + p^2)$                         |
|        | Schur   | $(n_1 + n_2)(n^2/2 + 3np + p^2)$                       |
|        | o-Hess. | $(n_1 + n_2)(n^2/2 + 3np + p^2)$                       |

"seq.," "Schur," "o-Hess" and "c-Hess" abbreviations, while "full" refers to the implementations described in previous sections.<sup>3</sup>

### III. ERROR ANALYSIS

In this section we analyze the effect of rounding errors on Kalman filtering in the four different implementations described above. The analysis is split in three parts: 1) what bounds can be obtained for the errors performed in step  $k$ ; 2) how do errors performed in step  $k$  propagate in subsequent steps; and 3) how do errors performed in different steps interact and accumulate. Although this appears to be the logical order in which one should treat the problem of error buildup in KF, we first look at the second aspect, which is also the only one that has been studied in the literature so far. Therefore, we first need the following lemma which is easily proved by inspection.

*Lemma 1:* Let  $A$  be a square nonsingular matrix with smallest singular value  $\sigma_{\min}$  and let  $E$  be a perturbation of the order of  $\delta = \|E\|_2 \ll \sigma_{\min}(A)$  with  $\|\cdot\|_2$  denoting the 2-norm. Then

$$(A + E)^{-1} = A^{-1} + \Delta_1 = A^{-1} - A^{-1}EA^{-1} + \Delta_2 \quad (24)$$

where

$$\|\Delta_1\|_2 \leq \delta / \sigma_{\min}(\sigma_{\min} - \delta) = O(\delta) \quad (25)$$

$$\|\Delta_2\|_2 \leq \delta^2 / \sigma_{\min}^2(\sigma_{\min} - \delta) = O(\delta^2). \quad (26)$$

Notice that when  $A$  and  $E$  are symmetric, these first- and second-order approximations (25) and (26) are also symmetric.

We now thus consider the propagation of errors from step  $k$  to step  $k + 1$  when no additional errors are performed during that update. We denote the quantities in computer with an upperbar, i.e.,  $\bar{P}_{k|k-1}$ ,  $\bar{x}_{k|k-1}$ ,  $\bar{G}_k$ ,  $\bar{S}_k$ ,  $\bar{T}_k$ ,  $\bar{R}_k^{e1/2}$ ,  $\bar{F}_k$ ,  $\bar{K}_k$ , or  $\bar{L}_k$ , depending on the algorithm.

For the CKF, let  $\delta P_{k|k-1}$  and  $\delta x_{k|k-1}$  be the accumulated errors in step  $k$ , then:

$$\bar{P}_{k|k-1} = P_{k|k-1} + \delta P_{k|k-1}, \quad \bar{x}_{k|k-1} = \hat{x}_{k|k-1} + \delta \hat{x}_{k|k-1}. \quad (27)$$

By using Lemma 1 for the inverse of  $\bar{R}_k^e = R_k^e + C_k \delta P_{k|k-1} C_k'$ , we find

$$[\bar{R}_k^e]^{-1} = [R_k^e]^{-1} - [R_k^e]^{-1} C_k \delta P_{k|k-1} C_k' [R_k^e]^{-1} + O(\delta^2). \quad (28)$$

<sup>3</sup> where the full implementation of the CKF exploits symmetry

From this, one then derives

$$\begin{aligned}\bar{K}_k &= A_k \bar{P}_{k|k-1} C_k' \bar{R}_k^{e-1} \\ \delta K_k &= A_k \delta P_{k|k-1} C_k' R_k^{e-1} \\ &\quad - A_k P_{k|k-1} C_k' R_k^{e-1} C_k \delta P_{k|k-1} C_k' R_k^{e-1} + O(\delta^2) \\ &= F_k \delta P_{k|k-1} C_k' R_k^{e-1} + O(\delta^2)\end{aligned}\quad (29)$$

where

$$F_k = A_k(I - P_{k|k-1} C_k' R_k^{e-1} C_k) = A_k - K_k C_k \quad (30)$$

and (assuming  $\bar{P}_{k|k-1}$  is not necessarily symmetric, which would, e.g., occur when applying (9) bluntly)

$$\begin{aligned}\bar{P}_{k+1|k} &= A_k(\bar{P}_{k|k-1} - \bar{P}'_{k|k-1} C_k' \bar{R}_k^{e-1} C_k \bar{P}_{k|k-1}) A_k' + B_k Q_k B_k' \\ \delta P_{k+1|k} &= A_k(\delta P_{k|k-1} - \delta P'_{k|k-1} C_k' R_k^{e-1} C_k P_{k|k-1} \\ &\quad - P_{k|k-1} C_k' R_k^{e-1} C_k \delta P_{k|k-1} \\ &\quad + P_{k|k-1} C_k' R_k^{e-1} C_k \delta P_{k|k-1} C_k' R_k^{e-1} C_k P_{k|k-1}) A_k' + O(\delta^2) \\ &= (A_k - K_k C_k) \delta P_{k|k-1} (A_k' - C_k' K_k') \\ &\quad + A_k(\delta P_{k|k-1} - \delta P'_{k|k-1}) C_k' K_k' + O(\delta^2) \\ &= F_k \delta P_{k|k-1} F_k' + A_k(\delta P_{k|k-1} - \delta P'_{k|k-1}) A_k' \\ &\quad - A_k(\delta P_{k|k-1} - \delta P'_{k|k-1}) F_k' + O(\delta^2).\end{aligned}\quad (31)$$

For the estimate  $\hat{x}_{k+1|k}$ , we have:

$$\begin{aligned}\bar{\hat{x}}_{k+1|k} &= \bar{F}_k \bar{\hat{x}}_{k|k-1} + \bar{K}_k y_k + D_k u_k \\ \delta \hat{x}_{k+1|k} &= F_k \delta \hat{x}_{k|k-1} + \delta F_k \hat{x}_{k|k-1} + \delta K_k y_k + O(\delta^2) \\ &= F_k \delta \hat{x}_{k|k-1} + \delta K_k (y_k - C_k \hat{x}_{k|k-1}) + O(\delta^2) \\ &= F_k [\delta \hat{x}_{k|k-1} + \delta P_{k|k-1} C_k' R_k^{e-1} (y_k - C_k \hat{x}_{k|k-1})] + O(\delta^2).\end{aligned}\quad (32)$$

When on the other hand,  $\delta P_{k|k}$  and  $\hat{x}_{k|k}$  are given, one derives analogously,

$$\begin{aligned}\delta P_{k+1|k+1} &= \bar{F}_k \delta P_{k|k} \bar{F}_k' + A_k(\delta P_{k|k} - \delta P'_{k|k}) A_k' \\ &\quad - A_k(\delta P_{k|k} - \delta P'_{k|k}) \bar{F}_k' + O(\delta^2) \quad (33) \\ \delta \hat{x}_{k+1|k+1} &= \bar{F}_k [\delta \hat{x}_{k|k} + \delta P_{k|k} A_k' C_{k+1}' R_{k+1}^{e-1} \\ &\quad \cdot (y_{k+1} - C_{k+1} A_k \hat{x}_{k|k})] + O(\delta^2) \quad (34)\end{aligned}$$

where  $\bar{F}_k = (I - P_{k+1|k} C_{k+1}' R_{k+1}^{e-1} C_{k+1}) A_k$  has the same spectrum as  $F_{k+1}$  in the time-invariant case, since  $F_{k+1} A_k = A_{k+1} \bar{F}_k$  [16].

We thus find that when  $\delta P_{k|k-1}$  or  $\delta P_{k|k}$  is symmetric, only the first term in (31) or (33) remains and the error propagation behaves roughly as

$$\|\delta P_{k+1|k}\|_2 \approx \|F_k\|_2^2 \cdot \|\delta P_{k|k-1}\|_2 = \gamma_k^2 \cdot \|\delta P_{k|k-1}\|_2 \quad (35)$$

$$\|\delta P_{k+1|k+1}\|_2 \approx \|\bar{F}_k\|_2^2 \cdot \|\delta P_{k|k}\|_2 = \bar{\gamma}_k^2 \cdot \|\delta P_{k|k}\|_2 \quad (36)$$

which are decreasing in time when  $F_k$  and  $\bar{F}_k$  are contractions (i.e., when  $\gamma_k$  and  $\bar{\gamma}_k < 1$ ). The latter is usually the case when the matrices  $A_k$ ,  $B_k$ ,  $C_k$ ,  $Q_k$ , and  $R_k$  do not vary too wildly in time [12]. For the time-invariant case one can improve on this by saying that  $F_k$  and  $\bar{F}_k$  tend to the constant matrices  $F_\infty$  and  $\bar{F}_\infty$ , respectively, with (equal) spectral radius  $\rho_\infty < 1$  and one then has for some appropriate matrix norm [17]:

$$\|\delta P_{k+1|k}\| \approx \rho_\infty^2 \cdot \|\delta P_{k|k-1}\| \quad (37)$$

$$\|\delta P_{k+1|k+1}\| \approx \rho_\infty^2 \cdot \|\delta P_{k|k}\| \quad (38)$$

for sufficiently large  $k$ . Notice that  $\rho_\infty$  is smaller than  $\gamma_\infty$  or  $\bar{\gamma}_\infty$ , hence, (37), (38) are better bounds than (35), (36). Using this, it then also follows from (37), (38) that all three errors  $\delta P_{k|k-1}$ ,  $\delta K_k$ , and  $\delta \hat{x}_{k|k-1}$  are decreasing in time when no additional errors are performed. The fact that past errors are weighted in such a manner is the main reason why many Kalman filters do not diverge in presence of rounding errors.

