Operator Factorization and Other Aspects of the Analysis of Linear Systems

Hagander, Per

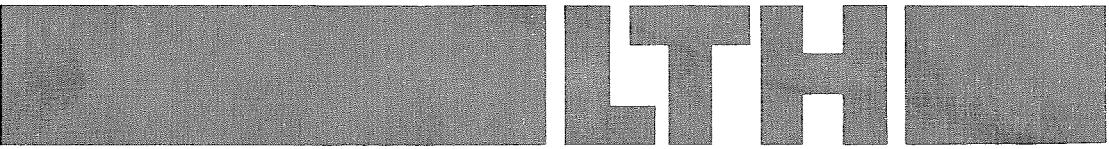
1973

Document Version:
Publisher's PDF, also known as Version of record

REPORT 7336
SEPTEMBER 1973

Operator Factorization and Other Aspects of the Analysis of Linear Systems

PER HAGANDER



Division of Automatic Control · Lund Institute of Technology

OPERATOR FACTORIZATION AND OTHER ASPECTS OF THE ANALYSIS OF LINEAR SYSTEMS

av

Per Hagander
tekn lic, Ssk

Akademisk avhandling som för avläggande av teknisk doktorsexamen vid tekniska fakulteten vid universitetet i Lund kommer att offentligen försvaras i sal M:A, Maskinhuset, Lunds Tekniska Högskola, fredagen den 7 december 1973 kl. 10.15.

To Barbro

AV Centralen
Lund 1973

OPERATOR FACTORIZATION AND OTHER ASPECTS OF THE ANALYSIS OF LINEAR SYSTEMS.

Applications of control theory appear in many branches of engineering, biology and economy. The basic questions are the same, although different aspects may be emphasized. Control theory discusses how to manipulate a process and how to describe the resulting behaviour. A system is called static if the current behaviour depends only on the current values of the control signals. For dynamical systems it may depend also on old values. Causality implies that future control actions must not have any influence on the present behaviour.

The aim of control is to use the available input signals to make the system behave as well as possible in spite of uncertainty and disturbances. A regulator determines the input signals from measurements of the process.

The formulation of mathematical models for a system usually involves many simplifications made in order to reduce complexity. Linearization is frequently used. Uncertainty can be taken into account by using random variables. The objective of the system is also formalized in mathematical terms.

Models for Linear Dynamical Systems.

There are mainly two types of descriptions of dynamical systems. If only the input-output behaviour is modelled it is natural to talk about an external description. The output at time t is then expressed as a weighted sum of the values of the input at past times, i.e.

$$y(t) = \int_{-\infty}^{t} h(t,s)u(s)ds \qquad (1)$$

for continuous time and

$$y(t) = \sum_{-\infty}^{t} h(t,s)u(s) \qquad (2)$$

for discrete time. The function h is called the weighting function.

For many processes there exists much knowledge of the internal behaviour. Physical laws may govern the dynamics, and it may be possible to define a number of variables, the state, that summarizes the influence of the past. The concept originates from classical mechanics, where positions and velocities are typical state variables.

First order linear differential or difference equations describing the change of the state are natural models. The measurements are assumed to be linear combinations of the state variables and the inputs. Vector and matrix notation is widely used. A continuous time system could thus be written

$$\begin{cases} \frac{d}{dt}x(t) = Ax(t) + Bu(t) & x(t_0) = x_0 \\ y(t) = Cx(t) + Du(t) \end{cases} \qquad (3)$$

and correspondingly in the discrete time case:

$$\begin{cases} x(t+1) = \phi x(t) + \Gamma u(t) & x(t_0) = x_0 \\ y(t) = \theta x(t) + Du(t) \end{cases} \qquad (4)$$

It is possible to obtain an input-output description

from all internal descriptions like (3) and (4). In fact, the models (1) and (2) are somewhat more general since it may not be possible to find a finite number of state variables.

The external model is just a linear map, an operator, that relates output signals to the input signals. If there is also an internal state variable description, it further specifies the operator.

Two Problems.

Having some noisy measurements it is a natural problem to try to describe the behaviour of a system, for instance by estimating the state. A good estimate should be close to the true value. It is therefore reasonable that the estimation error should have zero mean value and minimal variance. Another problem is to find inputs to drive a system back to a desired state, for instance zero, after some disturbance. For many linear systems there is an input signal, that makes the state zero in a very short time at the expense of control energy. A measure is usually defined that balances the deviation at some final time, t_1, against the required control energy. The behaviour during the trajectory may also be taken into account. The most common loss function is the quadratic form

$$J = x^T(t_1)Q_0x(t_1) + \int_{t_0}^{t_1} x^T(t)Q_1x(t)dt +$$

$$+ \int_{t_0}^{t_1} u^T(t)Q_2u(t)dt \qquad (5)$$

for continuous time and for discrete time

$$J = \sum_{t=t_0}^{t_1} x^T(t)Q_1x(t) + \sum_{t=t_0}^{t_1} u^T(t)Q_2u(t) \qquad (6)$$

where Q_0, Q_1 and Q_2 are symmetric and positive semi-definite. In order to be able to penalize infinite values of the inputs it is usually assumed that $Q_2 > 0$.

Methods for Solution.

Around 1940 Wiener [11] and Kolmogorov [9] solved these two problems using the weighting function representation. Wiener used calculus of variations. He found that the problem could be reduced to the solution of an integral equation, the Wiener-Hopf equation. Wiener solved the equation by function theoretic methods. In simple cases this reduces to the problem of factorizing a polynomial into a part with zeroes in the left half-plane and another part with zeroes in the right half-plane only (spectral factorization). Twenty years later Bellman [2] and Kalman [6] etc. applied dynamic programming and the Hamilton-Jacobi equation to the state variable representation and found a formally and computationally different approach. The problem reduced to an initial value problem for an ordinary nonlinear differential equation, the Riccati equation.

In this thesis the causal linear operator representation is used to formulate the two linear problems discussed above. The solution is obtained by factorizing certain operators in order to solve operator equations corresponding to the Wiener-Hopf equation. The state variable

description is utilized, and the Riccati equation gives the factorization directly in the time domain.

The thesis is composed of the following publications:

I. The Use of Operator Factorization for Linear Control and Estimation. Automatica 9 pp 623-31 (1973).

II. Inversion of a Dynamical System by an Operator Identity. Automatica 8 pp 361-362 (1972).

III. A New Proof and an Adjoint Filter Interpretation for Linear Discrete Time Smoothing. Report 7330, Lund Institute of Technology, Division of Automatic Control.

IV. Kalman Filters for Processes with Unknown Initial Values. Report 7332, Lund Institute of Technology, Division of Automatic Control.

V. Numerical Solution of $A^TS + SA + Q = 0$. Information Sciences 4 pp 35-50 (1972).

The first three parts are devoted to the operator approach to the control and estimation problems. Part I treats the continuous time problems, part II deals with the inversion of dynamical systems and part III demonstrates the operator technique for discrete time applied to the state estimation problem.

One problem of the linear theory that does not seem to have a satisfactory solution is estimation for processes with unknown initial state. This problem is discussed in part IV.

The application of the linear theory to any realistic

problem requires substantial numerical computation. Since the different approaches to the linear problem also lead to different computational methods it would be interesting to analyse the numerical aspects. No complete study has been made. Part V discusses a simplified problem, which contains many of the more general aspects.

Operator Notation.

Estimation and optimization problems are easy to solve for static systems. Dynamical systems can in fact also be considered as static systems in spaces of much larger dimensions. The discrete time systems on finite intervals, say $[0,t_1]$, can equivalently be analyzed using vectors like

$$\underline{x} = \begin{bmatrix} x(0) \\ x(t) \\ \vdots \\ x(t_1) \end{bmatrix}$$

The system (4) in weighting function form is:

$$\begin{cases} x(t) = \phi(t,0)x_0 + \sum_{s=0}^{t-1} \phi(t,s+1)\Gamma(s)u(s) \\ y(t) = \theta x(t) + Du(t) \end{cases}$$

It can also be written as

$$\begin{cases} \underline{x} = \underline{\mathcal{G}}x_0 + \underline{L}\underline{\Gamma}\underline{u} \\ \underline{y} = \underline{\theta}\underline{x} + \underline{D}\underline{u} \end{cases} \qquad (7)$$

using the block matrices:

$$\underline{L} = \begin{bmatrix} 0 & & \\ I & \ddots & 0 \\ \phi(t_1,1) & I & 0 \end{bmatrix} \qquad \underline{\Gamma} = \begin{bmatrix} \Gamma(0) & & 0 \\ & \ddots & \\ 0 & & \Gamma(t_1) \end{bmatrix}$$

$$\underline{\theta} = \begin{bmatrix} \theta(0) & & 0 \\ & \ddots & \\ 0 & & \theta(t_1) \end{bmatrix} \qquad \underline{g} = \begin{bmatrix} I \\ \phi(1,0) \\ \vdots \\ \phi(t_1,0) \end{bmatrix}$$

$$\underline{D} = \begin{bmatrix} D(0) & & 0 \\ & \ddots & \\ 0 & & D(t_1) \end{bmatrix}$$

The structure of the matrices should be noted. The matrix $\underline{L}$ is lower block triangular, a consequence of causality. For time constant ϕ it also has a band structure. The matrices $\underline{\Gamma}$, $\underline{\theta}$ and $\underline{D}$ are block diagonal. The signals $\underline{u}$, $\underline{x}$ and $\underline{y}$ could be considered to be functions on the discrete time interval $[0,t_1]$ instead of being vectors. The block matrices then correspond to linear operators.

Introducing appropriate block diagonal matrices $\underline{Q}_1$ and $\underline{Q}_2$ the loss function (6) can be written as

$$J = \underline{x}^T\underline{Q}_1\underline{x} + \underline{u}^T\underline{Q}_2\underline{u} \tag{8}$$

If the expression for $\underline{x}$ is inserted into (8) then

$$J = \underline{u}^T(\underline{Q}_2+\underline{\Gamma}^T\underline{L}^T\underline{Q}_1\underline{L}\underline{\Gamma})\underline{u} + 2\underline{u}^T\underline{\Gamma}^T\underline{L}^T\underline{Q}_1\underline{g}x_0 + x_0^T\underline{g}^T\underline{Q}_1\underline{g}x_0 \tag{9}$$

The dynamic optimization problem is thus rewritten as a static problem. The solution is easily obtained as

$$\underline{u} = - (\underline{Q}_2+\underline{\Gamma}^T\underline{L}^T\underline{Q}_1\underline{L}\underline{\Gamma})^{-1}\underline{\Gamma}^T\underline{L}^T\underline{Q}_1\underline{g}x_0 \tag{10}$$

provided that the inverse exists.

For the estimation problem let $\underline{u} = 0$ in (7) and add white noise terms $\underline{v}$ and $\underline{e}$ with block diagonal covariance matrices $\underline{R}_1$ and $\underline{R}_2$, $R_2 > 0$. Assume also that x_0 is a random vector with covariance matrix R_0. Thus

$$\begin{cases} \underline{x} = \underline{g}x_0 + \underline{L}\underline{v} \\ \\ \underline{y} = \underline{\theta}\underline{x} + \underline{e} \end{cases} \tag{11}$$

which is also a static problem. Assume for simplicity that all mean values are zero. The minimal variance linear estimate $\hat{\underline{x}}$ of $\underline{x}$ based on $\underline{y}$, i.e. $y(t)$, $t \in [0,t_1]$, can be expressed in terms of covariance matrices by the projection theorem

$$\hat{\underline{x}} = R_{\underline{x}\underline{y}}R_{\underline{y}}^{-1}\underline{y} = R_{\underline{x}}\underline{\theta}^T(\underline{\theta}R_{\underline{x}}\underline{\theta}^T+\underline{R}_2)^{-1}\underline{y} \tag{12}$$

provided that the matrix $R_{\underline{y}}$ is invertible.

Both optimization and estimation are thus solved as static problems in larger spaces. In fact, many problems in linear systems theory can be explained easily in this setting. Typical examples are controllability, observability and invertibility.

The continuous time problems are conceptually more difficult. Finite dimensional spaces are no longer sufficient. Vectors indexed by an interval of the real line could be used, but the function space interpretation is the most natural. The vector - matrix analogy is useful for intuition, so one can speak about triangular and diagonal operators.

The continuous time system description is introduced in part I. In analogy with (7) the system (3) is written

$$\begin{cases} \underline{x} = \underline{g}x_0 + \underline{L}\underline{B}\underline{u} \\ \underline{y} = \underline{C}\underline{x} + \underline{D}\underline{u} \end{cases}$$

It is shown that the adjoints $\underline{L}^*$ and $\underline{g}^*$ can be described by the dual system, and that the operator $\underline{L}$ is only left invertible. The analogues of (10) and (12) are also shown. Systems with nonsingular D are invertible dynamical systems, and in part II it is shown how the inverse can be obtained using operators. Let x_0 be zero, then the inverse system is

$$\underline{u} = (\underline{C}\underline{L}\underline{B}+\underline{D})^{-1}\underline{y} = (\underline{D}^{-1}-\underline{D}^{-1}\underline{C}\underline{M}\underline{B}\underline{D}^{-1})\underline{y} \qquad (13)$$

$$\underline{x} = \underline{L}\underline{B}\underline{u} = \underline{L}\underline{B}(\underline{C}\underline{L}\underline{B}+\underline{D})^{-1}\underline{y} = \underline{M}\underline{B}\underline{D}^{-1}\underline{y} \qquad (14)$$

where $\underline{M}$ is a new triangular operator. The result corresponds to an elementary matrix lemma. In parts I and III frequent use is made of this inversion lemma.

The advantage of reformulating the problems as static problems is the simple formal solutions (10) and (12). The drawback is that the recursive, dynamic nature of the solutions is lost. A fundamental idea of this thesis is to use the static formulation with its simple formal solution and then provide methods by which recursive equations for the solution can be obtained.

Solution Using Operator Factorization.

Both the estimation and the optimization problems require the solution of a linear equation (12) or (10). The usual way of solving such equations in finite spaces is by Gauss elimination, i.e. triangularization and recursive solution of a backward and a forward system.

In the discrete time the number of operations required for Gauss elimination would be of the order of $[n(t_1-t_0)]^3$. The structure of the two operators to invert in (12) and (10) for the estimation and control problems can, however, be exploited to reduce the computations. For example if $\underline{R}_1 = \underline{I}$, $\underline{R}_2 = \underline{I}$, $\underline{\Theta} = \underline{I}$, $\underline{C} = \underline{I}$, $R_0 = 0$, then $R_{\underline{y}}$ is given by

$$R_{\underline{y}} = \underline{I} + \underline{L}\underline{L}^* \tag{15}$$

both in the discrete and continuous time cases. A symmetric factorization of $R_{\underline{y}}$ into lower and upper triangular operators can be obtained using the Riccati equation. $R_{\underline{y}}$ can be written

$$R_{\underline{y}} = \underline{I} + \underline{L}\underline{L}^* = (\underline{I}+\underline{L}\underline{P})(\underline{I}+\underline{P}\underline{L}^*) \tag{16}$$

in the continuous time case and for discrete time

$$R_{\underline{y}} = \underline{I} + \underline{L}\underline{L}^* = (\underline{I}+\underline{P}+\underline{L}\underline{\phi}\underline{P})(\underline{I}+\underline{P})^{-1}(\underline{I}+\underline{P}+\underline{P}\underline{\phi}^T\underline{L}^T) \tag{17}$$

provided that $\underline{P}$ is the diagonal operator with P(t) satisfying the continuous or discrete time Riccati equation with zero initial value. These identities and their generalizations are shown in parts I and III.