The property (35)–(38) was already observed before [2], but for symmetric  $\delta P_{k|k-1}$ . However, if symmetry is removed, divergence may occur when  $A_k$  (i.e., the original plant) is unstable. Indeed, from (31), (33) we see that when  $A_k$  is unstable the larger part of the error is skew symmetric:

$$\delta P_{k+1|k} \approx A_k \cdot (\delta P_{k|k-1} - \delta P'_{k|k-1}) \cdot A_k' \quad (39)$$

$$\delta P_{k+1|k+1} \approx A_k \cdot (\delta P_{k|k} - \delta P'_{k|k}) \cdot A_k' \quad (40)$$

and the lack of symmetry *diverges* as  $k$  increases. This phenomenon is well known in the extensive literature about Kalman filtering and experimental experience has led to a number of different "remedies" to overcome it. The above first-order perturbation analysis in fact explains why they work.

1) A first method to avoid divergence due to the loss of symmetry when  $A_k$  is unstable, is to *symmetrize*  $\bar{P}_{k|k-1}$  or  $\bar{P}_{k|k}$  at each recursion of the CKF by averaging it with its transpose. This makes the errors on  $P$  symmetric, and hence the largest terms in (31), (33) disappear!

2) A second method to make the errors on  $P$  symmetric, simply computes only the *upper* (or lower) *triangular* part of these matrices, such as indicated by the implementation in Table I.

3) A third technique to avoid the loss of symmetry is the so-called (Joseph's) stabilized KF [3]. In this implementation, the set of equations for updating  $P$  are rearranged as follows:

$$P_{k+1|k} = F_k P_{k|k-1} F_k' + K_k R_k K_k' + B_k Q_k B_k'. \quad (41)$$

A similar first-order perturbation study as for the CKF above, learns that *no symmetrization* is required in order to avoid divergence since here the error propagation model becomes:

$$\delta P_{k+1|k} = F_k \delta P_{k|k-1} F_k' + O(\delta^2) \quad (42)$$

where there are no terms anymore related to the loss of symmetry.

Since for the moment we assume that *no additional errors* are performed in the recursions, one *inherently* computes the same equations for the SRKF as for the CKF. Therefore, starting with errors  $\delta S_k$  and  $\delta \hat{x}_{k|k-1}$  (29), (31), (32), (35), and (37) still hold, whereby now

$$\delta P_{k|k-1} = S_k \cdot \delta S_k' + \delta S_k \cdot S_k' + \delta S_k \cdot \delta S_k' \quad (43)$$

is clearly symmetric by construction. According to (31) this now ensures the convergence to zero of  $\delta P_{k|k-1}$ , and hence of  $\delta S_k$ ,  $\delta K_k$  and  $\delta \hat{x}_{k|k-1}$  if  $\gamma_k$  is sufficiently bounded in the time-varying case.

For the SRIF we start with errors  $\delta T_k$  and  $\delta \hat{x}_{k|k}$  and use the identity

$$\delta P_{k|k}^{-1} = T_k' \cdot \delta T_k + \delta T_k' \cdot T_k + \delta T_k' \cdot \delta T_k \quad (44)$$

$$\delta x_{k|k} = (T_k + \delta T_k)^{-1} \delta \xi_{k|k} \quad (45)$$

to relate this problem to the CKF as well. Here one apparently does *not* compute  $\hat{x}_{k+1|k+1}$  from  $\hat{x}_{k|k}$  and therefore one would expect *no propagation* of errors between them. Yet, such a propagation is *present* via the relation (45) with the errors on  $\delta \xi_{k+1|k+1}$  and  $\delta \xi_{k|k}$ , which *do* propagate from one step to another. This in fact is reflected in the recurrence (34) derived earlier. Since the SRIF update is *inherently equivalent* to an update of  $P_{k|k}$  and  $\hat{x}_{k|k}$  as in the CKF, the equations (33), (36) still hold where now the symmetry of  $\delta P_{k|k}$  is ensured because of (44). From this it follows that  $\delta P_{k|k}$  and  $\delta \hat{x}_{k|k}$ , and therefore also  $\delta T_k$

and  $\delta\hat{\epsilon}_{k|k}$ , converge to zero as  $k$  increases, provided  $\tilde{\gamma}_k$  is sufficiently bounded in the time-varying case.

Finally, for the CSRFB we start with errors  $\delta L_{k-1}$ ,  $\delta G_{k-1}$ ,  $\delta R_{k-1}$ , and  $\delta\hat{x}_{k|k-1}$ . Because of these errors, (16) is perturbed exactly as follows:

$$\begin{pmatrix} R_{k-1}^{e1/2} + \delta R_{k-1}^{e1/2} & C(L_{k-1} + \delta L_{k-1}) \\ G_{k-1} + \delta G_{k-1} & A(L_{k-1} + \delta L_{k-1}) \end{pmatrix} \cdot \bar{U}_2 \\ = \begin{pmatrix} R_k^{e1/2} + \delta R_k^{e1/2} & 0 \\ G_k + \delta G_k & L_k + \delta L_k \end{pmatrix} \quad (46)$$

where  $\bar{U}_2$  is also  $\Sigma_p$ -unitary. When  $\lambda = \|C \cdot L_{k-1}\| \ll \|R_{k-1}^{e1/2}\|$  (which is satisfied when  $k$  is sufficiently large), Lemma A.3 yields after some manipulations

$$\begin{pmatrix} \delta R_{k-1}^{e1/2} & C\delta L_{k-1} \\ \delta G_{k-1} & A\delta L_{k-1} \end{pmatrix} \cdot U_2 = \begin{pmatrix} \delta R_k^{e1/2} & 0 \\ \delta G_k & \delta L_k \end{pmatrix} + O(\delta \cdot \lambda). \quad (47)$$

Now the (1, 1) and (1, 2) blocks of  $U_2'$  are easily checked to be given by  $R_k^{e-1/2} \cdot R_{k-1}^{e1/2}$  and  $R_k^{e-1/2} \cdot C \cdot L_{k-1} \cdot \Sigma$ , respectively. From this, one then derives that for  $k$  sufficiently large

$$\begin{aligned} \delta R_k^{e1/2} &= \delta R_{k-1}^{e1/2} \cdot [R_k^{e-1/2} \cdot R_{k-1}^{e1/2}]' + C \cdot \delta L_{k-1} \\ &\quad \cdot [R_k^{e-1/2} \cdot C \cdot L_{k-1} \cdot \Sigma]' + O(\delta \cdot \lambda) \\ &= \delta R_{k-1}^{e1/2} \cdot [R_k^{e-1/2} \cdot R_{k-1}^{e1/2}]' + O(\delta \cdot \lambda) \end{aligned} \quad (48)$$

$$\begin{aligned} \delta G_k &= \delta G_{k-1} \cdot [R_k^{e-1/2} \cdot R_{k-1}^{e1/2}]' + A \cdot \delta L_{k-1} \\ &\quad \cdot [R_k^{e-1/2} \cdot C \cdot L_{k-1} \cdot \Sigma]' + O(\delta \cdot \lambda) \\ &= \delta G_{k-1} \cdot [R_k^{e-1/2} \cdot R_{k-1}^{e1/2}]' + O(\delta \cdot \lambda). \end{aligned} \quad (49)$$

Here again thus the errors  $\delta R_{k-1}^{e1/2}$  and  $\delta G_{k-1}$  are multiplied by the matrix  $[R_k^{e-1/2} \cdot R_{k-1}^{e1/2}]'$  at each step. When  $\Sigma$  is the identity matrix (i.e., when  $\text{inc } P_k$  is nonnegative) this is a contraction since  $R_k^e = R_{k-1}^e + C \cdot L_{k-1} \cdot L_{k-1}' \cdot C'$ . From this, we then derive similar formulas for the propagation of  $\delta K_k$  and  $\delta\hat{x}_{k+1|k}$ . Using Lemma 1 for the perturbation of the inverse in  $K_k = G_k \cdot R_k^{e-1/2}$ , we find

$$\begin{aligned} \delta K_k &= \delta G_k \cdot R_k^{e-1/2} - G_k \cdot R_k^{e-1/2} \cdot \delta R_k^{e1/2} \cdot R_k^{e-1/2} + O(\delta^2) \\ &= \delta G_k \cdot R_k^{e-1/2} - K_k \cdot \delta R_k^{e1/2} \cdot R_k^{e-1/2} + O(\delta^2). \end{aligned} \quad (50)$$

Using (49), (50) and the fact that for large  $k$ ,  $K_k = K_{k-1} + O(\lambda)$ , we then obtain

$$\begin{aligned} \delta K_k &= \delta G_{k-1} \cdot [R_k^{e-1/2} \cdot R_{k-1}^{e1/2}]' \cdot R_k^{e-1/2} \\ &\quad - K_{k-1} \cdot \delta R_{k-1}^{e1/2} \cdot [R_k^{e-1/2} \cdot R_{k-1}^{e1/2}]' \cdot R_k^{e-1/2} + O(\delta \cdot \lambda) \\ &= \delta G_{k-1} \cdot R_{k-1}^{e-1/2} \cdot [R_{k-1}^e \cdot R_{k-1}^{e-1}] \\ &\quad - K_{k-1} \cdot \delta R_{k-1}^{e1/2} \cdot R_{k-1}^{e-1/2} \cdot [R_{k-1}^e \cdot R_{k-1}^{e-1}] + O(\delta \cdot \lambda) \end{aligned} \quad (51)$$

which because of (50) decremented by 1 becomes

$$\delta K_k = \delta K_{k-1} \cdot [R_{k-1}^e \cdot R_{k-1}^{e-1}] + O(\delta \cdot \lambda). \quad (52)$$

Using (17) we then also obtain from this

$$\delta\hat{x}_{k+1|k} = F_k \cdot \delta\hat{x}_{k|k-1} + \delta K_k \cdot (y_k - C \cdot \hat{x}_{k|k-1}) + O(\delta^2). \quad (53)$$

For the same reason as above, the matrix  $[R_{k-1}^e \cdot R_{k-1}^{e-1}]$  is a contraction when  $\Sigma = I$ , which guarantees the convergence to zero of  $\delta K_k$  and  $\delta\hat{x}_{k+1|k}$ . Notice, however, that here the contraction becomes closer to the identity matrix as  $k$  increases,

which suggests that the inherent decaying of errors performed in previous steps will be less apparent for this filter. Besides that, nothing is claimed about  $\delta L_k$  or  $\delta P_{k+1|k}$ , but apparently these are less important for this implementation of the KF since they do not directly affect the precision of the estimate  $\hat{x}_{k+1|k}$ . Moreover, when  $\Sigma$  is not the identity matrix, the above matrix has norm larger than 1 and divergence may be expected. This has also been observed experimentally as shown in Section IV.