The continuous time solution in part I will now be demonstrated. Using $\underline{P}$ the operator $R_{\underline{xy}}$ is

$$R_{\underline{xy}} = \underline{L}\underline{L}^* = \underline{P}\underline{L}^* + \underline{L}\underline{P}(\underline{I}+\underline{P}\underline{L}^*)$$

so that

$$R_{\underline{xy}}R_{\underline{y}}^{-1} = \underline{L}\underline{L}^*(\underline{I}+\underline{L}\underline{L}^*)^{-1} = \underline{P}\underline{L}^*(\underline{I}+\underline{P}\underline{L}^*)^{-1}(\underline{I}+\underline{L}\underline{P})^{-1} + \underline{L}\underline{P}(\underline{I}+\underline{L}\underline{P})^{-1}$$

Inversion of dynamical systems gives

$$(\underline{I}+\underline{L}\underline{P})^{-1} = \underline{I} - \underline{M}\underline{P}, \quad \underline{L}\underline{P}(\underline{I}+\underline{L}\underline{P})^{-1} = \underline{M}\underline{P}, \quad \underline{L}^*(\underline{I}+\underline{P}\underline{L}^*)^{-1} = \underline{M}^*$$

where $\underline{M}$ is a new lower triangular operator like $\underline{L}$. Thus

$$R_{\underline{xy}}R_{\underline{y}}^{-1} = (\underline{L}\underline{P}+\underline{P}\underline{M}^*)(\underline{I}-\underline{M}\underline{P}) = \underline{M}\underline{P} + \underline{P}\underline{M}^*(\underline{I}-\underline{M}\underline{P}) \qquad (18)$$

gives the so called smoothing estimate. The discrete time estimate is given in part III. Eq. (18) represents causal and anticausal integral operators. Using forward and backward initial value differential equations it can be reformulated as by Bryson and Frazier [3]. It is also interesting that the operator $\underline{M}\underline{P}$ gives the filter estimate, i.e. the estimate of x(t) using y(s) only for $s \leq t$. This corresponds to the forward equation.

Instead of a high dimensional Gauss elimination, a simpler factorization and some dynamical system inversions directly give the formulas used in practice. The matrix Riccati equation with the dimension of the state plays the central role of the factorization. It is thus found how the internal structure, which is lost in the input output formulation, can be exploited to obtain recursive algorithms.

The continuous time regulator problem is treated in part I, including also restricted end point, deterministic disturbance known in advance and disturbances described by stochastic processes.

Earlier Work.

The operator notation, where the signals are considered as time functions, has been used before, e.g. by Bellman [1] and Kailath [5], but mostly descriptively. The connection between Riccati equations and Fredholm and Wiener-Hopf integral equations was shown by the work by Kalman and Bucy [8] and has been applied primarily in the field of two point boundary value problems by Schumitsky [10] for the factorization of Fredholm resolvents. Gohberg and Krein [4] have given an important theorem on the existence and uniqueness of such factorizations also for asymmetric problems. Kailath [5] has applied the resolvent identity to signal smoothing, i.e. estimation of $(\underline{y}-\underline{e})$, in his innovations approach to the estimation problem. Using the inverse of (16) in this way there are difficulties when $C \neq I$ or $R_0 \neq 0$. Some extensions of the solutions discussed here to infinite dimensional state spaces are possible. The extension to infinite time horizon also requires some caution. Stable systems can be handled without much extra effort. The factorization (16) directly corresponds to spectral factorization of the corresponding Laplace transforms. The system I+LP is stable and has a stable inverse.

Processes With Unknown Initial Values.

In the well-established field of linear quadratic control the estimation problem for a system with unknown initial state has not been satisfactorily solved. This problem is treated in part IV. The problem is the dual of optimal control for fixed end point. Intuitively speaking a "dead beat filter" which eliminates the effect of the unknown initial value should be applied. A special case was discussed early by Kalman [7] in connection with observability. In practice one usually solves the ordinary recursive equations starting with a large covariance. This method is numerically ill-conditioned. In part IV the discrete time optimal filter is ob-

tained by letting the initial covariance go to infinity. Two Riccati equations are then obtained, one for the error covariance and one for updating the unobservable subspace. A different type of solution obtained by duality is suggested for the continuous time case.

Numerical Aspects.

The numerical aspects of the problems discussed above are important for economic realization on digital computers, and by no means fully understood. One example, treated in part V, is the problem of solving S in the simple n×n matrix equation

$$A^TS + SA + Q = 0 \qquad (19)$$

This fundamental equation is often called the Lyapunov equation, since it can be used to construct Lyapunov functions. It also arises when evaluating steady state loss functions in linear control and covariance matrices for dynamical systems. More than ten different methods have been proposed emanating both from internal and external system descriptions. Part V gives a survey and classifies them into direct methods, transformation methods and iterative methods. The nine algorithms that seem to be the best ones are coded and tested on a batch of representative A and Q matrices. The direct methods which convert (19) to a n(n+1)/2 dimensional linear system thereby loosing much structure, are good for small problems, say n<7, but unfeasible at present stage for large problems because of large core requirement and long computing time. The iteration methods are the best for larger systems. The required computing time depends also on the numerical condition of the equation. The fastest algorithm is a transformation method but the accuracy is too bad. If the A matrix is in any special form like Jordan or Companion form or if the corresponding transformations would be of any value in the fur-

ther analysis of a problem, then it should be seriously considered to use a method that takes advantage of this structure.

Acknowledgements.

I want to express my sincere gratitude to my adviser Karl Johan Åström, who proposed the operator factorization problems, and who has constantly supported me with ideas and encouragement. I would also like to thank all my collegues at the Division of Automatic Control, who form a group with an open and stimulating atmosphere. Many improvements have resulted from our frequent discussions. The typing of the manuscripts with all their formulas was done by Gudrun Christensen and Kerstin Palmqvist, and I really appreciate their work. Finally I want to thank my wife Barbro for her understanding and support.

References.

[1] Bellman R.E.: Introduction to the Mathematical Theory of Control Processes, Vol. 1, Academic Press, New York (1967).

[2] Bellman R.E.: On a Class of Variational Problems, Q. Appl. Math. 14, 353-9 (1957).

[3] Bryson A.E., Frazier M.: Smoothing for Linear and Nonlinear Dynamic Systems, Aero. Sys. Div., Wright-Patterson AFB, Ohio, Tech. Rept., ASD-TDR-63-119 (1963).

[4] Gohberg I.C., Krein M.G.: On the Factorization of Operators in Hilbert Spaces, Acta Sci. Math. Szeged. 25, 90-123 (1964).

[5] Kailath T.: Application of a Resolvent Identity to a Linear Smoothing Problem, SIAM J. Control 7, 68-74 (1969).

[6] Kalman R.E.: Contributions to the Theory of Optimal Control, Bol. Soc. Mat. Mexicana 5, 102-119 (1960).

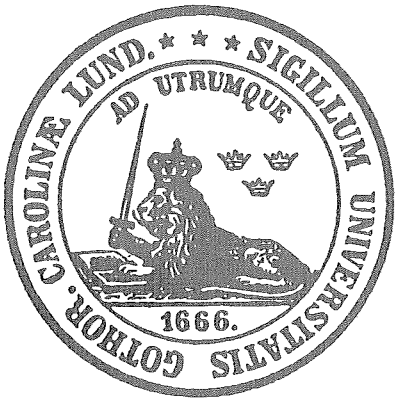
[7] Kalman R.E.: New Methods in Wiener Filtering Theory, in Proc. of 1st Symp. on Eng. Appl. of Random Function Theory and Probability, Bogdanoff J.L., Kozin F. (eds.), Wiley, New York (1963).

[8] Kalman R.E., Bucy R.S.: New Results in Linear Filtering and Prediction Theory, J. Basic Eng., Trans ASME Ser. D 83, 95-108 (1961).

[9] Kolmogorov A.N.: Interpolation and Extrapolation of Stationary Random Sequences, Bull. Moscow Univ. USSR, Ser. Math. 5 (1941).

[10] Schumitsky A.: On the Equivalence between Matrix Riccati Equations and Fredholm Resolvents, J. Comp. Syst. Sci. 2, 76-87 (1968).

[11] Wiener N.: The Extrapolation, Interpolation and Smoothing of Stationary Time Series and Engineering Applications, Wiley, New York (1942).


SIGILLUM UNIVERSITATIS GOTHOR. CAROLINÆ LUND. ★ ★ ★
AD UTRUMQUE
1666.

LTH

Automatica, Vol. 9, pp. 623–631. Pergamon Press, 1973. Printed in Great Britain.

Brief Paper

The Use of Operator Factorization for Linear Control and Estimation*

L'emploi de Factorisation d'Opérateur pour le Contrôle Linéaire et l'Estimation

Die Benutzung der Operator-Faktorisierung zur linearen Steuerung und Schätzung

Nспользование операторной факторизации для линейного управления и оценки

PER HAGANDER†

Summary—The linear filtering, prediction and smoothing problems as well as the linear quadratic control problems can very generally be formulated as operator equations using basic linear algebra.

The equations are of Fredholm type II, and they are difficult to solve directly.

It is shown how the operator can be factorized into two Volterra operators using a matrix Riccati equation. Recursive solution of these triangular operator equations is then obtained by two initial value differential equations.

The proofs of smoothing and optimal control under known disturbances are in this way especially clear and simple.

1. *Introduction*

THERE are many different ways to approach the linear estimation and control problems. In the original estimation formulation, due to Wiener, the problems were stated as minimization of quadratic functionals in the L_2-space. It was shown by Wiener that the minimization led directly to a linear integral equation for the weighting function or the transfer function of the optimal filter, the so-called Wiener–Hopf equation.

There are no intrinsic difficulties involved in extending this analysis to the time varying case. This leads to a time varying version of the Wiener–Hopf equation.

The most serious difficulty when trying to solve the Wiener–Hopf equation is that the problem is basically infinite dimensional. As has been shown by Bellman, Kalman and others the linear quadratic problems are significantly simplified for systems governed by ordinary differential equations. In such cases the difficulties can be reduced by use of initial value problems for some ordinary differential equations, Riccati equations, and the solutions to all relevant problems can be written in terms of solutions to the Riccati equation. This approach is for instance used in [1, 3, 4, 8–12, 16]. The problems associated with ordinary differential equations can of course easily be formulated as integral equations, the kernels of which can be expressed in terms of solutions to ordinary differential equations.

In this paper the integral equations, in more general form, are obtained by operator formulation in function spaces and basic linear algebra lemmas, almost the Wiener approach. The Riccati equation is then used to decompose the operators of the resulting Fredholm equations into causal and anticausal parts, so that the solutions are obtained in the usual form as differential equations. Thus the equivalence is displayed between the two approaches, and the role of the Riccati equation is emphasized. It also gives neat alternative proofs of the results in linear estimation and control including the smoothing case and optimal control under known disturbances.

The operator notation has been used for instance by BELLMAN [2] and MEE [14], and the correspondence between Riccati equations and Fredholm resolvents is explored by KALABA [7], SCHUMITZKY [15] and KAILATH [8–11], who also in [10] indicated the operator problems solved in this paper. The decomposition into causal and anticausal part has been done frequently by other methods and is in fact the spectral factorization of the Wiener theory. Kailath has labelled it the innovations approach of the estimation problems [8, 9, 11].

The disposition of the paper is such that the beginning of the second section is devoted to the operator notation and some useful results concerning these operators before the problems are formulated and written in the new notation. The operator factorization, the main result, is presented in Section 3. In the fourth section this main result is used to give the solutions of the stated problems.

2. *Definitions and formulation of the problems*

Notation, scalar products and adjoints. Linear control problems are often formulated using differential equations of first order, the state equations. The state at time t can be regarded as the value $x(t)$ of a function x on the interval $[t_0, t_1]$. Let this function be an element of the function space X.

Introduce a scalar product and a norm by

$$x \cdot y = \int_{t_o}^{t_1} x^T(t) y(t) \mathrm{d}t$$

$$\|x\|^2 = x \cdot x.$$

The solution x to the differential equation

$$\begin{cases} \dot{x} = Ax + Bu \\ x(t_o) = x_o \end{cases}$$

is

$$x(t) = \phi(t, t_o) x_o + \int_{t_o}^{t} \phi(t, s) B u(s) \mathrm{d}s$$

---

* Received 5 June 1972; Revised 5 December 1972. The original version of this paper was not presented at any IFAC meeting. It was recommended for publication in revised form by Associate Editor B. D. O. Anderson.

† The author is with the Division of Automatic Control, Lund Institute of Technology, Lund, Sweden.

that is the function x is a linear function of the vector x_o and the function u.

Define the operators g, L and B giving

$$x=gx_o+LBu. \tag{2.1}$$

Thus g is an operator from the state space R^n into X, B is an operator from a space of input functions to X and L is an operator in X.

Scalar products introduce adjoint operators, so that $x=L^*y$ means

$$x(t)=\int_t^{t_1}\phi^T(s,\,t)y(s)\mathrm{d}s.$$

Introduce also the operators $h:R^u\to X$; $x=hb$ *by*

$$x(t)=\phi^T(t_1,\,t)b$$

$$T_o\colon X\to R^n \text{ by } T_ox=x(t_o) \text{ and}$$

$$T_1\colon X\to R^n \text{ by } T_1x=x(t_1).$$

Let the following calculation serve as one example of possible operator manipulations.

Example. Express g and h in terms of L, T_o and T_1. The scalar product in R^n gives

$$g^*x\cdot a=x\cdot ga=\int_{t_o}^{t_1}x^T(t)\phi(t,\,t_o)a\mathrm{d}t=$$

$$\left(\int_{t_o}^{t_1}\phi^T(t,\,t_o)x(t)\mathrm{d}t\right)^Ta=T_oL^*x\cdot a.$$

Thus $g^*=T_oL^*$ and correspondingly $h^*=T_1L$.

Inversion. No inverse exists to the operator L. The operator $\frac{\mathrm{d}}{\mathrm{d}t}-A$ however, is a left hand inverse:

$$\left(\frac{\mathrm{d}}{\mathrm{d}t}-A\right)L=I$$

but

$$L\left(\frac{\mathrm{d}}{\mathrm{d}t}-A\right)x=x$$

is only true if $x(t_o)=0$, in fact

$$L\left(\frac{\mathrm{d}}{\mathrm{d}t}-A\right)=I-gT_o. \tag{2.2}$$

Correspondingly

$$\left(-\frac{\mathrm{d}}{\mathrm{d}t}-A^T\right)L^*=I$$

and

$$L^*\left(-\frac{\mathrm{d}}{\mathrm{d}t}-A^T\right)=I-hT_1.$$

Inversion of a dynamical system. The inverse system of

$$\begin{cases}x=Ax+Bu,\ x(t_0)=0\\ y=Cx+Du\end{cases} \tag{2.3}$$

or

$$y=(CLB+D)u$$

exists for a nonsingular D and gives according to [6]

$$u=(CLB+D)^{-1}y$$

$$=(D^{-1}-D^{-1}CMBD^{-1})y \tag{2.4}$$

where M fulfils

$$\left\{\frac{\mathrm{d}}{\mathrm{d}t}-(A-BD^{-1}C)\right\}M=I.$$

Space of stochastic processes. In order to treat the stochastic problems the space X must be extended to contain stochastic processes generated by linear Wiener process driven differentials, cf. [16], like

$$\mathrm{d}x=Ax\mathrm{d}t+Bu\mathrm{d}t+\mathrm{d}v;\ x(t_o)=x_o.$$

The operator L should then be replaced by $\dot{L}$ defined by the stochastic Ito integral

$$z=\dot{L}x;\ z(t)=\int_{t_o}^{t}\phi(t,\,s)\mathrm{d}x(s)=\int_{t_o}^{t}\phi(t,\,s)(Ax$$

$$+Bu)\mathrm{d}s+\int_{t_o}^{t}\phi(t,\,s)\mathrm{d}v(s). \tag{2.5}$$

Also other dot operators will appear in the sequel.

Note that the usual deterministic functions can be regarded as special cases of stochastic processes. This motivates the need for a pure integration operator, in the subspace of deterministic functions:

$$z=\int x;\ z(t)=\int_{t_o}^{t}x(s)\mathrm{d}s \tag{2.6}$$

with the adjoint

$$z=\int^* x;\ z(t)=\int_t^{t_1}x(s)\mathrm{d}s$$

so that

$$z=\dot{L}\int x=Lx.$$

Define also the scalar product in the space of stochastic processes

$$x\cdot y=E\int_{t_o}^{t_1}x^T(t)y(t)\mathrm{d}t \tag{2.7}$$

which is consistent with the scalar product in the deterministic subspace. Notice that the processes are not required to have zero mean function.

It is possible to show that interchanges of integration order are allowed also in the stochastic space, a fact that is used frequently.

Opitmal control problem. Reformulate the problem to minimize

$$J=\int_{t_o}^{t_1}(x^TQ_1x+u^TQ_2u)\mathrm{d}t+x^T(t_1)Q_ox(t_1) \tag{2.8}$$

under the constraint

$$\dot{x}=Ax+Bu,\ x(t_o)=x_o \tag{2.9}$$

using the introduced notation.

Define operators Q_1 and Q_2 suitably, then

$$J=x\cdot Q_1x+u\cdot Q_2u+T_1x\cdot Q_oT_1x$$
$$x=gx_o+LBu$$

or

$$J=(gx_o+LBu)\cdot Q_1(gx_o+LBu)+u\cdot Q_2u$$
$$+T_1(gx_o+LBu)\cdot Q_oT_1(gx_o+LBu)$$
$$=u\cdot(Q_2+B^T(L^*Q_1L+hQ_oh^*)B)u+$$
$$+2u\cdot B^T(L^*Q_1g+hQ_oT_1g)x_o$$
$$+x_o\cdot(T_oL^*Q_1g+T_ohQ_oT_1g)x_o$$
$$=u\cdot Pu+2u\cdot r+c.$$

The minimizing u is now easily obtained by completing squares. It is then necessary to solve

$$Pu=-r \tag{2.10}$$

where

$$P=Q_2+B^T(L^*Q_1L+hQ_oh^*)B.$$

Equation (2.10) corresponds to a Fredholm II integral equation.

Using the technique of Section 3, P can be factorized into a causal and an anticausal part, and (2.10) can be solved recursively by differential equations.

Linear estimation. Also the estimation problems of linear systems can be handled with the operator technique.

Consider

$$\begin{cases} \mathrm{d}x=Ax\mathrm{d}t+\mathrm{d}v,\ x(t_o)=x_o \\ \mathrm{d}y=Cx\mathrm{d}t+\mathrm{d}e,\ y(t_o)=0 \end{cases} \tag{2.11}$$

where x_o has zero mean value, covariance R_o, and where v and e are independent, zero mean, Wiener processes, independent of x_o, with incremental covariance $R_1\mathrm{d}t$ and $R_2\mathrm{d}t$ respectively. It is also assumed that R_2 is nonsingular. Rewrite

$$x=\dot{L}v+gx_o \tag{2.12}$$

according to (2.5).

Now find the best linear estimate $\hat{x}$ of x in the minimum variance sense. That is, find a linear operator $\dot{K}$ with

$$\hat{x}=\dot{K}y \tag{2.13}$$

meaning

$$\hat{x}(t)=\int_{t_o}^{t_1}k(t,\ s)\mathrm{d}y(s)=\int_{t_o}^{t_1}k(t,\ s)Cx(s)\mathrm{d}s$$
$$+\int_{t_o}^{t_1}k(t,\ s)\mathrm{d}e(s)$$

such that the variance of $\tilde{x}_i(t)=x_i(t)-\hat{x}_i(t)$ is minimized for all i and t.

Notice that this is the smoothing problem. The information available to form $\hat{x}$ is $\mathrm{d}y$ (or y) during the interval $[t_o,\ t_1]$. This includes the filter problem as a special case. Prediction can be handled with only minor extensions in the following.

Introduce a Hilbert space of one dimensional, zero mean stochastic variables with the scalar product

$$<\xi,\ \eta>=\mathrm{cov}(\xi,\ \eta)$$

and apply the projection theorem, e.g. [13, p. 51], for each component $x_i(t)$ at all times t. Observe that they have zero mean. Thus there exists a unique best estimator $\hat{x}_i(t)$ in the closure of the linear subspace spanned by $y_j(s)$, each component at each time regarded as a one dimensional stochastic variable. Moreover,

$$<\tilde{x}_i(t),\ y_j(s)>=0 \quad \forall t,\ s,\ i,\ j.$$

Now assume that $\hat{x}$ really belongs to the subspace, that is can be generated by a $\dot{K}$, a fact that will be verified in Section 4. Then the orthogonality condition results in the Fredholm equation corresponding to the Wiener–Hopf equation of the filtering case,

$$r_{xy}(t,\ s)=\int_{\tau=t_o}^{t_1}k(t,\ \tau)\mathrm{d}r_y(\tau,\ s) \quad \forall t,\ s$$

or

$$R_{xy}=\dot{K}R_y. \tag{2.14}$$

The covariance operators for the system (2.12) are

$$R_{xy}=\{LR_1L^*+gR_og^*\}C^T\int^*$$
$$R_y=\int C\{LR_1L^*+gR_og^*\}C^T\int^*+\int R_2\int^*$$

with $\int$ defined by (2·6). (2.6)

Thus from (2.14)

$$\{LR_1L^*+gR_og^*\}C^T=K[C\{LR_1L^*$$
$$+gR_og^*\}C^T+R_2] \tag{2.15}$$

with a structure similar to (2.10).

Separation. Regard a linear stochastic system

$$\begin{cases} \mathrm{d}x=Ax\mathrm{d}t+Bu\mathrm{d}t+\mathrm{d}v,\ x(t_o)=x_o \\ \mathrm{d}y=Cx\mathrm{d}t+\mathrm{d}e,\qquad\quad y(t_o)=0 \end{cases}$$

where v and e are defined as in (2.11). x_o has mean value m and covariance R_o.

Rewritten in operator notation this gives

$$\begin{cases} x=LBu+gx_o+\dot{L}v \\ y=\int C[LBu+gx_o+\dot{L}v]+e. \end{cases} \tag{2.16}$$

Now define the loss to be minimized under the constraint (2.16)

$$J=E\int_{t_o}^{t_1}[x^T(t)Q_1x(t)+u^T(t)Q_2u(t)]\mathrm{d}t$$
$$+Ex^T(t_1)Q_ox(t_1)$$

or with scalar products between stochastic processes as (2.7)

$$J=x\cdot Q_1x+u\cdot Q_2u+T_1x\cdot Q_oT_1x.$$

Rewrite using (2.16) and adjoint operators

$$J=u\cdot[Q_2+B^T(L^*Q_1L+hQ_oh^*)B]u$$
$$+2u\cdot B^T[(L^*Q_1g+hQ_oT_1g)x_o$$
$$+(L^*Q_1\dot{L}+hQ_oT_1\dot{L})v]$$
$$+\dot{L}v\cdot Q_1\dot{L}v+T_1\dot{L}v\cdot Q_oT_1\dot{L}v$$
$$+2x_o\cdot[g^*Q_1\dot{L}+T_ohQ_oT_1\dot{L}]v$$
$$+x_o\cdot[g^*Q_1g+T_ohQ_oT_1g]x_o. \quad (2.17)$$

In order to minimize with respect to u, integral equations of the same kind as (2.10) and (2.15) have to be solved.

3. *Main result, factorization using the Riccati equation*

The operators in (2.10), (2.15) and (2.17), which have to be inverted, are of the same structure. When neglecting boundary values, R_o or Q_o, they consist of the sum of one "diagonal" operator and the product of a "triangular", Volterra, operator and its adjoint.

Consider the simplified equation

$$(I+LL^*)x=y. \quad (3.1)$$

In analogy with the decomposition idea of linear algebraic equations (3.1) could be solved by recursions, differential equations, if $(I+LL^*)$ were rewritten as a product of a "triangular" operator and its adjoint. For instance, try to find a "diagonal" operator P, corresponding to mutliplication with a symmetric matrix $P(t)$, such that

$$I+LL^*=(I+LP)(I+LP)^* \quad (3.2)$$

or

$$LL^*=LP+PL^*+LPPL^* \quad (3.3)$$

and using

$$L\left(\frac{\mathrm{d}}{\mathrm{d}t}-A\right)=I-gT_o$$

and

$$\left(-\frac{\mathrm{d}}{\mathrm{d}t}-A^T\right)L^*=I$$

$$0=L\left[P\left(-\frac{\mathrm{d}}{\mathrm{d}t}-A^T\right)+\left(\frac{\mathrm{d}}{\mathrm{d}t}-A\right)P+PP-I\right]L^*+gT_oPL^*.$$

This is certainly true for P such that

$$0=\frac{\mathrm{d}}{\mathrm{d}t}P-P\frac{\mathrm{d}}{\mathrm{d}t}-PA^T-AP+PP-I \quad (3.4)$$

and

$$0=T_oP. \quad (3.5)$$

Since

$$\frac{\mathrm{d}}{\mathrm{d}t}P(t)x(t)=\dot{P}(t)x(t)+P(t)\frac{\mathrm{d}}{\mathrm{d}t}x(t)$$

an operator P fulfilling (3.4) and (3.5) must have $P(t)$ as the solution to the matrix Riccati equation

$$\dot{P}(t)=AP(t)+P(t)A^T+I-P(t)P(t)$$
$$P(t_o)=0. \quad (3.6)$$

This can be summarized into the main theorem:

Theorem. The operator $I+LL^*$ operating in a space of functions on $[t_o, t_1]$ with L defined by (2.1) *can be factorized into*

$$I+LL^*=(I+LP)(I+LP)^*$$

where the operator P means multiplication with the symmetric solution $P(t)$ of the matrix Riccati equation (3.6).□

The main theorem can now be generalized, giving the factorizations shown in Table 1 together with their associated, Riccati equations. First boundary values, I, are introduced then R_2, R_1 and R_o, or Q_2, Q_1 and Q_o, and finally the rectangular matrix C, or B. This gives formulas (7) and (8), that solve the estimation and control problems, as will be shown in the next section.

TABLE 1. OPERATOR FACTORIZATION USING RICCATI EQUATIONS

(1)	$I+LL^*=(I+LP)(I+LP)^*$	$\dot{P}=AP+PA^T+I-PP$	$P(t_o)=0$
(2)	$I+L^*L=(I+SL)^*(I+SL)$	$-\dot{S}=A^TS+SA+I-SS$	$S(t_1)=0$
(3)	$I+LL^*+gg^*=(I+LP)(I+LP)^*$	$\dot{P}=AP+PA^T+I-PP$	$P(t_o)=I$
(4)	$I+L^*L+hh^*=(I+SL)^*(I+SL)$	$-\dot{S}=A^TS+SA+I-SS$	$S(t_1)=I$
(5)	$R_2+LR_1L^*+gR_og^*=(R_2+LP)R_2^{-1}(R_2+LP)^*$	$\dot{P}=AP+PA^T+R_1-PR_2^{-1}P$	$P(t_o)=R_o$
(6)	$Q_2+L^*Q_1L+hQ_oh^*=(Q_2+SL)^*Q_2^{-1}(Q_2+SL)$	$-\dot{S}=A^TS+SA+Q_1-SQ_2^{-1}S$	$S(t_1)=Q_o$
(7)	$R_2+C(LR_1L^*+gR_og^*)C^T=(R_2+CLPC^T)R_2^{-1}(R_2+CLPC^T)^*$	$\dot{P}=AP+PA^T+R_1-PC^TR_2^{-1}CP$	$P(t_o)=R_o$
(8)	$Q_2+B^T(L^*Q_1L+hQ_oh^*)B=(Q_2+B^TSLB)^*Q_2^{-1}(Q_2+B^TSLB)$	$-\dot{S}=A^TS+SA+Q_1-SBQ_2^{-1}B^TS$	$S(t_1)=Q_o$
(9)	$M_2+L_1M_1L_2^*=(M_2+L_1R)M_2^{-1}(M_2+RL_2^*)$	$\dot{R}=A_1R+RA^T{}_2+M_1-RM_2^{-1}R$	$R(t_o)=0$

The solutions to the Riccati equations (1)–(8) exist uniquely as long as $R_1 \leq 0$ and $R_2 > 0$, or $Q_1 \leq 0$ $Q_2 > 0$, and with suitable restrictions on the time variation of the matrices A, C, R_1 and R_2, or the corresponding B, Q_1, Q_2. The asymmetric equation (9) is still a further generalization, but the solution R of the Riccati equation might diverge in finite time, so the factorization is only possible for time intervals where the solution exists.

For discrete time systems exactly the same program might be performed.

Let L be defined by $x = Lz$

$$x(t) = \sum_{s=0}^{t-1} \phi(t, s+1) z(s)$$

and define suitable scalar products.

Then the main theorem corresponds to

$$(I + LL^*) = (I + L\phi P)(P + I)^{-1}(I + L\phi P)^*$$

with $P(t)$ from

$$P(t+1) = \phi(t+1, t) P(t) \phi^T(t+1, t) + I$$
$$- \phi(t+1, t) P(t) [P(t) + I]^{-1} P(t) \phi^T(t+1, t)$$
$$P(t_0) = 0.$$

The generalizations are done correspondingly.

It is also interesting to note the connections with spectral factorization and transfer functions. In the frequency domain L corresponds to $G(s) = [sI - A]^{-1}$, L^* to $G^T(-s) = [-sI - A^T]^{-1}$, and the operations are just multiplication of transforms.