We now turn our attention to the numerical errors performed in one single step  $k$ . Bounds for these errors are derived in the following theorem.

**Theorem 1:** Denoting the norms of the absolute errors due to roundoff during the construction of  $P_{k+1|k}$ ,  $K_k$ ,  $\hat{x}_{k+1|k}$ ,  $S_k$ ,  $T_k$ ,  $P_{k+1|k+1}^{-1}$ , and  $\hat{x}_{k|k}$  by  $\Delta_p$ ,  $\Delta_k$ ,  $\Delta_x$ ,  $\Delta_s$ ,  $\Delta_t$ ,  $\Delta_{pinv}$ , and  $\Delta_x$ , respectively, we obtain the following upper bounds (where all norms are 2-norms):

1) CKF

$$\Delta_p \leq \epsilon_1 \cdot \sigma_1^2 / \sigma_p^2 \cdot \|P_{k+1|k}\|$$

$$\Delta_k \leq \epsilon_2 \cdot \sigma_1^2 / \sigma_p^2 \cdot \|K_k\|$$

$$\begin{aligned} \Delta_x \leq \epsilon_3 \cdot (\|F_k\| \cdot \|\hat{x}_{k|k-1}\| + \|K_k\| \cdot \|y_k\| + \|D_k\| \cdot \|u_k\|) \\ + \Delta_k \cdot (\|C_k\| \cdot \|\hat{x}_{k|k-1}\| + \|y_k\|) \end{aligned}$$

2) SRFB

$$\Delta_s \leq \epsilon_4 \cdot (1 + \sigma_1 / \sigma_p) \cdot \|S_{k+1}\| / \cos \phi_1$$

$$\Delta_p \leq \epsilon_5 \cdot (1 + \sigma_1 / \sigma_p) \cdot \|P_{k+1|k}\| / \cos \phi_1$$

$$\Delta_k \leq \epsilon_6 / \sigma_p \cdot (\sigma_1 / \sigma_p \cdot \|S_{k+1}\| + \sigma_1 \cdot \|G_k\| + \|S_{k+1}\| / \cos \phi_1)$$

$$\begin{aligned} \Delta_x \leq \epsilon_7 \cdot (\|F_k\| \cdot \|\hat{x}_{k|k-1}\| + \|K_k\| \cdot \|y_k\| + \|D_k\| \cdot \|u_k\|) \\ + \Delta_k \cdot (\|C_k\| \cdot \|\hat{x}_{k|k-1}\| + \|y_k\|) \end{aligned}$$

3) CSRFB

$$\Delta_k \leq \epsilon_8 \cdot \kappa(U_2) / \sigma_p \cdot (\sigma_1 / \sigma_p \cdot \|L_k\|$$

$$+ \sigma_1 \cdot \|G_k\| + \|L_k\| / \cos \phi_2)$$

$$\Delta_x \leq \epsilon_9 \cdot (\|F_k\| \cdot \|\hat{x}_{k|k-1}\| + \|K_k\| \cdot \|y_k\| + \|D_k\| \cdot \|u_k\|)$$

$$+ \Delta_k \cdot (\|C_k\| \cdot \|\hat{x}_{k|k-1}\| + \|y_k\|)$$

4) SRIF

$$\Delta_t \leq \epsilon_{10} \cdot \{\kappa(A_k) + \kappa(R_k^{1/2}) + \tau_1 / \tau_m \cdot [\kappa(Q_k^{1/2})$$

$$+ \kappa(A_k)]\} \cdot \|T_{k+1}\| / \cos \phi_3$$

$$\Delta_{pinv} \leq \epsilon_{11} \cdot \{\kappa(A_k) + \kappa(R_k^{1/2}) + \tau_1 / \tau_m \cdot [\kappa(Q_k^{1/2})$$

$$+ \kappa(A_k)]\} \cdot \|P_{k+1|k+1}^{-1}\| / \cos \phi_3$$

$$\Delta_p \leq \Delta_{pinv} \cdot \|P_{k+1|k+1}\|^2$$

$$\Delta_x \leq \epsilon_{12} \cdot \|D_k\| \cdot \|u_k\|$$

$$+ \Delta_t \cdot [\kappa^2(T_{k+1}) \cdot \|r_{k+1}\| + \kappa(T_{k+1}) \cdot \|\hat{x}_{k+1|k+1}\|$$

$$+ \|r_{k+1}\| / \cos \phi_4]$$

where  $\sigma_i$  and  $\tau_i$  are the  $i$ th singular value of  $R_k^{e1/2}$  and  $Q_{k+1}^{e-1/2}$ , respectively,  $\epsilon_i$  are constants close to the machine precision  $\epsilon$  and  $\cos \phi_i$  are defined as follows:

$$\cos \phi_1 = \|S_{k+1}\| / \|[G_k]S_{k+1}\|$$

$$\cos \phi_2 = \|L_k\| / \|[G_k]L_k\|$$

$$\cos \phi_3 = \|T_{k+1}\| / \left\| \begin{pmatrix} T_k A_k^{-1} \\ R_{k+1}^{-1/2} C_{k+1} \end{pmatrix} \right\|$$

$$\cos \phi_4 = \|r_{k+1}\| / \left\| \begin{pmatrix} \hat{x}_{k+1|k+1} \\ r_{k+1} \end{pmatrix} \right\|$$

and are usually close to 1.

*Proof:*

1) *CKF:* Using Lemma A.1 the errors performed when constructing the matrix  $R_k^e$  can be bounded by  $\epsilon_r \cdot \|R_k^e\|$  and those for its inverse by  $\epsilon_r \cdot \kappa(R_k^e) \cdot \|R_k^e\|$ . By again applying Lemma A.1 several times one finally obtains all bounds for  $\Delta_p$ ,  $\Delta_k$ , and  $\Delta_x$ , as given above.

2) *SRCF:* The bounds for  $\Delta_k$  and  $\Delta_x$  follow directly from Lemma A.2 since  $K_k$  and  $S_{k+1}$  are the least-squares solution and the residual, respectively, of the problem  $[A|B]$  where  $A'$  and  $B'$  are the top and bottom block rows of the prearray (12). The matrix  $A_1$  of Lemma A.2 is here the matrix  $R_k^{e1/2}$ . The bound for  $\Delta_p$  then follows directly from the bound for  $\Delta_k$  using (43) and the fact that  $\|S_{k+1}\|^2 = \|P_{k+1|k}\|$ . Finally, the bound from  $\Delta_x$  is obtained from the one for  $\Delta_k$  and from using Lemma A.1 several times.

3) *CSRF:* For the case  $\Sigma = I$ , one obtains the bound for  $\Delta_k$  as for the SRCF, from the observation that  $K_k$  is the least-squares solution of the problem  $[A|B]$  where  $A'$  and  $B'$  are the top and bottom block rows of the prearray (16). The matrix  $A_1$  of Lemma A.2 is here also the matrix  $R_k^{e1/2}$ . When  $\Sigma \neq I$ , this bound is multiplied by  $\kappa(U_2)$  from the following observation. We can use Lemma A.1 to bound the errors in constructing the prearray (which we call  $M_k$ ) by  $\epsilon_m \cdot \|M_k\|$ , and those in constructing the postarray (which we call  $N_k$ ) by

$$\epsilon_m \cdot \|U_2\| \cdot \|M_k\| = \epsilon_m \cdot \|U_2\| \cdot \|N_k\| \cdot U_2^{-1} \\ \leq \epsilon_m \cdot \kappa(U_2) \cdot \|N_k\|.$$

In terms of  $N_k$  we are now again in a problem of classical least-squares and errors in  $M_k$  and  $N_k$  are related by a factor  $\kappa(U_2)$ , hence the bound for  $\Delta_k$  for general  $\Sigma$ . The bound for  $\Delta_x$  is then obtained by repeatedly using Lemma A.1 as for the SRCF.

4) *SRIF:* As above  $T_{k+1}$  is the residual of a least-squares problem where  $A$  and  $B$  are the first and second block columns of the prearray (22). The relative backward errors ( $\delta_a$  and  $\delta_b$  in the Appendix) in these matrices  $A$  and  $B$  are, according to Lemma A.1, bounded by  $\kappa(R_k^{1/2}) + \kappa(A_k)$  and  $\kappa(Q_k^{1/2}) + \kappa(A_k)$ , respectively. Using this and Lemma A.2 we then obtain the bound for  $\Delta_r$ . The bound for  $\Delta_{pinv}$  is then obtained from that for  $\Delta_r$  using (44) and the fact that  $\|T_{k+1}\|^2 = \|P_{k+1|k}^{-1}\|$ . The bound for  $\Delta_p$  is then on its turn obtained from that for  $\Delta_p$  using Lemma 1. Finally,  $\hat{x}_{k+1|k+1}$  is the least-squares solution of the bottom  $2 \times 2$  block in the postarray, which on itself is a residue (much as  $T_{k+1}$ ) and is therefore only known with  $\Delta_r$  precision. Using Lemma A.2 we thereby obtain the bounds for  $\Delta_x$ .