Spectral factorization yields

$$[I + G(s) G^T(-s)] = [I + G(s) G^T(-s)]^+ [I + G(s) G^T(-s)]^-$$

where the causal operator $[I + G(s) G^T(-s)]^+$ corresponds to $[I + LP]$ in the time domain.

For discrete time systems the z-transform gives the same analogy.

4. *Alternative proofs for the linear estimation and control problems*

The solutions to the problems formulated in Section 2 are now easy consequences of the factorization theorem.

Optimization.

Corollary 1. *The criterion* (2.8) *is minimized under the constraints* (2.9) *by the control*

$$u = -Q_2^{-1} B^T S x \tag{4.1}$$

where $S(t)$ *is the solution to equation* (8) *of Table* 1, *giving the closed loop performance*

$$\left(\frac{\mathrm{d}}{\mathrm{d}t} - A + BQ_2^{-1} B^T S \right) x = 0 \tag{4.2}$$

$$T_0 x = x_0$$

and the loss

$$J = x_0 \cdot S(t_0) x_0.$$

Proof. Regard equation (2.10). The factorization 8 of Table 1 gives

$$P = Q_2 + B^T (LQ_1 L^* + hQ_0 h^*) B$$
$$= (Q_2 + B^T SLB)^* Q_2^{-1} (Q_2 + B^T SLB)$$
$$r = B^T [L^* Q_1 g + hQ_0 T_1 g] x_0$$
$$= (Q_2 + B^T SLB)^* Q_2^{-1} B^T S g x_0$$

and

$$T_0 [L^* Q_1 g + hQ_0 T_1 g] = T_0 (Sg + L^* SBQ_2^{-1} B^T Sg).$$

Both the dynamic system $(Q_2 + B^T SLB)$ and its adjoint are invertible since Q_2 is regular. Thus

$$P^{-1} r = (Q_2 + B^T SLB)^{-1} B^T S g x_0$$

and

$$J = (u + P^{-1} r) \cdot P (u + P^{-1} r) - r \cdot P^{-1} r + c$$
$$= [(Q_2 + B^T SLB) u + B^T S g x_0] \cdot Q_2^{-1} [(Q_2 + B^T SLB) u + B^T S g x_0] - B^T S g x_0 \cdot Q_2^{-1} B^T S g x_0 + x_0 T_0 (Sg + L^* SBQ_2^{-1} B^T Sg) x_0$$

or

$$J = (Q_2 u + B^T Sx) \cdot Q_2^{-1} (Q_2 u + B^T Sx) + x_0 \cdot T_0 S g x_0 = (u + Q_2^{-1} B^T Sx) \cdot Q_2 (u + Q_2^{-1} B^T Sx) + x_0 \cdot S(t_0) x_0$$

J is certainly minimized for

$$u + Q_2^{-1} B^T Sx = 0. \quad \square$$

Now it is also simple to minimize J under the restriction that linear combinations of the final state are fixed:

$$Nx(t_1) = b. \tag{4.3}$$

Introduce u_f and x_f, the solution of the free end point problem

$$u_f = -P^{-1} r = -(Q_2 + B^T SLB)^{-1} B^T S g x_0$$
$$= -Q_2^{-1} B^T S x_f$$

and x_f satisfies the homogenous equation (4.2).

Define v by

$$v = u + Q_2^{-1} B^T Sx = u - u_f + Q_2^{-1} B^T S (x - x_f)$$
$$= (I + Q_2^{-1} B^T SLB)(u - u_f)$$

and thus

$$x = LB(u - u_f) + x_f = LB(I + Q_2^{-1} B^T SLB)^{-1} v + x_f = \left(\frac{\mathrm{d}}{\mathrm{d}t} - A + BQ_2^{-1} B^T S \right)^{-1} Bv + x_f.$$

Let

$$U=NT_1\left(\frac{\mathrm{d}}{\mathrm{d}t}-A+BQ_2^{-1}B^TS\right)^{-1}B$$

and

$$y=b-NT_1x_f. \tag{4.4}$$

Then the constraint (4.3) can be expressed by

$$Uv=y$$

and the loss (2.8) is

$$J=v\cdot Q_2v+x_o\cdot S(t_o)x_o.$$

When $UQ_2^{-1}U^*$ is invertible, a well-known lemma [3] says that the minimal J

$$J=y\cdot(UQ_2^{-1}U^*)^{-1}y+x_o\cdot S(t_o)x_o$$

is obtained for

$$v=Q_2^{-1}U^*(UQ_2^{-1}U^*)^{-1}y.$$

Introduce the fundamental matrix $\Psi(t, s)$ for (4.2) then

$$W=UQ_2^{-1}U^*=N\int_{t_o}^{t_1}\Psi(t_1, s)BQ_2^{-1}B^T\Psi^T(t_1,s)\,\mathrm{d}sN^T, \tag{4.5}$$

where the integral is the controllability Gramian of the closed loop system (4.2),
and

$$v(t)=Q_2^{-1}B^T\Psi^T(t_1, t)N^TW^{-1}y.$$

Summarize this in

Corollary 2. *The criterion* (2.8) *is minimized under the constraints* (2.9) *and* (4.3) *by the control*

$$u=-Q_2^{-1}B^TSx+Q_2^{-1}B^T\xi$$

where ξ *is given by*

$$\begin{cases}\left(-\dfrac{\mathrm{d}}{\mathrm{d}t}-A^T+SBQ_2^{-1}B^T\right)\xi=0\\ T_1\xi=N^TW^{-1}y\end{cases} \tag{4.6}$$

and y *and* W *defined by* (4.4) *and* (4.5). *The closed loop performance is described by*

$$\begin{cases}\left(\dfrac{\mathrm{d}}{\mathrm{d}t}-A+BQ_2^{-1}B^TS\right)x=BQ_2^{-1}B^T\xi\\ T_ox=x_o\end{cases} \tag{4.7}$$

and the loss is

$$J=x_o\cdot S(t_o)x_o+y\cdot W^{-1}y.\square$$

Remark. It is interesting to notice how the solution consists of a factorization into one causal part (4.7) and one anticausal part (4.6). The latter vanishes when there are no restrictions on the final state.

Another simple extension of Corollary 1 is to let the system (2.9) be corrupted by a known deterministic disturbance ω so that [1]

$$\dot{x}=Ax+Bu+\omega,\ x(t_o)=x_o \tag{4.8}$$

or

$$x=gx_o+LBu+L\omega.$$

Thus the criterion (2.8) gives

$$\begin{aligned}J=&u\cdot(Q_2+B^T(L^*Q_1L+hQ_oh^*)B)u\\&+2u\cdot B^T\{(L^*Q_1g+hQ_oT_1g)x_o+(L^*Q_1L\\&+hQ_oh^*)\omega\}+\omega\cdot(L^*Q_1L+hQ_oh^*)\omega\\&+2x_o\cdot(T_oL^*Q_1L+T_ohQ_oh^*)\omega+x_o\cdot(T_oI^*Q_1g\\&+T_ohQ_oT_1g)x_o=u\cdot Pu+2u\cdot q+d.\end{aligned} \tag{4.9}$$

Factorize P and q using equation (8) of Table 1,

$$\begin{aligned}P=&Q_2+B^T(L^*Q_1L+hQ_oh^*)B=(Q_2\\&+B^TSLB)^*Q_2^{-1}(Q_2+B^TSLB)\\q=&B^T\{(L^*Q_1g\\&+hQ_oT_1g)x_o+(L^*Q_1L+hQ_oh^*)\omega\}=(Q_2\\&+B^TSLB)^*Q_2^{-1}B^TS(L\omega+gx_o)+B^TL^*S\omega\end{aligned}$$

and using that P is composed of two invertible dynamical systems and thus invertible

$$J=(u+P^{-1}q)\cdot P(u+P^{-1}q)-q\cdot P^{-1}q+d$$

which is certainly minimized with respect to u for

$$\begin{aligned}u=-p^{-1}q=&-(Q_2+B^TSLB)^{-1}\{B^TS(L\omega+gx_o)\\&+Q_2(Q_2+B^TL^*SB)^{-1}B^TL^*S\omega\}.\end{aligned}$$

But using an inversion lemma similar to (2.4)

$$(Q_2+B^TL^*SB)^{-1}B^TL^*=Q_2^{-1}B^T\left(-\frac{\mathrm{d}}{\mathrm{d}t}-A^T+SBQ_2^{-1}B^T\right)^{-1}.$$

Introduce η by

$$\left(-\frac{\mathrm{d}}{\mathrm{d}t}-A^T+SBQ_2^{-1}B^T\right)\eta=S\omega,\quad \eta(t_1)=0 \tag{4.10}$$

then

$$(Q_2+B^TSLB)u=-B^TS\,(^TL\omega+gx_o)+B^T\eta$$

or using (4.8)

$$u=-Q_2^{-1}B^T(Sx+\eta).$$

This can now be summarized in

Corollary 3. *The criterion* (2.8) *is minimized under the constraint* (4.8) *by the control*

$$u=-Q_2^{-1}B^T(Sx+\eta)$$

where n *is defined by* (4.10).
The closed loop performance is described by

$$\left(\frac{\mathrm{d}}{\mathrm{d}t}-A+BQ_2^{-1}B^TS\right)x=\omega-BQ_2^{-1}B^T\eta \quad (4.11)$$

$$T_0x=x_0.\square$$

Remark. Notice how the solution consists of a causal part (4.11) and an anticausal part (4.10). The latter vanished in Corollary 1. The control obviously contains feed forward.

Estimation. The smoothing estimate of (2.11) given by (2.13) and (2.15), is solved using equation (7) of Table 1.

Corollary 4. *The smoothing estimate* $\hat{x}(t|t_1)$ *for the system* (2.11) *is obtained by first calculating the filtering estimate* $\hat{x}(t|t)$ *from the information available at time* t

$$\mathrm{d}\hat{x}(t|t)=(A-PC^TR_2^{-1}C)\hat{x}(t|t)\mathrm{d}t+PC^TR_2^{-1}\mathrm{d}y(t)$$

$$\hat{x}(t_0|t_0)=0. \quad (4.12)$$

The innovations $\mathrm{d}\tilde{y}(t)=\mathrm{d}y(t)-Cx\,(t|t)\mathrm{d}t$ *during* $[t, t_1]$ *are then calculated by integrating* (4.12) *up to* t_1.
The adjoint equation

$$\begin{cases} -\mathrm{d}\lambda(t)=(A-PC^TR_2^{-1}C)^T\lambda(t)\mathrm{d}t \\ \qquad\qquad +C^TR_2^{-1}\mathrm{d}\tilde{y}(t) \\ \lambda(t_1)=0 \end{cases} \quad (4.13)$$

is solved from t_1 *and backwards and finally*

$$\hat{x}(t|t_1)=\hat{x}(t|t)+P(t)\lambda(t). \quad (4.14)$$

Proof. Rewrite (2.15) using equation (7) of Table 1.

$$C(LR_1L^*+gR_0g^*)C^T+R_2=(R_2+CLPC^T)R_2^{-1}(R_2+CLPC^T)^*$$

$$(LR_1L^*+gR_0g^*)C^T=PL^*C^T+LPC^TR_2^{-1}(R_2+CLPC^T)^*.$$

Notice that both (R_2+CLPC^T) and the adjoint system are invertible since R_2 is invertible! Thus

$$K=[LPC^T+PL^*C^T(R_2+CPL^*C^T)^{-1}R_2](R_2+CLPC^T)^{-1}$$

Introduce

$$K_f=LPC^T(R_2+CLPC^T)^{-1}=\left(\frac{\mathrm{d}}{\mathrm{d}t}-A+PC^TR_2^{-1}C\right)^{-1}PC^TR_2^{-1} \quad (4.15)$$

where the last equality is obtained by an inversion lemma similar to (2.4). Notice that $T_1L^*=0$ giving $T_1K_f=T_1K$, and $\hat{x}_f(t_1)$ defined by the stochastic integral

$$\hat{x}_f=\dot{K}_fy$$

is thus the filter estimate of $x(t_1)$.

Similar to (4.15) and (2.4)

$$L^*C^T(R_2+CPL^*C^T)^{-1}R_2=\left[-\frac{\mathrm{d}}{\mathrm{d}t}-A^T+C^TR_2^{-1}CP\right]^{-1}C^T$$

and

$$(R_2+CLPC^T)^{-1}=R_2^{-1}-R_2^{-1}C\left(\frac{\mathrm{d}}{\mathrm{d}t}-A+PC^TR_2^{-1}C\right)^{-1}PC^TR_2^{-1}$$

giving

$$K=K_f+P\left[-\frac{\mathrm{d}}{\mathrm{d}t}-A^T+C^TR_2^{-1}CP\right]^{-1}C^TR_2^{-1}[I-CK_f].$$

Finally define λ so that

$$\hat{x}=\dot{K}y=\dot{K}_fy+P\lambda=\hat{x}_f+P\lambda$$

which is fulfilled for λ according to the adjoint dynamics (4.13).$\square$

Remark. Notice how the smoothing estimate consists of a factorization into one causal part (4.12) and one anticausal part (4.13). The latter vanishes in the filter case.

Notice also that the estimation problems for the innovations representation of (2.11), cf. [8, 9, 11, 14],

$$\begin{cases} \mathrm{d}\hat{x}_f=A\hat{x}_f\mathrm{d}t+PC^TR_2^{-1}\mathrm{d}\tilde{y} \\ \hat{x}_f(t_0)=0 \end{cases}$$

$$\mathrm{d}y=C\hat{x}_f\mathrm{d}t+\mathrm{d}\tilde{y}$$

is already factorized. Furthermore the smoothing and filtering estimates are equal. There is no randomness in the initial condition. If there were, a full factorization would be necessary.

Separation. Minimization of the functional (2.17) very much resembles the minimization of (4.9) in Corollary 3, but instead of the deterministic component $L\omega$, (2.17) contains the stochastic term $v\dot{L}$.

$$\begin{aligned} J=&u\cdot[Q_2+B^T(L^*Q_1L+hQ_0h^*)B]u \\ &+2u\cdot B^T(L^*Q_1g+hQ_0T_1g)x_0 \\ &+2u\cdot B^T(L^*Q_1\dot{L}+hQ_0T_1\dot{L})v \\ &+\dot{L}v\cdot Q_1\dot{L}v+T_1\dot{L}v\cdot Q_0T_1\dot{L}v \\ &+2x_0\cdot[g^*Q_1\dot{L}+T_0hQ_0T_1\dot{L}]v \\ &+x_0\cdot[g^*Q_1g+T_0hQ_0T_1g]x_0 \\ &=u\cdot Pu+2u\cdot q_1+2u\cdot q_2+d_1. \end{aligned}$$

Factorize as before using equation (8) of Table 1

$$P=(Q_2+B^TSLB)^*Q_2^{-1}(Q_2+B^TSLB)$$

$$q_1=(Q_2+B^TSLB)^*Q_2^{-1}B^TSgx_0$$

$$q_2=B^T\int_t^{t_1}\phi^T(s,t)S(s)\mathrm{d}v(s)+(Q_2+B^TSLB)^*Q_2^{-1}B^TS\dot{L}v.$$

Thus since P is invertible

$$J=[u+P^{-1}(Q_2+B^TSLB)^*Q_2^{-1}B^TS(gx_o+\dot{L}v)]\cdot P[u+P^{-1}(Q_2+B^TSLB)Q_2^{-1}B^TS(gx_o+\dot{L}v)]-(Q_2+B^TSLB)^*Q_2^{-1}B^TS(gx_o+\dot{L}v)\cdot P^{-1}(Q_2+B^TSLB)^*Q_2^{-1}B^TS(gx_o+\dot{L}v)+d_1+E\int_{t_o}^{t_1}u^T(t)\left[B^T\int_t^{t_1}\phi^T(s,t)S\mathrm{d}v(s)\right]\mathrm{d}t. \quad (4.16)$$

Now assume that an admissible $u(t)$ is a function of the state $x(t)$, possibly of old values, not of future values to any extent. Consequently there may be no dependence between an admissible $u(t)$ and the increment $\mathrm{d}v(s)$, $s>t$ Thus the last term of (4.16) vanishes for all admissible u. Ito integrals are being used, and the increments $\mathrm{d}v(t)$ have zero mean. Now insert the expression for P

$$J=[Q_2u+B^TSLBu+B^TS(gx_o+\dot{L}v)]\cdot Q_2^{-1}[Q_2u+B^TSLBu+B^TS(gx_o+\dot{L}v)]-B^TS(gx_o+\dot{L}v)\cdot Q_2^{-1}B^TS(gx_o+\dot{L}v)+d_1. \quad (4.17)$$

Corollary 5. (*Complete state information.*) *Let the admissible controls, $u(t)$, be functions of the state up to time t. The functional* (2.17) *is minimized for the system* (2.16) *by the control*

$$u=-Q_2^{-1}B^TSx. \quad (4.18)$$

Proof. Since (4.18) is an admissible control (4.17) is certainly minimized for this control. □

To obtain the minimal loss d_1 should be decomposed. Strictly done without introducing white noise this is rather technical in the introduced notation. Then cf. [16]

$$J_1=m^TS(t_o)m+trS(t_o)R_o+\int_{t_o}^{t_1}trS(t)R_1(t)\mathrm{d}t \quad (4.19)$$

which is larger than the loss that would be obtained with feed forward if the disturbance was known in advance. A combination of Corollaries 3 and 5 is useful to handle noise with a mean value different from zero.

If complete state information is not available, (4.18) is not an admissible control.

Corollary 6. (*Incomplete state information.*) *Let admissible controls $u(t)$ be linear functions of the output up to time t. Then* (2.17) *is minimized for the system* (2.16) *by the control*

$$u=-Q_2^{-1}B^TS\hat{x}_f$$

where f is the filter estimate of x (cf. Corollary 4).

Proof. The expression (4.17) is still valid since $u(t)$ and $\mathrm{d}v(s)$, $s\leq t$ are independent and $\mathrm{d}v$ zero mean. Rewrite (4.17) using

$$x=\hat{x}_f+\tilde{x}_f.$$

Then

$$J=(u+Q_2^{-1}B^TS\hat{x}_f)\cdot Q_2(u+Q_2^{-1}B^TS\hat{x}_f)+2u\cdot B^TS\tilde{x}_f+B^TS\tilde{x}_f\cdot Q_2^{-1}B^TS\tilde{x}_f+d_1-B^TS(gx_o+\dot{L}v)\cdot Q_2^{-1}B^TS(gx_o+\dot{L}v). \quad (4.20)$$

But $\tilde{x}_f(t)$ has zero mean value and is independent of y up to time t and consequently of an admissible $u(t)$

$$u\cdot B^TS\tilde{x}_f=0$$

J is then certainly minimized* for the control (4.19). □

The minimal loss can be expressed using J_1 from (4.19)

$$J_2=B^TS\tilde{x}_f\cdot Q_2^{-1}B^TS\tilde{x}_f+J_1=\int_{t_o}^{t_1}trPSBQ_2^{-1}B^TS\mathrm{d}t+J_1 \quad (4.21)$$

where the last equality is obtained using

$$P(t)=E\tilde{x}_f(t)\tilde{x}_f^T(t).$$

5. *Conclusions*

The linear estimation and linear quadratic control problems are very generally formulated using operator notation. Solutions can be obtained using completion of squares and the projection theorem, resulting in Fredholm integral equations, and it is shown for the finite order system case how the Riccati equation decomposes the equation into two Volterra equations.

The well known solution of the problems are obtained as the recursive solution of these initial value differential equations.

The proofs of smoothing and optimal control under known disturbances are in this way especially clear and simple.

References

[1] M. Athans and P. L. Falb: *Optimal Control.* McGraw-Hill, New York (1966).

[2] R. Bellman: *Introduction to the Mathematical Theory of Control Processes*, Vol. 1. Academic Press, New York (1967).

[3] R. W. Brockett: *Finite Dimensional Linear Systems.* Wiley, New York (1970).

[4] A. E. Bryson and Y.-C. Ho: *Applied Optimal Control.* Ginn, London (1969).

[5] P. Hagander: Linear Control and Estimation Using Operator Factorization, Report 7114. Division of Automatic Control, Lund Institute of Technology, Lund, Sweden (1971).

[6] P. Hagander: Inversion of a dynamical system by an operator identity. *Automatica* **8,** 361–362 (1972).

[7] H. H. Kagiwada, R. E. Kalaba and B. J. Vereeke: The invariant imbedding numerical method for Fredholm integral equations with degenerated kernels. *J. Approx. Theory* **1,** 355–364 (1968).

[8] T. Kailath: An innovations approach to least-squares estimation. Part I: Linear filtering in additive white noise. *IEEE* **AC-13,** 646–655 (1968).

[9] T. Kailath and P. Frost: An innovations approach to least-squares estimation. Part II: Linear smoothing in additive white noise. *IEEE* **AC-13,** 655–660 (1968).

[10] T. Kailath: Application of a resolvent identity to a linear smoothing problem. *SIAM J. Control* **7,** 68–74 (1969).

[11] T. Kailath and A. Geesey: An innovations approach to least squares estimation. Part IV: Recursive estimation given lumped covariance functions. *IEEE* **AC-16,** 720–727 (1971).

* The information obtainable does not change with the input because of the restriction to linear feedback.

[12] R. KALMAN and R. S. BUCY: New results in linear filtering and prediction theory. *J. Basic Eng., Trans. ASME., Series D* **83,** 95–108 (1961).
[13] D. G. LUENBERGER: *Optimization by Vector Space Methods.* Wiley, New York (1969).
[14] D. H. MEE: Optimal feedback gains for the linear system—quadratic cost problem. *Int. J. Control* **13,** 179–187 (1971).
[15] A. SCHUMITSZKY: On the equivalence between matrix Riccati equations and Fredholm resolvents. *J. Comp. Syst. Sci.* **2,** 76–87 (1968).
[16] K. J. ÅSTRÖM: *Introduction to Stochastic Control Theory.* Academic Press, New York (1970).

Résumé—Les problèmes de filtration linéaire, de prédiction et d'adoucissement ainsi que ceux de contrôle linéaire quadratique peuvent généralement être formulés comme équations d'opérateur utilisant l'algèbre linéaire fondamentale.

Les équations sont du type Fredholm II et sont difficiles à résoudre directement.

Il est montré comment l'operatuer peut être factorisé en deux opérateurs Volterra utilisant une équation de matrice Riccati. On obtient ensuite la solution récursive de ces équations d'opératuer triangulaires par deux équations différentielles à valeur initiale.

Les prueves d'adoucissement et de contrôle optimal dans des conditions de dérangements connus sont ainsi spécialement claires et simples.

Zusammenfassung—Die Probleme der linearen Filterung, der Vorhersage und der Glättung, ebenso wie die der linear-quadratischen Steuerung können unter Benutzung der Grundlagen der linearen Algebra sehr allg—ein als Operatorgleichungen formuliert werden.

Die Gleichungen sind vom Fredholmschen Typ II und schwierig zu lösen.

Gezeigt wird, wie der Operator in zwei Volterra-Operatoren faktorisiert werden kann, wobei eine Riccatische Matrix-Gleichung benutzt wird. Eine rekursive Lösung dieser Dreiecks-Operatorgleichungen wird dann durch zwei Anfangswert-Differentialgleichungen erhalten.

Die Prüfung von Glättung und optimaler Steuerung bei bekannten Störungen ist auf diese Art besonders klar und einfach.

Резюме—Проблемы линейной фильтрации, предсказания, сглаживания, как и проблемы линейного квадратичного управления, можно в достаточно общем виде описать операторными уравнениями с использованием основных законов линейной алгебры.

Уравнения относятся ко 2-ому типу уравнений фридхолма и весьма трудны для непосредственного решения.

Показано как оператор может быть с помощью матричного уравнения Рикатти, факторизован в два оператора Вольтера.

Рекурсивное решение этого треугольного операторного уравнения получается затем с помощью двух дифференциальных уравнений начального порядка. Таким путем весьма просты проверки сглаживания и оптимальность управления при известных возмущениях.

Automatica, Vol. 8, pp. 361–362. Pergamon Press, 1972. Printed in Great Britain.

Correspondence Item

Inversion of a Dynamical System by an Operator Identity*

Inversion d'un système dynamique par une identité d'opérateurs

Inversion eines dynamischen Systems durch eine Operatoridentität

Инверсия динамической системы с помощью операторной идентичности

PER HAGANDER†

Summary—Inversion of a linear dynamical system is shown to be an operator equivalence to the well-known matrix lemma:

$$(D+CLB)^{-1}=[D^{-1}-D^{-1}C(L^{-1}+BD^{-1}C)^{-1}BD^{-1}]$$

CONSIDER the system

$$\begin{cases}\dot{x}(t)=Ax(t)+Bu(t), & x(t_0)=a\\ y(t)=Cx(t)+Du(t).\end{cases} \tag{1}$$

If (1) has the same number of inputs and outputs and if the matrix D is regular, the inverse system [1] is easily obtained as

$$\begin{cases}u(t)=-D^{-1}Cx(t)+D^{-1}y(t)\\ \dot{x}(t)=(A-BD^{-1}C)x(t)+BD^{-1}y(t), & x(t_0)=a.\end{cases} \tag{2}$$

Now specialize to $a=0$, and let $\varphi(t, s)$ and $\psi(t, s)$ be the fundamental matrices corresponding to (1) and (2). Define the operators L and M in the space of continuous vector functions by

$$x=Lz, \qquad x(t)=\int_{t_0}^{t}\varphi(t, s)z(s)\mathrm{d}s$$

$$x=Mz, \qquad x(t)=\int_{t_0}^{t}\psi(t, s)z(s)\mathrm{d}s$$

and regard the matrices B, C, D and D^{-1} as operators in the function space.

The input-output relation of the system (1) can then be written as

$$y=(CLB+D)u. \tag{3}$$

Notice that

$$\left(\frac{\mathrm{d}}{\mathrm{d}t}-A\right)Lz=z$$

and therefore define

$$L^{-1}=\frac{\mathrm{d}}{\mathrm{d}t}-A. \tag{4}$$

Correspondingly we have for M

$$\left[\frac{\mathrm{d}}{\mathrm{d}t}-(A-BD^{-1}C)\right]Mz=z$$

and the inverse of M can thus be defined as

$$M^{-1}=\frac{\mathrm{d}}{\mathrm{d}t}-(A-BD^{-1}C)=L^{-1}+BD^{-1}C. \tag{5}$$

Insert (5) into (2) giving

$$x=(L^{-1}+BD^{-1}C)^{-1}BD^{-1}y$$

and

$$u=\{D^{-1}-D^{-1}C(L^{-1}+BD^{-1}C)^{-1}BD^{-1}\}y. \tag{6}$$

A comparison of (6) with (3) now gives the identity:

$$(D+CLB)^{-1}=\{D^{-1}-D^{-1}C(L^{-1}+BD^{-1}C)^{-1}BD^{-1}\}. \tag{7}$$

For matrix inversion this identity is well-known, see Ref. [2].

The inversions of L and M and especially $(L^{-1}+BD^{-1}C)$ are here defined only formally. In order to be more strict the initial values have also to be considered.

* Received 15 November 1971. The original version of this paper was not presented at any IFAC meeting. It was recommended for publication in revised form by Associate Editor B. Anderson.

† The author is with the Division of Automatic Control, Lund Institute of Technology, Lund, Sweden.

This work was supported by the Swedish Board for Technical Development (contract 70-337/U270).

References

[1] M. K. SAIN and J. L. MASSEY: Invertibility of linear timeinvariant dynamical systems. *IEEE Trans. Aut. Control* **AC-14,** 141–149 (1969).

[2] T. E. FORTMAN: A matrix inversion identity. *IEEE trans. aut. control* **AC-15,** 599 (1970).

Résumé—L'article montre que l'inversion d'un système dynamique linéaire est une équivalence d'opérateurs au lemme bien connu de matrices:

$$(D+CLB)^{-1}=[D^{-1}-D^{-1}C(L^{-1}+BD^{-1}C)^{-1}BD^{-1}]$$

Zusammenfassung—Die Inversion eines linearen dynamischen Systems ist, wie gezeigt wird, eine Operatoräquivalenz zu dem wohlbekannten Matrix-Lemma

$$(D+CLB)^{-1}=[D^{-1}-D^{-1}C(L^{-1}+BD^{-1}C)^{-1}BD^{-1}]$$

Резюме—Показано что инверсия динейной динамической системы представляет собой операторную эквивалентность хорошо знакомой матричной леммы

$$(D+CLB)^{-1}=[D^{-1}-D^{-1}C(L^{-1}+BD^{-1}C)^{-1}BD^{-1}]$$

A NEW PROOF AND AN ADJOINT FILTER INTERPRETATION FOR LINEAR DISCRETE TIME SMOOTHING

PER HAGANDER

Report 7330 September 1973
Lund Institute of Technology
Division of Automatic Control

A NEW PROOF AND AN ADJOINT FILTER INTERPRETATION FOR LINEAR DISCRETE TIME SMOOTHING

Per Hagander

Abstract

Linear discrete time systems, usually formulated using difference equations, can also be described by operators, which is more general. The covariances for a stochastic system are expressed as operators, and the solution of the fixed interval smoothing problem is obtained by use of the projection theorem:

$$\hat{x} = R_{xy} R_y^{-1} y$$

The computation of $\hat{x}$ is conveniently done if R_y can be factored into two Volterra (triangular) operators. It is shown how this factorization can be carried out using the Riccati equation, so that the estimate can be expressed as two adjoint coupled filters, the Bryson-Frazier formulation.

From the operator identity used for factorization it is seen that the one step ahead predictor is fundamental. Both the forward backward difference equations and the weighting function representation are presented, and the weighting function is shown to be the error covariance of the one step ahead predictor.