Here again we should point out that all bounds hold for several norms when appropriately adapting the constants  $\epsilon_i$  (see the Appendix). ■

These bounds are crude simplifications of the complicated process of rounding errors in linear algebra, but are often a good indication of what can go wrong in these problems (see, e.g., [18] and [19]). This will be investigated more precisely in the experimental analysis of Section IV. It is interesting to note here that the bounds derived in Theorem 1 disprove in a certain sense a result that was often used to claim the numerical supremacy of the SRF's, namely that the sensitivity of  $P_{k+1|k}$ ,  $K_k$  and  $\hat{x}_{k+1|k}$  (which according to Theorem 1 depends mainly on the singular values of  $R_k^e$ ) as computed by the SRF's is the square root of that of the same quantities computed via the CKF (see, e.g., [6], end of Section III). As far as the error analysis is concerned, this can only be claimed for  $P_{k+1|k}$  and *not* for  $K_k$  or  $\hat{x}_{k+1|k}$ , as follows from a quick comparison of the CKF and the SRF's in Theorem 1.

Therefore, we conclude that for situations that allow the application of the CKF, the SRF's *do not necessarily improve* the calculations of the Kalman gain or filtered estimates, although such a behavior is often observed. Counterexamples are given in Section IV.

Note also that when  $\kappa(R_k^e) = 1$  all quantities are computed with roughly the same accuracy in the CKF and the SRCF. This particular situation arises, e.g., when appropriately scaling the output measurements (this is also a known technique [3] to improve the performance of the CKF) or when using the "sequential processing" technique [8], described in the Introduction. This is also investigated in Section IV.

*Corollary 1:* The above theorem also gives bounds on the errors due to model deviations  $\delta A_k$ ,  $\delta B_k$ ,  $\delta C_k$ ,  $\delta D_k$ ,  $\delta Q_k$ , and  $\delta R_k$ , assuming that the latter are sufficiently small, as follows. Let  $\eta$  be the relative size of these errors, i.e.,  $\|\delta M\| \leq \eta \|M\|$  for  $M$  equal to each of the above model matrices, then the above bounds hold when replacing the  $\epsilon_i$  by numbers  $\eta_i$  which are now all of the order of  $\eta$ .

*Proof:* The model errors can indeed be interpreted as backward errors on the matrices  $A_k$ , etc., but then on a machine of precision  $\eta$ . The same analysis then holds, but with  $\epsilon$  replaced by  $\eta$ . ■

Note that other modeling errors, such as bias errors on the input signals, discretization errors, etc., do not fall under this category and a separate analysis or treatment is required for each of them (see, e.g., [10], [7]).

The above theorem is now used together with the analysis of the propagation of errors through the recursion of the KF to yield bounds on the total error of the different filters at a given step  $k$ , which we denote by the prefix  $\delta_{tot}$  instead of  $\delta$ .

For this we first turn to the (symmetrized) CKF. For the total error  $\delta_{tot} P_{k+1|k}$  we then have according to (29), (31), (33), (35) and Theorem 1 (for any consistent norm [21]):

$$\|\delta_{tot} P_{k+1|k}\| \leq \gamma_k^2 \cdot \|\delta_{tot} P_{k|k-1}\| + \bar{\Delta}_p \quad (54)$$

$$\|\delta_{tot} K_k\| \leq c_1 \cdot \gamma_k \cdot \|\delta_{tot} P_{k|k-1}\| + \bar{\Delta}_k \quad (55)$$

$$\|\delta_{tot} \hat{x}_{k+1|k}\| \leq \gamma_k \cdot \{\|\delta_{tot} \hat{x}_{k|k-1}\| + c_2 \cdot \|\delta_{tot} P_{k|k-1}\|\} + \bar{\Delta}_x \quad (56)$$

Here the upperbar on the  $\Delta$ 's indicate that these are not the exact bounds of Theorem 1 (which are derived under the assumption that the computations up to step  $k$  are exact), but analogous bounds derived for the perturbed results stored in computer at step  $k$ . Under the assumption that at step  $k$  the accumulated errors are still of the order of the local errors performed in one step (i.e., those estimated in Theorem 1), one easily finds that the  $\Delta$ - and  $\bar{\Delta}$ -quantities are  $O(\delta^2)$ -close to each other. It is thus reasonable to assume that they are equal to each other. Denoting by  $\Delta_{tot}$  the norm of the corresponding matrix  $\delta_{tot}$ , then finally yields

$$\begin{pmatrix} \Delta_{tot} P_{k+1|k} \\ \Delta_{tot} K_k \\ \Delta_{tot} \hat{x}_{k+1|k} \end{pmatrix} \leq \gamma_k \cdot \begin{pmatrix} \gamma_k & 0 & 0 \\ c_1 & 0 & 0 \\ c_2 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} \Delta_{tot} P_{k|k-1} \\ \Delta_{tot} K_{k-1} \\ \Delta_{tot} \hat{x}_{k|k-1} \end{pmatrix} + \begin{pmatrix} \Delta_p \\ \Delta_k \\ \Delta_x \end{pmatrix} \quad (57)$$

where the inequality is meant elementwise. From this one then easily sees that the total errors will remain of the order of the local errors as long as the norms  $\gamma_k$  do not remain too large for a long period of time. This is also confirmed by the experimental results of the next section. For a time-invariant system,  $\gamma_k$  can be replaced by  $\rho_k$ —if the norm is chosen appropriately as discussed in (37)—which then becomes eventually smaller than 1. Comparable results can also be stated about the  $\gamma_k$  if the time variations in the model are sufficiently smooth.

Using the above inequality recursively from 0 to  $\infty$ , one finally

obtains

$$\begin{pmatrix} \Delta_{\text{tot}} P_{\infty} \\ \Delta_{\text{tot}} K_{\infty} \\ \Delta_{\text{tot}} \hat{x}_{\infty} \end{pmatrix} \leq \begin{pmatrix} 1/(1-\hat{\gamma}^2) & 0 & 0 \\ c_1 \hat{\gamma}/(1-\hat{\gamma}^2) & 1 & 0 \\ c_2 \hat{\gamma}/((1-\hat{\gamma}^2)(1-\hat{\gamma})) & 0 & 1/(1-\hat{\gamma}) \end{pmatrix} \cdot \begin{pmatrix} \Delta_p \\ \Delta_k \\ \Delta_x \end{pmatrix} \quad (58)$$

if  $\hat{\gamma} < 1$ , where  $\hat{\gamma}$  is the largest of the  $\gamma_k$ 's. When  $\gamma_k$  tends to a fixed value  $\gamma_{\infty}$  it is easily shown that  $\hat{\gamma}$  can be replaced by  $\gamma_{\infty}$  in (58), since the contributing terms to the summation are those with growing index  $k$ . For a *time-invariant* system, finally, this can then be replaced by  $\rho_{\infty}$  as was remarked earlier, and the condition  $\hat{\gamma} = \rho_{\infty} < 1$  is then always satisfied.

For the SRCF, one uses the relation to the CKF (as far as the propagation of errors from one step to another is concerned) to derive (58) in an analogous fashion, but now with  $\Delta_p$ ,  $\Delta_k$ , and  $\Delta_x$  appropriately adapted for the SRCF as in Theorem 1. For the SRIF one also obtains analogously the top and bottom inequalities of (57) for  $\Delta_p$  and  $\Delta_x$  adapted for the SRIF as in Theorem 1 and where now  $\hat{\gamma}$  is the largest of the  $\tilde{\gamma}_k$ 's. Upon convergence, the same remarks hold as above for replacing  $\hat{\gamma}$  by  $\tilde{\gamma}_{\infty}$  and  $\rho_{\infty}$ . Finally, for the CSRF, we can only derive from (52), (53) a recursion of the type

$$\begin{pmatrix} \Delta_{\text{tot}} K_k \\ \Delta_{\text{tot}} \hat{x}_{k+1|k} \end{pmatrix} \leq \begin{pmatrix} \beta_k & 0 \\ c_2 & \gamma_k \end{pmatrix} \cdot \begin{pmatrix} \Delta_{\text{tot}} K_{k-1} \\ \Delta_{\text{tot}} \hat{x}_{k|k-1} \end{pmatrix} + \begin{pmatrix} \Delta_k \\ \Delta_x \end{pmatrix} \quad (59)$$

where  $\beta_k = \|R_{k-1}^e \cdot R_k^{e-1}\|_2$ . Recursive summation of these inequalities as was done to obtain (58), only converge here—for both  $\Delta_{\text{tot}} K_{\infty}$  and  $\Delta_{\text{tot}} \hat{x}_{\infty}$ —when the  $\beta_k$  increase sufficiently slow to 1 as  $k$  grows. We remark here that these are only upper bounds (just as the bounds for the other filters), but the fact that they may diverge does indeed indicate that for the CSRF numerical problems are more likely to occur.

Notice that the first-order analysis of the section collapses when  $O(\delta^2)$  and  $O(\delta)$  errors become comparable. According to Lemma 1, this happens when  $\kappa(R_k^e) \approx 1/\delta$ , but in such cases it is highly probable that divergence will occur for all filter implementations.

#### IV. EXPERIMENTAL EVALUATION OF THE DIFFERENT KF'S

In this section we show a series of experiments reflecting the results of our error analysis. For these examples the upper bounds for numerical roundoff developed in the previous section are reasonably close to the true error build up.

##### A. Experimental Setup

The simulations are performed for a realistic flight-path reconstruction problem, described in [10]. The *numerical difficulties* observed in a preliminary experimental analysis with the CKF [11] showed that this case study is ideally suited to validate the theoretical analysis of Section III. Reversely, it demonstrates how this first-order perturbation study contributes in understanding and solving these difficulties. In order to shed more light on the trouble spots of some of the filters, we have "artificially" modified the realistic conditions of our problem (see Table II). We then show that the behavior of the different filters can be predicted by the error analysis of Section IV. This analysis indicated the following parameters as being relevant for the error propagation in the four different KF implementations we considered.