1. I

The
func
can
line
spac
demo
equa
tain
solv

2. N

Cons

x(t+

The
can
as a
line

x =

with
(2.1

Usir

Sinc

Defi

x_1 •)

givi

L^*:

Defi

$T_o x$

1. Introduction

The two approaches to linear estimation problems, the Wiener filter using covariance functions and the Kalman filter directly using difference or differential equations can be unified by use of the Riccati equation. [7,12]. In [7] the continous time linear control and estimation problems were analysed using operators in function spaces. The same technique is applicable in the discrete time case. This is demonstrated here on the smoothing problem. The projection theorem gives an equation in covariance operators from which the difference equations are obtained by operator factorization using the discrete Riccati equation. The resolvent identity searched for by Kailath and Frost [9] is thus presented.

2. Notations

Consider a discrete time system for $t\in[t_o,t_1]$

$$x(t+1) = \phi(t+1,t)x(t) + v(t), \qquad x(t_o) = x_o \tag{2.1}$$

The state of (2.1) $x(t)$, at all times t during the discrete time interval $[t_o,t_1]$, can be formed as a long vector with $n(t_1-t_o+1)$ elements, but it can also be regarded as a time function on $[t_o,t_1]$. The difference equation (2.1) can be formulated using linear operators in a space X of such functions:

$$x = gx_o + Lv$$

with $L: X \to X$ and $g: R^n \to X$. The operator formulation is more general than (2.1), and (2.1) thus introduces special structure on the operators.

Using the long vector interpretation these operators are in fact large matrices. Since L is causal it corresponds to a lower block triangular matrix.

Define in X the scalar product

$$x_1 \cdot x_2 = \sum_{t=t_o}^{t_1} x_1^T(t)x_2(t)$$

giving the adjoint of L:

$$L^*: X \to X; \quad z = L^*x, \quad z(t) = \sum_{s=t+1}^{t_1} \phi^T(s,t+1)x(s), \qquad \phi(t,s) = \prod_{i=s}^{t-1} \phi(i+1,i)$$

Define also the functions from X to R^n:

$$T_o x = x(t_o), \qquad T_1 x = x(t_1)$$

3. Linear, Stochastic, Time Discrete Systems

The space X can be extended to contain stochastic processes generated by linear systems driven by white noise. Such a Hilbert space is often used in the theory of stochastic processes, cf [4]. Let v and e of

$$x(t+1) = \phi x(t) + v(t), \qquad x(t_o) = x_o$$
$$y(t) = \theta x(t) + e(t) \tag{3.1}$$

be zero mean, independent white noise with covariances R_1 and R_2 ($R_2>0$), and let x_o have zero mean value, covariance R_o and be independent of v and e. The operators L and g are directly generalized. A new scalar product

$$x_1 \cdot x_2 = E \sum_{t=t_o}^{t_1} x_1^T(t)x_2(t)$$

gives the same adjoints. Notice that the deterministic functions constitute a subspace.

The covariance operator of x is easily obtained from the reformulation of (3.1).

$$x = Lv + gx_o$$

using the matrix point of view:

$$R_x = LR_1L^* + gR_og^*$$

where R_1 is now an operator in X (or a block diagonal matrix).
Moreover, $R_{xy} = R_x\theta^T$ and $R_y = \theta R_x\theta^T + R_2$. θ^T is a diagonal operator with $\theta^T(t)$ in the diagonal.

4. Smoothing Estimate

All linear estimates of x based on $\{y(t_o),\ldots,y(t_1)\}$ can be written

$$\hat{x} = Fy$$

If the operator F is such that the error variance in each component $\hat{x}_i(t)$ is minimized then $\hat{x}$ is the smoothing estimate of x.

Using the projection theorem [4], F must satisfy

$$R_{xy} = FR_y$$

In the continuous time case it has been shown that the Riccati equation decomposes operators like R_y into a product of a causal and an anti-causal part [7,8,12]. The algebra of the discrete time case is more involved and a corresponding identity has not been obtained previously, cf[9].

In its most simplified form the discrete time identity can be formulated as

Theorem 1: Let $P(t)$ be the solution of

$$P(t+1) = \phi P(t)\phi^T + I - \phi P(t)(I+P(t))^{-1}P(t)\phi^T$$
$$P(t_o) = 0$$

then

$$I + LL^* = (I+P+L\phi P)(I+P)^{-1}(I+P+P\phi^T L^*)$$

where P is a blockdiagonal operator with $P(t)$ in the diagonal.

Proof: With the forward and backward shift operators q and q^{-1} defined by $qx(t)=x(t+1)$, $qx(t_1)=0$, $q^{-1}x(t)=x(t-1)$, $q^{-1}x(t_o)=0$, it is easy to prove that

$$L(q-\phi) = I - gT_o \tag{4.1}$$

$$(q^{-1}-\phi^T)L^* = I - (gT_o)^* \tag{4.2}$$

so that the proposition

$$I + LL^* = I + P + L\phi P + P\phi^T L^* + L\phi P(P+I)^{-1}P\phi^T L^*$$

could be written

$$I+LL^* = I+L(q-\phi)P(q^{-1}-\phi^T)L^*+L\phi P(q^{-1}-\phi^T)L^*+L(q-\phi)P\phi^T L^*+L\phi P(P+I)^{-1}P\phi^T L^*+$$
$$+ gT_oP + L(q-\phi)P(gT_o)^* + L\phi P(gT_o)^* + gT_oP\phi^T L^*$$

When $T_oP = 0$ this requires

$$L\{I - qPq^{-1} + \phi P\phi^T - \phi P(I+P)^{-1}P\phi^T\}L^* = 0$$

which is true for $P(t)$ from the Riccati equation. □

The problem of decomposing R_y is solved by a generalization of Theorem 1.

Corollary 1: Let P be defined by

$$P(t+1) = \phi P(t)\phi^T + R_1 - \phi P\theta^T(\theta P\theta^T+R_2)^{-1}\theta P\phi^T$$
$$P(t_o) = R_o \tag{4.3}$$

then

$$R_x = L\phi P + P + P\phi^T L^* + L\phi P\theta^T(\theta P\theta^T+R_2)^{-1}\theta P\phi^T L^*$$
$$R_{xy} = P(I+\phi^T L^*)\theta^T + L\phi P\theta^T(\theta P\theta^T+R_2)^{-1}(\theta P\theta^T+R_2+\theta P\phi^T L^*\theta^T)$$
$$R_y = [\theta P\theta^T+R_2+\theta L\phi P\theta^T](\theta P\theta^T+R_2)^{-1}[\theta P\theta^T+R_2+\theta P\phi^T L^*\theta^T]$$

Proof: The only difficulty compared with Theorem 1 is the initial value. Just prove that

$$g^* = T_o(\phi^T L^* + I)$$

and use $T_oP = R_oT_o$. □

The smoothing formula can now be obtained using inversion of operators.

Theorem 2: The smoothing estimate for the system (3.1) is given by

$$\hat{x}(t|t_1) = \hat{x}(t|t-1) + P(t)\lambda(t-1) \qquad t_o \le t \le t_1$$

where $\hat{x}(t|t-1) = \hat{x}_p(t)$ is the one step ahead predictor.

$$\hat{x}_p(t+1) = \psi(t+1,t)\hat{x}_p(t) + K(t)y(t), \qquad \hat{x}_p(t_o) = 0 \tag{4.4}$$

$$\psi(t+1,t) = \phi(t+1,t) - K(t)\theta(t) \tag{4.5}$$

$$K = \phi P\theta^T(\theta P\theta^T+R_2)^{-1} \tag{4.6}$$

and λ the solution to an adjoint equation

$$\lambda(t-1) = \psi^T(t+1,t)\lambda(t) + \theta^T(\theta P\theta^T+R_2)^{-1}(y(t)-\theta\hat{x}_p(t)), \quad \lambda(t_1) = 0 \tag{4.7}$$

$P(t)$ is defined by (4.3).

Proof: Since the operator $[(\theta P\theta^T+R_2) + \theta L\phi P\theta^T]$ represents an invertible dynamical system [6], R_y is also invertible and using corollary 1 and (4.2):

$$\hat{x} = \{P(I+\phi^T L^*)\theta^T(\theta P\theta^T+R_2+\theta P\phi^T L^*\theta^T)^{-1}(\theta P\theta^T+R_2)+L\phi P\theta^T\}(\theta P\theta^T+R_2+\theta L\phi P\theta^T)^{-1}y$$

Introduce an operator M analogously to L:

$$(Mx)(t) = \sum_{s=t_o}^{t-1} \psi(t,s+1)x(s) \quad , \quad \psi(t,s) = \sum_{i=s}^{t-1} \psi(i+1,i) \tag{4.8}$$

with ψ from (4.5). Using the same technique as in [6] it is possible to prove the following operator identities:

$$(\theta P\theta^T+R_2+\theta L\phi P\theta^T)^{-1} = (\theta P\theta^T+R_2)^{-1} - (\theta P\theta^T+R_2)^{-1}\theta MK$$

$$L\phi P\theta^T(\theta P\theta^T+R_2+\theta L\phi P\theta^T)^{-1} = MK$$

$$L^*\theta^T(\theta P\theta^T+R_2+\theta P\phi^T L^*\theta^T)^{-1} = M^*\theta^T(\theta P\theta^T+R_2)^{-1}$$

with K from (4.6), so that

$$\hat{x} = MKy + P(I-\theta^T K^T M^*+\phi^T M^*)\theta^T(\theta P\theta^T+R_2)^{-1}(I-\theta MK)y = MKy + P(I+\psi^T M^*)\theta^T(\theta P\theta^T+R_2)^{-1}(I-\theta MK)y$$

Note that $\hat{x}_p$ and λ defined by (4.4) and (4.7) can be written as

$$\hat{x}_p = MKy \tag{4.9}$$

$$\lambda = M^*\theta^T(\theta P\theta^T+R_2)^{-1}(y-\theta\hat{x}_p) \tag{4.10}$$

which proves the theorem. □

5. Adjoint Filter Interpretation

Introduce

$$\tilde{x}_p(t) = x(t) - \hat{x}_p(t)$$

$$\tilde{y}_p(t) = y(t) - \theta\hat{x}_p(t)$$

and the covariance function $P(t,s)$ of $\tilde{x}_p$, cf [2].

$$P(t,s) = \begin{cases} \psi(t,s)P(s) & t > s \\ P(t) & t = s \\ P(t)\psi^T(s,t) & t < s \end{cases} \tag{5.1}$$

and define

$$y_m = \theta^T R_2^{-1} y \tag{5.2}$$

$$\nu_m = \theta^T (\theta P \theta^T + R_2)^{-1} \tilde{y}_p \tag{5.3}$$

The adjoint filter form of the smoothing estimate was derived in [9] for the continuous time case. A corresponding formula for discrete time can be formulated:

Corollary 2: The smoothing estimate for the system (3.1) is given by

$$\hat{x}(t) = \hat{x}(t|t_1) = \sum_{s=t_o}^{t-1} P(t,s) y_m(s) + \sum_{s=t}^{t_1} P(t,s) \nu_m(s) \tag{5.4}$$

with $P(t,s)$, y_m and ν_m defined by (5.1), (5.2) and (5.3).

Proof: First notice that

$$K = \phi P \theta^T (\theta P \theta^T + R_2)^{-1} = (\phi - \phi P \theta^T (\theta P \theta^T + R_2)^{-1} \theta) P \theta^T R_2^{-1} = (\phi - K\theta) P \theta^T R_2^{-1} \tag{5.5}$$

Hence from (4.9)

$$\hat{x}_p(t) = \sum_{s=t_o}^{t-1} \psi(t,s) P(s) \theta^T R_2^{-1} y(s) = \sum_{s=t_o}^{t-1} P(t,s) y_m(s)$$

and from (4.10)

$$P(t)\lambda(t-1) = P(t) \sum_{s=t}^{t_1} \psi^T(s,t) \theta^T (\theta P \theta^T + R_2)^{-1} \tilde{y}_p(s) = \sum_{s=t}^{t_1} P(t,s) \nu_m(s)$$

which proves the Corollary. □

Remark: The fixed point smoothing problem is directly solved from eq (5.4).

$$\begin{aligned} &\hat{x}(t|s+1) = \hat{x}(t|s) + B(s+1)\nu_m(s+1) \qquad s \geq t-1 \\ &B(t) = P(t) \\ &B(s+1) = B(s)\psi^T(s+1,s) = B(s)[\phi^T(s+1,s) - \theta^T(s) K^T(s)] \end{aligned} \tag{5.6}$$

Eq (5.6) could also be used to evaluate the fixed interval estimate $\hat{x}(t|t_1)$. This recursion is on the stability boundary but it has computational advantages since it can be performed in parallel with the one step ahead predictor and the Riccati equation. Variants of Theorem 2 and (5.6) have been presented earlier, see [3,9]. Some other formulations contain an unstable recursion or inversions to be performed in each time step [1,5,10,11].

6. Conclusions

This note presents a new proof of the discrete time smoothing problem by means of an operator identity searched for by Kailath and Frost [9]. It also gives the adjoint formula in the sense of [9, p 656]. The main difference between earlier derivations and this one is that the estimation is done once for all directly in the function

space giving an analogue of the Wiener Hopf equation, and not performed recursively in time.

The role of the Riccati equation in the factorization of the critical operator R_y is made clear, and the importance of the one step ahead predictor is more obvious.

References

[1] B.D.O. Anderson and S. Chirarattananon: New Linear Smoothing Formulas, IEEE Trans. Automat. Contr vol AC-17, pp 160-161, Feb 1972.

[2] K.J. Åström: Introduction to Stochastic Control Theory. Academic Press, New York 1970.

[3] A.E. Bryson and M. Frazier: Smoothing for Linear and Nonlinear Dynamic Systems, Proc. Optimum Sys. Synthesis Conf., U.S. Air Force Tech. Rept. ASO-TOR-63-119, Feb 1963.

[4] H. Cramér and M.R. Leadbetter: Stationary and Related Stochastic Processes, John Wiley, New York 1967.

[5] D.C. Fraser and J.E. Potter: The Optimum Linear Smoother as a Combination of Two Optimum Linear Filters, IEEE Trans. Automat Contr vol AC-14, pp 387-390, Aug 1969.

[6] P. Hagander: Inversion of a Dynamical System by an Operator Identity, Automatica 8, p 361-362, May 1972.

[7] P. Hagander: The Use of Operator Factorization for Linear Control and Estimation, Automatica 9, Sept 1973.

[8] T. Kailath: Application of a Resolvent Identity to a Linear Smoothing Problem, SIAM J Control vol 7, pp 68-74, Feb 1969.

[9] T. Kailath and P.Frost: An Innovation Approach to Least-Squares Estimation, Part II: Linear Smoothing in Additive White Noise, IEEE Trans. Automat. Contr. vol AC-13, pp 655-660, Dec 1968.

[10] J.S. Meditch: Stochastic Optimal Linear Estimation and Control, McGraw Hill, New York 1969.

[11] H.E. Rauch, F. Tung, and C.T. Striebel: Maximum Likelihood Estimates of Linear Dynamic Systems, AIAA J. vol 3, pp 1445-1450, Aug 1965.

[12] A. Schumitsky: On the Equivalence between Matrix Riccati Equations and Fredholm Resolvents, Journal of Computer and System Sciences, vol 2, pp 76-87, June 1968.