- 1) The initial condition for the error covariance matrix  $P_{0|0-1}$ .
- 2) The condition number  $\kappa(R_k^e)$  of the innovation signal covariance matrix. In our example, it turns out that  $\kappa(R_k)$  approximately determines  $\kappa(R_k^e)$  during the whole run. This is partly due to the fact that  $\|P_{k|k-1}\|$  is small compared to  $\|R_k\|$ .
- 3) The spectral norm  $\gamma_k$  and radius  $\rho_k$  of the matrix  $F_k$ . This can be affected by "weighting" of the system matrix  $A_k$  by a factor  $c_w$ .

4) The condition number  $\kappa(Q_k)$  of the process noise covariance matrix.

5) The condition number  $\kappa(A_k)$  of the system state transition matrix. This is affected by the choice of a state-space coordinate system.

6) The condition number  $\kappa(S_k)$  of the Choleski factor of the error covariance matrix. This parameter is hard to estimate *a priori*.

These are also the parameters we tried to influence in our experimental setup as given in Table II.

To study roundoff errors in single precision, mixed precision computations were carried out and *double precision* results are considered to be exact. The roundoff errors on three different quantities that result from a KF were considered in the simulations, namely:

- 1) on the state error covariance matrix  $P$ , denoted by

$$\Delta_{\text{tot}} P_{k|k-1} = \|P_{k|k-1} - \bar{P}_{k|k-1}\| = \|\delta_{\text{tot}} P_{k|k-1}\|;$$

- 2) on the Kalman gain  $K$ , denoted by  $\Delta_{\text{tot}} K_k = \|K_k - \bar{K}_k\| = \|\delta_{\text{tot}} K_k\|;$

- 3) on the reconstructed state quantities  $\hat{x}_{k|k-1}$  or  $\hat{x}_{k|k}$ , denoted by

$$\Delta_{\text{tot}} x_k = \|\hat{x}_{k|k-1} - \bar{x}_{k|k-1}\| \text{ or } \|\hat{x}_{k|k} - \bar{x}_{k|k}\|.$$

In the experiments, the total roundoff error  $\Delta_{\text{tot}}$  in (57) and (59) is approximated by the Frobenius norm of the difference between the single and double precision quantities, which are, respectively, denoted by  $(\bar{\cdot})$  and  $(\cdot)$ . For the state error covariance matrix  $P_{k|k-1}$  this approximation becomes  $\Delta_{\text{tot}} P_{k|k-1} = \|P_{k|k-1} - \bar{P}_{k|k-1}\| = \|\delta_{\text{tot}} P_{k|k-1}\|$ . It is noted that the SRIF does not require the Kalman gain  $K_k$  explicitly to compute the filtered state quantities. Therefore, the second parameter will not be considered for this implementation.

Since the accuracy of the first two quantities determines the accuracy of the reconstructed state, a first analysis can be *restricted* to these quantities. If conditions can be formulated under which accuracy degradation of these two quantities occurs, extensive simulation tests with input and output time histories of the real (or simulated) system become obsolete.

Because of the inclusion of the CSRF, only the *time-invariant* case will be considered here. The SRCF and the SRIF algorithms are closely related from a numerical point of view. They are, therefore, first compared to the CKF and second to the CSRF.

##### B. Comparing the SRCF/SRIF with the CKF

The experimental conditions of the different tests are listed in Table II. From the theoretical analysis of Section III it follows that the relevant parameters that influence the reliability of the CKF are  $\kappa(R_k^e)$  and  $\rho(F_k)$ , the spectral radius of  $F_k$ . Two tests were performed to analyze their effect. The magnitudes of the variables  $\kappa(R^e)$  and  $\rho(F)$  are very close to the values of  $\kappa(R)$  and  $\rho(A)$  given in Table II. Two tests were performed to analyze their effect. The results of these tests are plotted in Fig. 1. From this figure the following observations are made.

- 1) Test 1—Fig. 1(a): ( $\rho(A) = 1.0$  and  $\kappa(R) = 10^2$ ).

Since symmetry of the error state covariance matrix  $P$  is not preserved by the CKF, the roundoff error propagation model for the local error  $\delta P_{k|k-1}$ , given by (31), learns that divergence of roundoff errors on  $P$ , and hence on  $K$  will occur if the original system is *unstable*. This experiment confirms this divergence phenomenon also when  $\rho(A) = 1.0$ , as is the case for the considered flight-path reconstruction problem [10]. Furthermore, it is observed from Fig. 1(a) that the error on  $P$  with the CKF is almost completely determined by the *loss of symmetry*, computed by  $\|\bar{P}_{k|k-1} - P'_{k|k-1}\| = \Delta_{\text{sym}} P_{k|k-1}$ .

As indicated in the previous section, different methods have been proposed to solve this problem. One particular class of methods consists of forcing the error on the state covariance



TABLE II  
TEST CONDITIONS TO EVALUATE THE DIFFERENT KF  
IMPLEMENTATIONS

| quantity         | Test1                | Test2                | Test3                |
|------------------|----------------------|----------------------|----------------------|
| $\kappa(A)$      | $1.46 \cdot 10^4$    | $1.46 \cdot 10^4$    | $1.46 \cdot 10^4$    |
| $\kappa(Q)$      | 1.0                  | 1.0                  | 1.0                  |
| $\kappa(R)$      | $9.90 \cdot 10^2$    | $9.90 \cdot 10^2$    | 1.0                  |
| $\kappa(R^e/2)$  | 7.53                 | 8.06                 | 1.02                 |
| $c_w$            | 1.0                  | 0.9                  | 1.0                  |
| $\rho(P_\infty)$ | $9.80 \cdot 10^{-1}$ | $9.00 \cdot 10^{-1}$ | $9.80 \cdot 10^{-1}$ |
| $P_{0 -1}$       | $\neq 0$             | $\neq 0$             | $\neq 0$             |
| $\ P_\infty\ $   | $2.06 \cdot 10^{-1}$ | $1.97 \cdot 10^{-3}$ | $6.96 \cdot 10^{-2}$ |
| $\ K_\infty\ $   | $4.37 \cdot 10^{-1}$ | $2.50 \cdot 10^{-1}$ | $1.29 \cdot 10^{-1}$ |
| $\ S_\infty\ $   | $1.56 \cdot 10^{-1}$ | $4.91 \cdot 10^{-2}$ | $2.98 \cdot 10^{-1}$ |

| quantity         | Test 4               | Test 5               | Test 6               |
|------------------|----------------------|----------------------|----------------------|
| $\kappa(A)$      | $1.46 \cdot 10^4$    | $2.36 \cdot 10^6$    | $1.46 \cdot 10^4$    |
| $\kappa(Q)$      | 1.0                  | $1.39 \cdot 10^6$    | $9.99 \cdot 10^5$    |
| $\kappa(R)$      | 1.0                  | $1.02 \cdot 10^5$    | 3.4                  |
| $\kappa(R^e/2)$  | 1.02                 | $1.35 \cdot 10^1$    | 1.67                 |
| $c_w$            | 1.0                  | 1.0                  | 1.0                  |
| $\rho(P_\infty)$ | $9.80 \cdot 10^{-1}$ | $9.00 \cdot 10^{-1}$ | $9.95 \cdot 10^{-1}$ |
| $P_{0 -1}$       | $= 0$                | $\neq 0$             | $\neq 0$             |
| $\ P_\infty\ $   | $6.96 \cdot 10^{-2}$ | $1.98 \cdot 10^{-3}$ | $3.36 \cdot 10^{-3}$ |
| $\ K_\infty\ $   | $1.29 \cdot 10^{-1}$ | 1.21                 | $1.81 \cdot 10^{-2}$ |
| $\ S_\infty\ $   | $2.98 \cdot 10^{-1}$ | $4.49 \cdot 10^{-2}$ | $6.09 \cdot 10^{-2}$ |

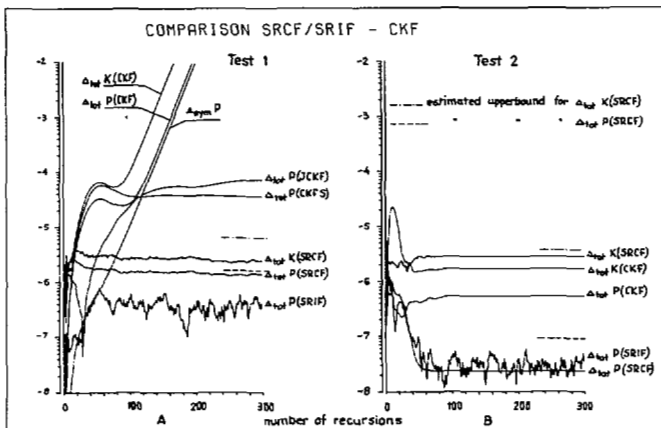


Fig. 1. Comparison of the SRCF/SRIF and the CKF (a)-(b).

matrix to become symmetric, which is done here by averaging the off-diagonal elements of  $P$  after each recursion. The behavior of  $\Delta_{\text{tot}} P_{k|k-1}$  for this implementation, denoted by CKF(S) in Fig. 1(a), clearly indicates that this implementation becomes again competitive, even when the original system is unstable. A similar effect has also been observed when computing only the upper triangular part of  $P$ . On the other hand, the behavior of  $\Delta_{\text{tot}} P_{k|k-1}$  for Joseph's stabilized CKF, denoted by (J)CKF in Fig. 1(a), confirms that the roundoff errors do not diverge even when the symmetry of  $P$  is not retained. We also observe from Fig. 1(a) that the roundoff error on  $P$  with these modified CKF remains higher (a factor 10) than the SRCF/SRIF

2) Test 2—Fig. 1(b): ( $\rho(A) = 0.9$  and  $\kappa(R) = 10^2$ ).