KALMAN FILTERS FOR PROCESSES WITH UNKNOWN INITIALVALUES

PER HAGANDER

Report 7332 September 1973
Lund Institute of Technology
Division of Automatic Control

KALMAN FILTERS FOR PROCESSES WITH UNKNOWN INITIAL VALUES.

P. Hagander

ABSTRACT.

A Kalman filter needs an à priori statistics for the initial state. It is shown how the filter should be started if some part of the initial state is totally unknown. The duality with optimal control with end point constraints is very useful both for proofs and intuition.

The usual way of starting with a very large covariance has very bad numerical properties. The optimal discrete time filter is determined by two "Riccati equations", one matrix to keep track of the bias until the unknown initial value is observable, and one matrix for the error covariance.

In continuous time the estimation is more complicated. The whole system becomes observable at once. After an initial discontinuity a usual Kalman filter could be started, but the gain would be almost infinite. It is therefore suggested how the estimate should be calculated using a separation into two estimates. The optimal linear stochastic control is also discussed.

TABLE OF CONTENTS

1. INTRODUCTION.

State estimation for linear stochastic systems is a well established theory, see e.g. [2, 5]. The Kalman filter requires, however, a known statistics for the initial state, which is often not available. The suboptimal solution recommended for instance by Sorensen [9] seems to be the most accepted substitute. The covariance of the initial state is assumed to be a unit matrix times a scalar, which is large compared with all other covariances. The resulting estimate and especially the assumed error covariance may, however, deviate considerably from the correct values. A serious fact is also that the Riccati equation (discrete or continuous time) will be ill conditioned if the initial covariance is very large.

Here it will be shown how to obtain the best linear unbiased estimate, i.e. the minimal variance estimate under the constraint that the expected value of the error should be zero for all initial states. Of course, it is only possible to obtain such an estimate for observable systems.

Such estimates were discussed by Kalman [5] for the case with only measurement noise, but they do not seem to have received much attention since, probably because the algebra is repellent. In the context of least squares parameter estimation, there has appeared some related work, e.g. [1].

In the following section the problem will be defined rigorously, and the solution for discrete time is obtained by letting the initial covariance reach infinity. The bad numerical properties of large covariances will be obvious, and the necessity of two Riccati equations is demonstrated.

In order to give a better understanding of the estimate and the necessary matrices, another derivation is made using a se-

paration into two estimates. In Section 3 the separation is shown, giving one filter for the stochastic terms and one for the unknown initial terms that has only measurement noise. The latter case is treated in Section 4, where recursive equations are given also before the system has become observable. It is shown in Section 5 how the two filters combine to one, the same as obtained in Section 2.

The filter is a time variable "dead beat" filter, which is especially simple in the single output case. Two matrix recursions are needed to get the filter gain. One matrix is the error covariance, and the other spans the bias of the estimate. As soon as the unknown initial value is observable the bias is zero, and the filter will continue as the usual Kalman filter.

In Section 6 the separation principle is applied, and the linear stochastic regulator problem is discussed.

Finally, it is shown that the problem is the dual of optimal control for fixed end state, and the solution for continuous time is obtained in this way. The difficulties with the differential equation formulation are discussed for time invariant systems, and a smoothing type algorithm is recommended for continuous time.

2. PROBLEM FORMULATION AND MAIN THEOREM.

Consider the discrete time system

$$\begin{aligned} x(t+1) &= \phi x(t) + v(t) \qquad x(t_0) = x_0 \\ y(t) &= \theta x(t) + e(t) \end{aligned} \tag{2.1}$$

where v and e are uncorrelated white noise sequences with covariances R_1 and R_2. For the initial state probability will be introduced in various degrees:

$$x_0 = x_0^S + x_0^N$$

The only assumption about x_0^N is that it is restricted to a subspace spanned by the full column rank rectangular matrix N^T:

$$x_0^N = N^T\xi \qquad \xi \text{ arbitrary}$$

x_0^S is uncorrelated with v and e and has zero mean value and covariance R_0^S. A very natural assumption, which will nowhere be used, is that x_0^S is restricted to a subspace disjoint from the range space of N^T.

Best linear unbiased estimate:

Introduce Y_t as the function y(s), $s \in [t_0,\ldots,t]$. It is now interesting to express Y_t as a linear function of ξ and α_t, a process into which all introduced random variables are collected. α_t has zero mean value and covariance Q_t.

$$Y_t = W_t\xi + \alpha_t \tag{2.2}$$

A linear unbiased estimate of ξ is a function F_t of Y_t such that

$$EF_tY_t = \xi$$

for all values of ξ. The minimal variance unbiased estimator is given by the well-known Gauss-Markov Theorem, see e.g. [1, 7], provided that Q_t is nonsingular

$$\hat{\xi}(t) = \left(W_t^T Q_t^{-1} W_t\right)^{-1} W_t^T Q_t^{-1} Y_t \qquad (2.3)$$

It should be noted that if ξ is assumed to be a random variable independent of all other introduced random variables, with covariance $\lambda^{-2}I$ and zero mean value, the Projection Theorem, e.g. [7], gives

$$\hat{\xi} = R_{\xi Y} R_Y^{-1} Y = \lambda^{-2} I W^T \left(W\lambda^{-2} I W^T + Q\right)^{-1} Y = \left(W^T Q^{-1} W + \lambda^2 I\right)^{-1} W^T Q^{-1} Y \qquad (2.3a)$$

and (2.3) is obtained as the limit when $\lambda \to 0$, infinite initial covariance.

This demonstrates the equivalence between the minimal variance estimate for infinite à priori covariance and the minimal variance unbiased estimate.

If the system (2.1) has only white measurement noise, (2.3) directly solves the linear unbiased estimate of $x(t)$. Since $x(t)$ is a deterministic linear function of ξ, it follows that

$$\hat{x}(t|t) = \phi(t,t_0) N^T \hat{\xi}(t) \qquad (2.4)$$

where

$$\phi(t,t_0) = \prod_{s=t_0}^{t-1} \phi(s+1,s)$$

Eq. (2.3) is also much simpler than in the general case because α and thus Q are especially simple. The inverse required in (2.3) is an observability Gramian.

In the general case (2.1) it is more complicated to obtain $\hat{x}(t|t)$. The system can be written as

$$\begin{cases} Y_t = W_t\xi + \alpha_t \\ x(t) = \phi(t,t_0)N^T\xi + \beta(t) \end{cases} \tag{2.5}$$

where α and β are correlated, and the state estimate will be

$$\hat{x}(t|t) = \phi(t,t_0)N^T\hat{\xi} + \hat{\beta}(t) \tag{2.6}$$

Operator formulas like in [4] could be used to evaluate (2.6), or the problem could be converted to a problem with only measurement noise, see Section 3. A third possibility is to use the limit argument of (2.3a) applied to the usual Kalman problem.

The infinite covariance limit of the Kalman filter.

If again x_0^N of (2.1) is a stochastic variable independent of x_0^S, v and e and with covariance $\lambda^{-2}N^TN$, a minimal variance unbiased estimate for the original problem could then be obtained by letting λ go to zero. The usual Kalman filter gives the minimal variance estimate:

$$\hat{x}(t|t) = \hat{x}(t|t-1) + K(t)[y(t) - \theta\hat{x}(t|t-1)] \qquad (2.7)$$

$$\hat{x}(t+1|t) = \phi\hat{x}(t|t) \qquad \hat{x}(t_0|t_0-1) = 0$$

$$K(t) = P(t)\theta^T\left(\theta P(t)\theta^T + R_2\right)^{-1}$$

$$P(t+1) = \phi[P(t) - K(t)\theta P(t)]\phi^T + R_1$$

$$P(t_0) = R_0 = R_0^S + \lambda^{-2}N^TN$$

Theorem 1: The minimal variance unbiased linear estimate for (2.1) is obtained by (2.7) and K(t) from

$$K = \Lambda\theta^T(\theta\Lambda\theta^T)^+[I - (\theta P_m\theta^T+R_2)[A(\theta P_m\theta^T+R_2)A]^+] +$$

$$+ P_m\theta^T[A(\theta P_m\theta^T+R_2)A]^+ \qquad (2.8)$$

with

$$A = I - (\theta\Lambda\theta^T)^+(\theta\Lambda\theta^T) \qquad (2.9)$$

$$\Lambda(t+1) = \phi[\Lambda - \Lambda\theta^T(\theta\Lambda\theta^T)^+\theta\Lambda]\phi^T \qquad \Lambda(t_0) = N^TN \qquad (2.10)$$

$$P_m(t+1) = R_1 + \phi\left\{(I-K\theta)P_m(I-K\theta)^T + KR_2K^T\right\}\phi^T$$

$$P_m(t_0) = R_0^S \qquad (2.11)$$

M^+ denotes the Moore Penrose pseudo inverse of M.

The estimate will be unbiased only if $\Lambda(t) = 0$.

Proof: Use induction in t to show that

$$P(t) = \lambda^{-2}\Lambda(t) + P_m(t)$$

which is true for $t = t_0$.

Introduce the full rank decompositions

$$\theta\Lambda(t)\theta^T = U^TU \qquad \theta P_m(t)\theta^T = V^TV \qquad V^TV + R_2 = H^TH$$

and consider the inverse

$$[\theta P\theta^T+R_2]^{-1} = [\lambda^{-2}U^TU+H^TH]^{-1}$$

which could be rewritten using a pseudo inverse formula given by Cline [3], see also [1].

$$= [(\bar{H}^T\bar{H})^+ + \lambda^2(I-\bar{H}^+H)(U^TU)^+(I-\bar{H}^+H)^T -$$

$$- \lambda^4(I-\bar{H}^+H)(U^TU)^+H^TQM(\lambda)QH(U^TU)^+(I-\bar{H}^+H)^T]$$

with

$$A = I - U^+U \qquad \bar{H} = HA \qquad Q = I - \bar{H}\bar{H}^+ \qquad M(\lambda) = [I+\lambda^2QH(U^TU)^+H^TQ]^{-1}$$

Note also that

$$UA = 0 \qquad U[\bar{H}^T\bar{H}]^+ = 0 \qquad U(I-\bar{H}^+H) = U$$

so that

$$K = P\theta^T[\theta P\theta^T+R_2]^{-1} = \Lambda\theta^T(U^TU)^+(I-\bar{H}^+H)^T+P_m\theta^T(\bar{H}^T\bar{H})^+ +$$

$$+ \lambda^2P_m\theta^T(I-\bar{H}^+H)(U^TU)^+(I-\bar{H}^+H)^T -$$

$$- \lambda^2\Lambda\theta^T(U^TU)^+H^TQM(\lambda)QH(U^TU)^+(I-\bar{H}^+H)^T + O(\lambda^4)$$

and

$$K\theta P = \lambda^{-2}\Lambda\theta^T(U^TU)^+\theta\Lambda + \Lambda\theta^T(U^TU)^+(I-\bar{H}^+H)^T\theta P_m +$$

$$+ P_m\theta^T(\bar{H}^T\bar{H})^+\theta P_m + P_m\theta^T(I-\bar{H}^+H)(U^TU)^+\theta\Lambda -$$

$$- \Lambda\theta^T(U^TU)^+H^TQM(\lambda)QH(U^TU)^+\theta\Lambda + O(\lambda^2)$$

Thus for very small λ, $P(t+1)$ could be written

$$P(t+1) = \lambda^{-2}\Lambda(t+1) + P_m(t+1) \qquad (2.12)$$

with $\Lambda(t+1)$ from (2.10)

$$\Lambda(t+1) = \phi\left(\Lambda - \Lambda\theta^T(\theta\Lambda\theta^T)^+\theta\Lambda\right)\phi^T$$

$$P_m(t+1) = R_1 + \phi\{P_m - \Lambda\theta^T(U^TU)^+(I-\bar{H}^+H)^T\theta P_m -$$
$$- P_m\theta^T(I-\bar{H}^+H)(U^TU)^+\theta\Lambda - P_m\theta^T(\bar{H}^T\bar{H})^+\theta P_m +$$
$$+ \Lambda\theta^T(U^TU)^+H^TQM(\lambda)QH(U^TU)^+\theta\Lambda\}\phi^T$$

and

$$K(t) = \Lambda\theta^T(U^TU)^+(I-\bar{H}^+H)^T + P_m\theta^T(\bar{H}^T\bar{H})^+$$

which directly gives (2.8).

The induction is completed.

It is also clear from (2.12) that P will be large if Λ is not zero. In the limit this means that the system is not unbiased.

In order to prove (2.11) note that

$$(I-K\theta)P_m(I-K\theta)^T + KR_2K^T = P_m - K\theta P_m - P_m\theta^TK^T + K(\theta P_m\theta^T+R_2)K^T$$

So it remains to show that

$$P_m\theta^T(\bar{H}^T\bar{H})^+\theta P_m + \Lambda\theta^T(U^TU)^+H^TQM(0)QH(U^TU)^+\theta\Lambda = KH^THK^T$$

But rewrite the right hand side using:

$$P_m\theta^T(\bar{H}^T\bar{H})^+H^TH(\bar{H}^T\bar{H})^+\theta P_m = P_m\theta^T(\bar{H}^T\bar{H})^+\theta P_m$$

$$P_m\theta^T(\bar{H}^T\bar{H})^+H^TH[I - (\bar{H}^T\bar{H})^+H^TH](U^TU)^+\theta\Lambda = 0$$

$$\Lambda\theta^T(U^TU)^+[I - H^TH(\bar{H}^T\bar{H})^+]H^TH[I - (\bar{H}^T\bar{H})^+H^TH](U^TU)^+\theta\Lambda =$$
$$= \Lambda\theta^T(U^TU)^+[H^TH - H^TH(\bar{H}^T\bar{H})^+H^TH](U^TU)^+\theta\Lambda$$

and the second term of the left hand side:

$$H^TQM(0)QH = H^T(I-\bar{H}\bar{H}^+)H = H^TH - H^T\bar{H}(\bar{H}^T\bar{H})^+\bar{H}^TH =$$
$$= H^TH - H^TH(\bar{H}^T\bar{H})^+H^TH$$

which completes the proof of (2.11) and the whole theorem.

□

Example 2.1:

$$\phi = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \qquad \theta = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \qquad R_1 = 0 \qquad R_2 = 1$$

$$P(0) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \lambda^{-2} \end{bmatrix} \qquad K(0) = \begin{bmatrix} 1/2 \\ 0 \\ 0 \end{bmatrix}$$

$$P - K\theta P = \begin{bmatrix} 1/2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \lambda^{-2} \end{bmatrix} \qquad P(1) = \begin{bmatrix} 3/2 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \lambda^{-2} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

$$K(1) = \begin{bmatrix} 3/5 \\ 2/5 \\ 0 \end{bmatrix} \qquad P - K\theta P = \begin{bmatrix} 0.6 & 0.4 & 0 \\ 0.4 & 0.6 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \lambda^{-2} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

$$P(2) = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 0.6 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \lambda^{-2} \begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}$$

$$K(2) = \left\{ \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix} + \lambda^{-2} \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} \right\} \cdot \frac{1}{3 + \lambda^{-2}} \to \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}$$

$$P - K\theta P \approx (\lambda^{-2} - \lambda^{-2}) \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 1 \end{bmatrix} + \begin{bmatrix} 1 & 2 & 1 \\ 2 & 8.6 & 5 \\ 1 & .5 & 3 \end{bmatrix} \qquad \lambda \to 0$$

which implies subtraction of very large numbers. This is avoided by theorem 1. The two terms of $P - K\theta P$ are stored separately in Λ and P_m.