If we make the original system stable, the CKF is numerically stable. Moreover, the accuracy with which the Kalman gain is computed is of the same order as that of the SRCF. This is in contrast with a general opinion that SRF's improve the calculations of the Kalman gain or filtered estimates [6], [3]. We can state that they do not make accuracy poorer. From Fig. 1(b) it is observed that only the error covariance matrix  $P$  is computed more accurately, which confirms the upperbounds for the roundoff errors as given in Section III. Summarizing these bounds, we

obtained in Section III the following recurrences:

$$\Delta_{\text{tot}} P_{k+1|k} \leq \gamma_k^2 \cdot \Delta_{\text{tot}} P_{k|k-1} + \Delta_p \quad (60)$$

$$\Delta_{\text{tot}} K_k \leq c \cdot \gamma_k \cdot \Delta_{\text{tot}} P_{k|k-1} + \Delta_k \quad (61)$$

where the upperbounds for the local errors  $\Delta_p$  and  $\Delta_k$  are given in Theorem 1, parts 1 and 2, for the CKF and the SRCF.

A comparison of the (a) and (b) bounds indicates that when the accuracy of the Kalman gain is considered no preference should exist for the SRF's to the CKF when  $A_k$  is *stable and time-invariant*. (For situations where  $A_k$  has eigenvalues on or outside the unit circle, the CKF has to be changed, e.g., to the CKF(S) implementation.) However, the experimental results demonstrate that for the latter conditions the loss of accuracy with a CKF(S) is still higher than the SRF's. This is also generally observed for other SRF variants such as the UDU' filters [7]. Here we only want to draw attention to the clear difference to be expected (and also reflected by the experiments) between the accuracy of  $P_{k|k-1}$  and  $K_k$  in the CKF(S) implementation with respect to those of SRF filters.

### C. Comparison SRCF/SRIF with CSRF

The upperbound for the roundoff errors of the Kalman gain and the state estimate  $\hat{x}_{k+1|k}$  computed by the CSRF (for large  $k$ ) can be summarized as follows:

$$\Delta_{\text{tot}} K_k \leq \beta_k \cdot \Delta_{\text{tot}} K_{k-1} + \Delta_k \quad (62)$$

$$\Delta_{\text{tot}} \hat{x}_{k+1|k} \leq c \cdot \Delta_{\text{tot}} K_{k-1} + \gamma_k \Delta_{\text{tot}} \hat{x}_{k|k-1} + \Delta_x \quad (63)$$

with the upperbounds for the local errors  $\Delta_k$  and  $\Delta_x$  given in Theorem 1, part 3. This model indicates that the error propagation is convergent when  $\beta_k = \|R_{k-1}^e \cdot (R_k^e)^{-1}\| < 1$ , which is the case only if the signature matrix  $\Sigma$  is the identity matrix  $I$ . Note that the error variation  $\Delta_{\text{tot}} K_k$  is now weighted by  $\beta_k$  (instead of  $\gamma_k$  for the other filters), which even for  $\Sigma = I$  becomes very close to 1 for large  $k$ . This is also the main reason of the poor numerical behavior of this filter. When  $\Sigma \neq I$  (which depends on the choice of  $P_{0|-1} \neq 0$ )  $\beta_k$  is larger than 1 and  $\kappa(U_2)$  may also become large. Both these phenomena have a negative influence on the above bounds and may eventually cause divergence. In addition to the *numerical sensitivity* introduced by the choice of  $P_{0|-1}$ , it also *influences the efficiency* of the CSRF implementation. This is indicated by the parameter ( $n_1 + n_2$ ) in Table I, which lies in the interval  $[n, \min(m, p)]$ .

The influence of the choice of  $P_{0|-1}$  is analyzed by the following two tests.

1) Test 3—Fig. 2(a) ( $P_{0|-1} \neq 0$ ,  $\rho(A) = 1.0$  and  $\kappa(R) = 1.0$ ).

The choice of  $P_{0|-1} \neq 0$  influences the CSRF implementation *negatively*. First, in this experiment the computational efficiency decreases in comparison to the case  $P_{0|-1} = 0$ , discussed in the following test. This is because ( $n_1 + n_2$ ) in Table I becomes greater than  $p$  or  $m$ . This was the case for all the tests performed with  $P_{0|-1} \neq 0$ . Second, the transformations used in each recursion to triangularize the prearray become  $\Sigma$ -unitary, i.e., having a condition number  $> 1$ . This is due to the fact that  $\text{inc } P_0$  is not definite. From Fig. 2(a) this negative effect is clearly observed. Both the error levels on  $P$  and  $K$  are a factor  $10^2$  larger than for the SRCF or SRIF. For the covariance type algorithms considered here, it is observed that the error on the Kalman gain is always higher than the error on the state error covariance matrix. This is partly due to the extra calculation  $G_k(R_k^e)^{-1/2}$  needed for the Kalman gain, where the condition number of  $(R_k^e)^{1/2}$  determines the loss of accuracy.

2) Test 4—Fig. 2(b) ( $P_{0|-1} = 0$ ,  $\rho(A) = 1.0$  and  $\kappa(R) = 1.0$ ).

For this case  $\text{inc } P_0 = B \cdot Q \cdot B'$  is positive definite, causing the transformations used in each recursion to be *unitary*. On the other



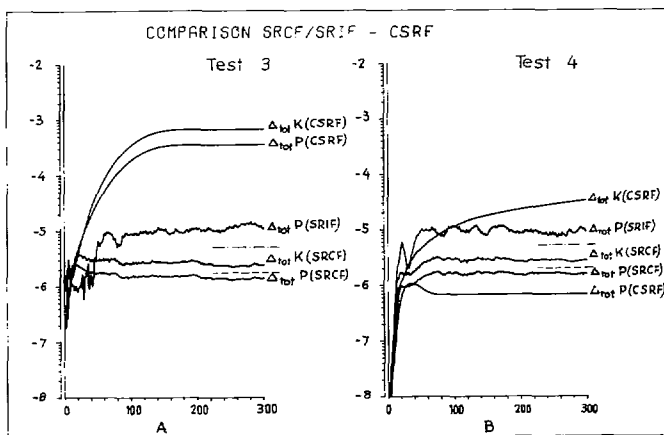


Fig. 2. Comparison of the SRCF/SRIF and the CSRF (a)–(b).

hand ( $n_1 + n_2$ ) in Table I is equal to  $m$ , what makes the CSRF “slightly” more efficient compared to the new SRCF/SRIF implementation based on the “condensed” system representations.

From the experimental results in Fig. 2(b) we observe that the error on  $P$  is very small, while the error on  $K$  is much higher than for the SRCF calculations. Furthermore, the errors on  $K$  with the CSRF increase very slowly because the coefficient  $\beta_k$  becomes very close to 1. This is due to the fact that for the CSRF roundoff errors are carried along on three matrices, namely  $G_k$ ,  $(R_k^e)^{1/2}$ , and  $L_k$ , while for the SRCF/SRIF errors are carried along only on the square roots of  $P$  or  $P^{-1}$ . For the error on  $L_k$  (supposing inc  $P_k$  factored as  $L_k L_k'$ ) this effect does not cause the errors on  $P_k$

$$P_k = \sum_{i=0}^{k-1} L_i L_i' + P_0 \quad (64)$$

to accumulate because  $L_k$  converges rapidly enough to zero such that the accumulated errors on  $P_k$ :

$$\Delta_{\text{tot}} P_k = \sum_{i=0}^k L_i \cdot \Delta_{\text{tot}} L_i' + \Delta_{\text{tot}} L_i \cdot L_i' \quad (65)$$

also convergences if the  $\Delta_{\text{tot}} L_i$  are not too large. The absolute value of the total error on  $(R_k^e)^{1/2}$  and  $G_k$  remain much higher. This is clearly reflected in the loss of accuracy in the calculation of  $K_k$  by  $G_k (R_k^e)^{-1/2}$ .

Generally, the CSRF is less reliable than the SRCF/SRIF combination. For zero initial conditions of the state error covariance matrix maximal reliability can be achieved with the CSRF. Therefore, for situations where  $n \gg m$ , the CSRF may be preferred because of its increased computational efficiency despite its loss of accuracy. We stress the fact that this property is only valid for the time-invariant case. Modifications of the CSRF exist taking into account certain time-varying effects [15], e.g., for the process noise covariance matrix  $Q$ . This, however, induces again an increased computational complexity.

#### D. Comparison of the SRCF and the SRIF

In the previous experiments the SRCF/SRIF combination performed equally well. In this section a further analysis is made to compare both implementations. Using the error model that indicates the upperbound for the roundoff errors made during one SRIF recursion:

$$\Delta_{\text{tot}} P_{k+1|k+1} \leq \tilde{\gamma}_k^2 \cdot \Delta_{\text{tot}} P_{k|k} + \Delta_p \quad (66)$$

$$\Delta_{\text{tot}} \hat{x}_{k+1|k+1} \leq \tilde{\gamma}_k \cdot \Delta_{\text{tot}} \hat{x}_{k|k} + \tilde{c}_2 \tilde{\gamma}_k \cdot \Delta_{\text{tot}} P_{k|k} + \Delta_x \quad (67)$$

with the upperbounds of the local errors  $\Delta_p$  and  $\Delta_x$  given in Theorem 1, part 4, learns that besides  $\kappa(R_k)$  and  $\rho(F_k)$ , other system parameters influence the roundoff error accumulation in the SRIF. The effect of these parameters is analyzed in the following tests.

#### 1) Test 5—Fig. 3(a).

In this test very large condition numbers for  $A$ ,  $Q$ , and  $R$  (see Table II), are considered. As expected, this indeed causes the error on  $P$  to be much higher (a factor  $10^3$ ) for the SRIF than for the SRCF. As in test 2, the large value of  $\kappa(R)$  again causes a great loss in the accuracy of the Kalman gain calculation in the SRCF. The level of roundoff errors on  $K$  indeed becomes a factor  $10^2$  larger than the roundoff level of  $P$ .

In this test we analyzed the deterioration of the error covariance matrix by the SRIF implementation by (fairly unrealistic) large condition numbers. In many practical situations, the effect of high  $\kappa(Q_k^e)$  and  $\kappa(R_k^e)$  can be relaxed by scaling, rearranging the system matrices or using scalar measurement and/or input updates [12]. Furthermore, we observed in the experiments that a high  $\kappa(Q_k)$  did not result in a high  $\kappa(Q_k^e)$ , which is in contrast with what was observed for  $\kappa(R_k^e)$ . However, the effect of a high  $\kappa(A_k)$  is much harder to control and as we have seen may influence the accuracy of the SRIF negatively. We repeat here that this is due to a careful choice of the problem coefficients [here  $\kappa(A_k)$  and  $\kappa(Q_k)$ ] in order to put forward the dependency on these parameters.

#### 2) Test 6—Fig. 3(b).

For this test, the measurement error statistics were taken from real flight-test measurement calibrations [16]. This results in the following forms for the process noise covariance matrix  $Q$ , respectively, the measurement noise covariance matrix  $R$ :

$$Q = \text{diag} \{8 \cdot 10^{-6}, 5 \cdot 10^{-5}, 5 \cdot 10^{-8}\}, R = \text{diag} \{5 \cdot 10^{-2}, 2 \cdot 10^{-1}\}. \quad (68)$$

The relevant parameters for the roundoff error propagation are listed in Table II. In Fig. 3(b) the simulated error  $\Delta_{\text{tot}} x$  on the state calculations is plotted for both filter implementations. Here, the error level with the SRIF is significantly higher than that for the SRCF, while  $P$  is computed with roughly equal accuracy. This is due to the high condition number of  $T_k$  (obtained by the test conditions given in Table II) in the calculation of the filtered state with the SRIF by (23).

### V. COMPARISON OF THE DIFFERENT FILTERS

In this section we compare the different filter implementations based on the error analysis of Section III strengthened by the simulation study of Section IV and the complexity analysis of Section II-E.

We first look at the time-varying case (hence excluding the CSRF). According to the error bounds of Theorem 1, it appears that the SRCF has the lowest estimate for the local errors generated in a single step  $k$ . The accumulated errors during subsequent steps is governed by the norms  $\gamma_k$  for all three filters in a similar fashion (at least for the error on the estimate)—this of course under the assumption that a “symmetrized” version of the CKF or the stabilized CKF is considered. From these modifications, the implementation computing only the upper (or lower) triangular part of the state error covariance matrix is the most efficient. The experiments of Section IV with the realistic flight path reconstruction problem indeed demonstrate that the CKF, the SRCF, and the SRIF seem to yield a comparable accuracy for the estimates  $\hat{x}_{k+1|k}$  or  $\hat{x}_{k+1|k+1}$ , unless some of the “influential” parameters in the error bounds of Theorem 1 become critical. This is, e.g., true for the SRIF which is likely to give worse results when choosing matrices  $A_k$ ,  $R_k$ , or  $Q_k$  that are hard to invert. As far as  $R_k$  or  $Q_k$  is concerned, this is in a sense an artificial disadvantage since in some situations the inverses  $R_k^{-1}$

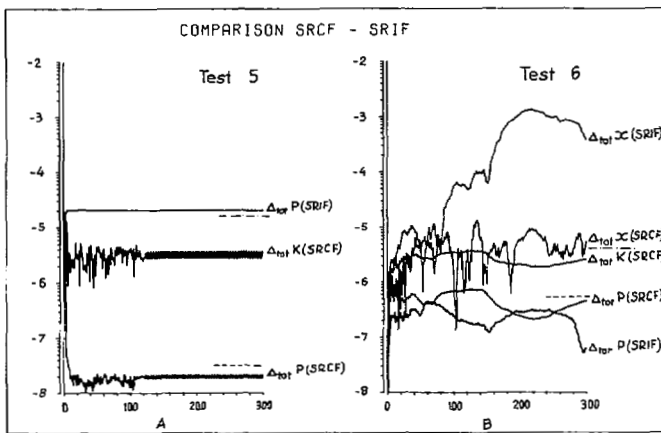


Fig. 3. Comparison of the SRCF and the SRIF (a)-(b).

and  $Q_k^{-1}$  are the given data and the matrices  $R_k$  and  $Q_k$  have then to be computed. This then would of course disadvantage the SRCF. In [23] it is shown that the problems of inverting covariances can always be bypassed as well for the SRIF as for the SRCF. The problem of inverting  $A_k$ , on the other hand, is always present in the SRIF.

For the computational cost, the SRCF/SRIF have a marginal advantage over the CKF when  $n$  is significantly larger than  $m$  and  $p$  (which is a reasonable assumption in general), even when computing the upper (or lower) triangular part of  $P$  with the CKF. Moreover, preference should go to the SRCF (respectively, SRIF) when  $p < m$  (respectively,  $p > m$ ), with a slight preference for the SRCF when  $p = m$ . As is shown in [10], [14], condensed forms or even the CSRF can sometimes be used in the time-varying case as well, when, e.g., only some of the matrices are time-varying or when the variations are structured. In that case the latter two may yield significant savings in computing time. Similarly, considerable savings can be obtained by using sequential processing [7] when diagonal covariances are being treated (which is often the case in practice).

For the time-invariant case, the same comments as above hold for the accuracy of the CKF, SRCF, and SRIF. The fourth candidate, the CSRF, has in general a much poorer accuracy than the other three. This is now not due to pathologically chosen parameters, but to the simple fact that the accumulation of rounding errors from one step to another is usually much more significant than for the three other filters, as was pointed out in Section III. This is particularly the case when the signature matrix  $\Sigma$  is not the identity matrix, which may then lead to divergence as shown experimentally in Section IV.

As for the complexity, the Hessenberg forms of the SRCF and the SRIF seem to be the most appealing candidate, except when the coefficient  $n_1 + n_2$  in Table I for the CSRF is much smaller than  $n$ . This is, e.g., the case when the initial covariance  $P_{0|0}$  is zero, in which case the CSRF becomes the fastest of all four filters. Although the Schur implementations of the SRCF and SRIF are almost as fast as the Hessenberg implementations, they also have the small additional disadvantage that the original state-space transformation  $U$  for condensing the model to Schur form is more expensive than that for the other condensed forms and that the (real) Schur form is not always exactly triangular but may contain some  $2 \times 2$  "bumps" on the diagonal (corresponding to complex eigenvalues of the real matrix  $A$ ). Finally, the c-Hess. form of the SRIF (given in Table I) requires a more complex initial transformation  $U$ , since it is constructed from the pair  $(A^{-1}, A^{-1}B)$  which also may be numerically more delicate due to the inversion of  $A$ .

As a general conclusion, we recommend the SRCF, and its observer-Hessenberg implementation in the time-invariant case, as the optimal choice of KF implementation because of its good

balance of reliability and efficiency. Other choices may of course be preferable in some specific cases because of special conditions that would then be satisfied.

## VI. CONCLUDING REMARKS

In this paper we have analyzed four different KF algorithms and some new variants for their reliability and computational efficiency. We note here that our implementations may differ substantially from similarly named algorithms described in [7]. The comparison is based on an error analysis and an operation count where we have made full use of possible savings in the time-invariant case.

From the error models a better insight is also obtained about which parameters influence the error propagation in the different KF algorithms that have been investigated. For the CKF and the SRCF these are the condition number of the innovation signal covariance matrix  $R_k^e$  and the spectral norm (radius) of the filter state transition matrix  $F_k$ , while for the SRIF the relevant parameters are the condition numbers of  $R_k$ ,  $Q_k$ ,  $Q_k^e$ ,  $A_k$  and of the Choleski factor  $T_k$  and the spectral norm (radius) of the filter state transition matrix  $\tilde{F}_k$ . For the CSRF the choice of the initial error covariance matrix  $P_{0|0}$  matrix and of the condition number of the innovation signal covariance matrix  $R_k^e$  become critical. This influence is also verified by a simulation study on the flight-path reconstruction problem [10] given in Section IV.

Further extensions of these techniques to the problem of computing other estimates  $\hat{x}_{k|j}$  (e.g., for smoothing) or using mixed representations of covariances (see, e.g., [23]) can also be considered. These mixed representations have the advantage that they allow for singular covariance matrices  $Q_k$ ,  $R_k$ , or  $P_{k|k-1}$  and even for singular information matrices  $I_{k|k} = P_{k|k}^{-1}$ , thereby avoiding any use of generalized inverses.

## A. APPENDIX

Here we briefly recall the propagation of rounding errors in some basic problems in linear algebra. The norm used is the 2-norm.

Let the matrix-vector pair  $(A, b)$  be known with relative precision  $\delta_a$  and  $\delta_b$ , respectively,

$$\delta_a = \|\delta A\|/\|A\|, \quad \delta_b = \|\delta b\|/\|b\|$$

then we have the following lemma (assuming  $A$  to be invertible).

**Lemma A.1 [20]:** The errors on the  $A \cdot b$  and  $A^{-1} \cdot b$  can be bounded by

$$\|(\overline{A \cdot b}) - (A \cdot b)\| \leq (\delta_a + \delta_b) \cdot \|A\| \cdot \|b\| + O(\delta^2)$$

$$\|(\overline{A^{-1} \cdot b}) - (A^{-1} \cdot b)\| \leq \delta_a \cdot \kappa(A) \cdot \|A^{-1} \cdot b\|$$

$$+ \delta_b \cdot \|A^{-1}\| \cdot \|b\| + O(\delta^2).$$

When the errors  $\delta_a$  and  $\delta_b$  are the backward errors of the above problem solved on a computer with machine precision  $\epsilon$ , then the above bounds are reasonably well approximated by

$$\|(\overline{A \cdot b}) - (A \cdot b)\| \leq \epsilon_1 \cdot \|A\| \cdot \|b\| \approx \epsilon_2 \cdot \|A \cdot b\|$$

$$\|(\overline{A^{-1} \cdot b}) - (A^{-1} \cdot b)\| \leq \epsilon_3 \cdot \kappa(A) \cdot \|A^{-1} \cdot b\|$$

where all  $\epsilon_i$  are of the order of  $\epsilon$ . ■

The above approximation implies that no serious cancellations occur in the product  $A \cdot b$ , which in general is a reasonable assumption.

Let  $A$  now be a  $m \times n$  matrix of rank  $m < n$  and transform the compound matrix  $[A|b]$  by a unitary transformation  $Q$  as follows:

$$Q \cdot [A|b] = \begin{pmatrix} A_1 & | & b_1 \\ 0 & & b_2 \end{pmatrix}$$

where  $A_1$  is now invertible. Then we have the following lemma (where  $\cdot^+$  denotes the generalized inverse of a matrix).

**Lemma A.2 [19]:** The errors on the least-squares solution  $A^+ \cdot b$  and the residual  $b_2$  can be bounded by

$$\begin{aligned} \|\overline{(A^+ \cdot b)} - (A^+ \cdot b)\| &\leq \delta_a \cdot \{\kappa(A) \cdot \|A^+ \cdot b\| \\ &\quad + \kappa(A) \cdot \|A^+\| \cdot \|b_2\|\} + \delta_b \cdot \|A^+\| \cdot \|b\| + O(\delta^2) \\ \|\overline{(b_2)} - (b_2)\| &\leq \delta_a \cdot \kappa(A) \cdot \|b\| + \delta_b \cdot \|b\| + O(\delta^2). \end{aligned}$$

When the errors  $\delta_a$  and  $\delta_b$  are the backward errors of the above problems solved on a computer with machine precision  $\epsilon$ , then they are both of the order of  $\epsilon$  and the above bounds are reasonably well approximated by

$$\begin{aligned} \|\overline{(A^+ \cdot b)} - (A^+ \cdot b)\| &\leq \epsilon_4 \cdot \{\kappa(A) \cdot \|A^+ \cdot b\| \\ &\quad + \kappa(A) \cdot \|A^+\| \cdot \|b_2\| + \|A^+\| \cdot \|b_2\|/\cos \phi\} \\ \|\overline{(b_2)} - (b_2)\| &\leq \epsilon_5 \cdot \{1 + \kappa(A)\} \cdot \|b_2\|/\cos \phi \end{aligned}$$

where all  $\epsilon_i$  are of the order of  $\epsilon$  and  $\cos \phi = \|b_2\|/\|b\|$ . ■  
We terminate with perturbation bounds on the QR-factorization of a matrix  $A$ .

**Lemma A.3 [22]:** Let  $A = Q \cdot R$ , where  $A$  has full column rank  $n$ ,  $Q' \cdot Q = I_n$  and  $R$  is upper triangular. Then for a small perturbation  $\bar{A}$  of  $A$  there exist perturbations  $\bar{Q}$  and  $\bar{R}$  of the factors, such that

$$\bar{A} = \bar{Q} \cdot \bar{R}$$

and

$$(A - \bar{A}) = Q \cdot (R - \bar{R}) + \Delta$$

with

$$\|\Delta\| \leq [\delta_a \cdot \kappa(A)]^2 \cdot \|A\|.$$

When the error  $\delta_a$  is the backward error of the above decomposition solved on a computer with machine precision  $\epsilon$ , then it is of the order of  $\epsilon$  and the above bound becomes

$$\|\Delta\| \leq \epsilon_6^2 \cdot \kappa(A)^2 \cdot \|A\|.$$

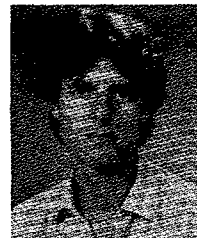
A similar result can also be found for a "skew decomposition," i.e., where  $Q' \cdot \Sigma \cdot Q = \Sigma$ , for some signature matrix  $\Sigma$ . ■

Although all these bounds are written for the 2-norm, they also hold for several other norms, up to a constant which is close to 1 and can therefore be absorbed in the  $\epsilon_i$ . This is, e.g., important when deriving the above bounds for a matrix  $B$  instead of a vector  $b$ . This is done by using the bounds for each column  $b_i$  of the matrix  $B$  and combining these bounds into a bound involving the norm of  $B$ , for which in this case the Frobenius norm is a natural choice [21]. These mixed bounds (as far as norms are concerned) can then again be formulated in terms of one norm only, by again adapting the  $\epsilon_i$  appropriately.

REFERENCES

[1] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Trans. ASME. (J. Basic Eng.)*, vol. 82D, pp. 34-45, Mar. 1960.  
 [2] A. H. Jazwinski, *Stochastic Processes and Filtering Theory*. New York: Academic, 1970.  
 [3] G. J. Bierman and C. L. Thornton, "Numerical comparison of Kalman filter algorithms: Orbit determination case study," *Automatica*, vol. 13, pp. 23-35, 1977.  
 [4] A. E. Bryson, "Kalman filter divergence and aircraft motion estimators," *Int. J. Guidance Contr.*, vol. 1, no. 1, pp. 71-79, 1978.  
 [5] J. E. Potter and R. G. Stern, "Statistical filtering of space navigation measurements," in *Proc. 1963 AIAA Guidance Contr. Conf.*, 1963.  
 [6] P. G. Kaminski, A. Bryson, and S. Schmidt, "Discrete square root

filtering: A survey of current techniques," *IEEE Trans. Automat. Contr.*, vol. AC-16, pp. 727-736, Dec. 1971.  
 [7] G. J. Bierman, *Factorization Methods for Discrete Sequential Estimation*. New York: Academic, 1977.  
 [8] J. Mendel, "Computational requirements for a discrete Kalman filter," *IEEE Trans. Automat. Contr.*, vol. AC-16, no. 6, pp. 748-758, 1971.  
 [9] T. Kailath, "Some new algorithms for recursive estimation in constant linear systems," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 750-760, Nov. 1973.  
 [10] M. H. Verhaegen, "A new class of algorithms in linear system theory, with application to real-time aircraft model identification," Ph.D. dissertation, Catholic Univ. Leuven, Leuven, Belgium, Nov. 1985.  
 [11] M. H. Verhaegen and P. Van Dooren, "An efficient implementation of square root filtering: Error analysis, complexity and simulation on flight-path reconstruction," in *Proc. INRIA Conf. Anal. Optimiz. Syst.*, Nice, June 1984, Springer-Verlag, vol. 62-63, pp. 250-267.  
 [12] B. D. O. Anderson and J. B. Moore, *Optimal Filtering*, (Information and System Sciences Series). Englewood Cliffs, NJ: Prentice-Hall, 1979.  
 [13] G. H. Golub, "Numerical methods for solving linear least squares problems," *Numerische Mathematik*, vol. 7, pp. 206-216, 1965.  
 [14] M. Morf and T. Kailath, "Square-root algorithms for least squares estimation," *IEEE Trans. Automat. Contr.*, vol. AC-20, pp. 487-497, Aug. 1975.  
 [15] M. Gentleman, "Least squares computations by Givens transformations without square roots," *JIMA*, vol. 12, pp. 329-336, 1973.  
 [16] F. R. Gantmacher, *The Theory of Matrices*. New York: Chelsea, 1960.  
 [17] R. E. Bellman, *Matrix Analysis*, 2nd ed., New York: McGraw-Hill, 1968.  
 [18] G. W. Stewart, "On the perturbation of pseudo-inverses, projections and linear least square problems," *SIAM Rev.*, vol. 19, pp. 634-662, Oct. 1977.  
 [19] A. van der Sluis, "Stability of the solution of linear least squares problems," *Numerische Mathematik*, vol. 23, pp. 241-254, 1975.  
 [20] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*. Oxford: Clarendon, 1965.  
 [21] G. W. Stewart, *Introduction to Matrix Computations*. New York: Academic, 1973.  
 [22] G. W. Stewart, "Perturbation bounds for the QR factorization of a matrix," *SIAM J. Numer. Anal.*, vol. 14, pp. 509-518, June 1977.  
 [23] C. C. Paige, "Covariance matrix representation in linear filtering," in Special Issue of Contemporary Mathematics on Linear Algebra and its Role in Systems Theory, AMS, 1985.



**Michel Verhaegen** was born in Antwerp, Belgium, on September 2, 1959. He received the engineering degree in aeronautics (Cum Laude) from the Delft University of Technology, Delft, The Netherlands, in 1982 and the doctoral degree in applied sciences from the Catholic University of Leuven, Leuven, Belgium, in 1985.

From 1982 to 1985 he was holder of an IWONL Research Assistantship in the Department of Electrical Engineering of the Catholic University of Leuven. He is currently holder of a Postdoctoral Fellowship of the National Research Council that enables him to conduct Research at the NASA Ames Laboratory, Moffett Field, CA. His research interests are mainly in the interdisciplinary domain between numerical analysis and linear system theory with emphasis on identification and filtering.



**Paul Van Dooren** (S'79-M'80) was born in Tienen, Belgium, on November 5, 1950. He received the engineering degree in computer science and the doctoral degree of applied sciences, both from the Catholic University of Leuven, Leuven, Belgium, in 1974 and 1979, respectively.

From 1974 to 1977 he was Assistant in the Department of Applied Mathematics and Computer Science of the Catholic University of Leuven. He was a Research Associate at the University of Southern California in 1978-1979, a Postdoctoral Fellow at Stanford University in 1979-1980, and a Visiting Fellow at the Australian National University in 1985. He is currently with the Philips Research Laboratory, Brussels, Belgium. His main interests lie in the areas of numerical linear algebra, linear system theory, digital signal processing, and parallel algorithms.

Dr. Van Dooren received the Householder Award IV in 1981. He is an Associate Editor of *Systems and Control Letters* and of the *Journal of Computational and Applied Mathematics*.