In order to be able to use calculation by hand the example is very much simplified, and the illconditioness might

seem reasonable, but for a real system there is no significance left after a few such subtractions. The gain K will contain serious errors, and the real error covariance will not decrease although P does.

Example 2.2: θ is changed to

$$\theta = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \qquad R_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Then

$$K(0) = \theta^T \cdot 1/2 \qquad P - K\theta P = \begin{bmatrix} 1/2 & 0 & 0 \\ 0 & 1/2 & 0 \\ 0 & 0 & \lambda^{-2} \end{bmatrix}$$

$$P(1) = \frac{1}{2}\begin{bmatrix} 2 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \lambda^{-2}\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

$$K(1) = \left\{\frac{1}{2}\begin{bmatrix} 2 & 1 \\ 1 & 1 \\ 0 & 0 \end{bmatrix} + \lambda^{-2}\begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}\right\}\left(\begin{bmatrix} 2 & 1/2 \\ 1/2 & 3/2 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & \lambda^{-2} \end{bmatrix}\right)^{-1} \rightarrow$$

$$\rightarrow \frac{1}{4}\begin{bmatrix} 2 & 0 \\ 0 & 1 \\ -1 & 1 \end{bmatrix} \qquad \lambda \rightarrow 0 \quad .$$

after very illconditioned operations!

Comments: The new formula (2.8) may be seen as a way of improving the numerical condition of the calculations by keeping track of the large terms using the new matrix Λ. When Λ has become zero the filter is identical with the

usual Kalman filter, and the error covariance is $P = P_m$. Note that (2.11) is the form of the covariance updating formula that is valid for any K. It is not possible to rewrite (2.11) as a simple Riccati equation.

The interpretation of the optimal gain K in (2.8) is most obvious in the single output case. Then A is either 1 or 0, and

$$K = \begin{cases} \Lambda\theta^T(\theta\Lambda\theta^T)^{-1} & \text{if } A = 0 \\ P_m\theta^T(\theta P_m\theta^T+R_2)^{-1} & \text{if } A = 1 \end{cases} \qquad (2.13)$$

If A = 0 the measurement contributes to the observability of ξ, and if A = 1 the information is only used to improve the current estimate just as in a Kalman filter.

Often the whole x_0 is unknown, i.e. N is a square matrix. Then A = 0 and $K = \Lambda\theta^T(\theta\Lambda\theta^T)^{-1}$ until t = n and $\Lambda(n) = 0$, where n is the order of the system. All the first n measurements contribute to the observability. The filter could be called a dead-beat filter with time varying gain in analogy with dead-beat controllers. The influence of the unknown initial value on the estimation error at t = n is zero. The estimate $\hat{x}(n|n-1)$ is unbiased. In this special case there are in fact no other unbiased estimates at t = n. Even the time constant dead-beat filter would have given the same $\hat{x}(n|n-1)$. It can be shown that the gain of the time constant filter is $K_c = K(n-1) = \Lambda(n-1)\theta^T(\theta\Lambda(n-1)\theta^T)^{-1}$. When the dead-beat filter is not unique, there is freedom left to minimize the error covariance. This is why the more complicated expressions (2.8) and (2.13) should be used in the multi-output case and when only part of the initial state is unknown. The filters are still time variable dead-beat filters!

The simple formulas (2.13) can be used also in the multi-output case, if the noise of the elements in the output vector are uncorrellated, i.e. if R_2 is diagonal. The elements can be used one at a time to update the estimate.
In order to give an interpretation of Λ and of the estimates for times before $\Lambda(t) = 0$, the theorem will be re-derived using a separation into two estimates, one for the stochastic terms and one with only measurement-noise for the unknown initial value.

3. SEPARATION INTO TWO ESTIMATES.

The formula from Section 2:

$$\hat{x}(t|t) = \phi(t,t_0)N^T\hat{\xi} + \hat{\beta}(t)$$

shows that $\hat{x}$ can be written as a sum of two estimates. It seems reasonable to separate the state into a stochastic term and a deterministic ξ-term like in (2.5):

$$x = x_1 + x_2 \tag{3.1}$$

$$\begin{cases} x_1(t+1) = \phi x_1(t) + v(t) & x_1(t_0) = x_0^S \\ y_1(t) = \theta x_1(t) + e(t) \end{cases} \tag{3.2}$$

$$\begin{cases} x_2(t+1) = \phi x_2(t) & x_2(t_0) = x_0^N = N^T\xi \\ y_2(t) = \theta x_2(t) \end{cases} \tag{3.3}$$

The Kalman filter for (3.2) would be

$$\hat{x}_1(t+1|t) = \phi\hat{x}_1(t|t-1) + \phi K_\Pi(t)[y_1(t) - \theta\hat{x}_1(t|t-1)]$$

$$\hat{x}_1(t_0|t_0-1) = 0$$

$$K_\Pi(t) = \Pi(t)\theta^T\left(\theta\Pi(t)\theta^T + R_2\right)^{-1} \tag{3.4}$$

$$\Pi(t+1) = \phi[\Pi(t) - K_\Pi\theta\Pi(t)]\phi^T + R_1 \quad \Pi(t_0) = R_0^S \tag{3.5}$$

or in operator form

$$\hat{x}_1(t+1|t) = R_{x_1(t+1)Y_{1t}}R_{Y_{1t}}^{-1}Y_{1t}$$

Since y_1 is not available for measurement it is interesting to define $\hat{x}_{\Pi}$ as the same linear operator applied to y instead:

$$\hat{x}_{\Pi}(t+1|t) = R_{x_1(t+1)Y_{1t}} R^{-1}_{Y_{1t}} Y_t$$

or

$$\hat{x}_{\Pi}(t+1|t) = \phi\hat{x}_{\Pi}(t|t-1) + \phi K_{\Pi}(t)[y(t) - \theta\hat{x}_{\Pi}(t|t-1)]$$

$$\hat{x}_{\Pi}(t_0|t_0-1) = 0 \qquad (3.6)$$

Assume for a moment that ξ is a stochastic variable independent of v, e and x_0^S. Then by the projection theorem

$$\hat{x}(t+1|t) = R_{x(t+1)Y_t} R^{-1}_{Y_t} Y_t$$

which will be expressed in $\hat{x}_{\Pi}(t+1|t)$. Drop the time indices:

$$\hat{x} = (R_{x_1Y_1}+R_{x_2Y_2})R^{-1}_{Y}Y = R_{x_1Y_1}[R^{-1}_{Y_1}-R^{-1}_{Y_1}R_{Y_2}R^{-1}_{Y}]Y + R_{x_2Y_2}R^{-1}_{Y}Y =$$

$$= \hat{x}_{\Pi}+(R_{x_2Y_2}-R_{x_1Y_1}R^{-1}_{Y_1}R_{Y_2})R^{-1}_{Y}Y$$

by linearity. Since (3.2) and (3.3) are independent $R_{x_2Y_2} = R_{x_2Y}$ and $R_{Y_2} = R_{Y_2Y}$ and since

$$\hat{x}_{\Pi} - \hat{x}_1 = R_{x_1Y_1}R^{-1}_{Y_1}Y_2$$

it follows that

$$R_{x_2Y} - R_{x_1Y_1}R^{-1}_{Y_1}R_{Y_2Y} = R_{zY}$$

if z is defined as

$$z(t) = x_2(t) - \hat{x}_{\Pi}(t|t-1) + \hat{x}_1(t|t-1) \tag{3.7}$$

The projection theorem gives $\hat{z} = R_{zY}R_Y^{-1}Y$ so that

$$\hat{x} = \hat{x}_{\Pi} + \hat{z} = \hat{x}_{\Pi} + R_{zY}R_Y^{-1}Y \tag{3.8}$$

Introduce also the <u>equivalent</u> measurement η

$$\eta = y - \theta\hat{x}_{\Pi} \tag{3.9}$$

It can be shown that z and η satisfy a simple dynamic system:

<u>Theorem 2</u>: z and η defined by (3.7) and (3.9) satisfy the system

$$\begin{cases} z(t+1) = \phi(I-K_{\Pi}\theta)z(t) \qquad z(t_0) = x_0^N = N^T\xi \\ \eta(t) = \theta z(t) + \varepsilon(t) \end{cases} \tag{3.10}$$

where ε is white noise with covariance $(\theta\Pi(t)\theta^T + R_2)$. K_{Π} and Π are defined by (3.4) and (3.5).

<u>Proof</u>:

$$\begin{aligned} z(t+1) &= x_2(t+1) - \hat{x}_{\Pi}(t+1|t) + \hat{x}_1(t+1|t) = \\ &= \phi[x_2(t) - \hat{x}_{\Pi}(t|t-1) + \hat{x}_1(t|t-1)] - \\ &\quad - \phi K_{\Pi}(t)[y(t) - \theta\hat{x}_{\Pi}(t|t-1) - y_1(t) + \theta\hat{x}_1(t|t-1)] = \\ &= \phi z(t) - \phi K_{\Pi}(t)[y_2(t) + \theta z(t) - \theta x_2(t)] = \\ &= [\phi - \phi K_{\Pi}(t)\theta]z(t) \end{aligned}$$

$$z(t_0) = x_2(t_0) - \hat{x}_{\Pi}(t_0|t_0-1) + \hat{x}_1(t_0|t_0-1) = x_2(t_0)$$

$$\eta(t) = y(t) - \theta\hat{x}_{\Pi}(t|t-1) = y_1(t) - \theta\hat{x}_1(t|t-1) + \theta z(t)$$

The innovations $\varepsilon(t) = y_1(t) - \theta\hat{x}_1(t|t-1)$ are white with covariance $R_2 + \theta\Pi(t)\theta^T$.

□

If the covariance of ξ goes to infinity, $\hat{x}_{\Pi}$ is not affected. The original problem with unknown ξ could thus be resumed. $\hat{x}$ is the minimal variance unbiased estimate of x, if $\hat{z}$ is so of z. Theorem 2 is not influenced by the different interpretations of ξ.

Define the estimation errors $\tilde{x}$, $\tilde{x}_1$ and $\tilde{z}$. Then

$$\tilde{x} = x - \hat{x} = x_1 + x_2 - \hat{x}_{\Pi} - \hat{z} = \tilde{x}_1 + \tilde{z} \qquad (3.11)$$

Since $\tilde{x}_1$ and $\tilde{z}$ are uncorrelated, define Σ as the covariance of $\tilde{z}$ so that

$$\text{Cov}\,\tilde{x}(t|t-1) = P(t) = \text{cov}\,\tilde{x}_1(t|t-1) + \text{cov}\,\tilde{z}(t|t-1) =$$

$$= \Pi(t) + \Sigma(t) \qquad (3.12)$$

In order to get recursive formulas for the estimate $\hat{x}$, formulas must be obtained for $\hat{z}$. The system (3.10) contains only measurement noise. z is deterministic, but ξ is unknown. Such systems will be discussed in the next section.

4. MEASUREMENT NOISE ONLY.

The estimation problem is now brought back to the special case, $v = 0$, $x_0^S = 0$.

In Section 2 the Gauss Markov Theorem was used to express $\hat{\xi}$, given Y_t. The relations (2.2) and (2.3) are now simple since

$$y(t) = \theta\phi(t,t_0)N^T\xi + e(t)$$

and

$$W_t^T Q_t^{-1} W_t = \sum_{s=t_0}^{t} N\phi^T(s,t_0)\theta^T R_2^{-1}\theta\phi(s,t_0)N^T = NM_{t+1}N^T \qquad (4.1)$$

$$W_t^T Q_t^{-1} Y_t = \sum_{s=t_0}^{t} N\phi^T(s,t_0)\theta^T R_2^{-1} y(s) = N\lambda_{t+1} \qquad (4.2)$$

by obvious definitions of M and λ, so that

$$\hat{\xi}(t) = (NM_{t+1}N^T)^{-1}N\lambda_{t+1}$$

Minimal bias estimates: It is possible to get an unbiased estimate $\hat{\xi} = FY$ only if (4.1) is invertible. If not it is only possible to estimate some linear combinations of ξ without bias. Those components of ξ that lie in the null-space of W cannot be estimated without bias. There is, however, freedom left to decide, without knowledge of ξ, in what complementary subspace the estimate should be unbiased. Any such estimate is called a minimum bias estimate. If the rows of W are linear dependent, the freedom should be used to minimize the variance of the estimate. A minimal variance minimal bias linear estimate is thus obtained by the orthogonal pseudo inverse, see also [8].

Thus

$$\hat{\xi}_m = (NMN^T)^+ N\lambda_{t+1} \qquad (4.3)$$

for which the components of ξ in the range space of W^T will be estimated without bias.

By (2.4) a good state estimate is

$$\hat{x}_m(t|t) = \phi(t,t_0)N^T\hat{\xi}_m(t)$$

For the degenerate case when Q_t is not invertible similar formulas exist [1] but they will not be considered here. The proofs and the interpretation in the sequel would be more complicated although the final result may be similar.

The estimates can also be obtained by some minimization of the mean square error V in some norm $\| \ \|_q$ induced by a quadratic form $x^T q x$

$$V = E\| \xi - FY \|_q^2 = E\| \xi - FW\xi - Fe \|_q^2 = \| \xi - FW\xi \|_q^2 + E\| Fe \|_q^2$$

which, of course, cannot be done directly if nothing is known about ξ. The min max estimate is

$$\hat{\xi}_q = q^{-1}NMN^T(NMN^T q^{-1} NMN^T)^+ N\lambda \qquad (4.3a)$$

V is minimized for the worst ξ. The bias term is first minimized, then the variance term.

Note that $\hat{\xi}_q = \hat{\xi}_m$ if $q = I$. If ξ is observable $\hat{\xi}_q = \hat{\xi}_m$ for all q.

It is easily verified that $\hat{\xi}_m$ is the minimal variance unbiased estimate of $NMN^T(NMN^T)^+\xi$, and $\tilde{\xi}_m$ defined by

$$\tilde{\xi}_m = NMN^T(NMN^T)^+\xi - \hat{\xi}_m \tag{4.4}$$

has covariance $(NMN^T)^+$. The bias of $\hat{\xi}_m$ as an estimate of ξ is

$$E(\xi - \hat{\xi}_m) = [I - NMN^T(NMN^T)^+]\xi \tag{4.5}$$

Introduce also $\tilde{x}_m$

$$\tilde{x}_m(t|t) = \phi(t,t_0)N^T\tilde{\xi}_m(t) \tag{4.6}$$

with covariance $\phi(t,t_0)N^T(NMN^T)^+N\phi^T(t,t_0)$.

Example:

$$y = W\xi + e = [1 \quad 1]\xi + e \qquad Ee = 0 \qquad Ee^2 = 1$$

$$M = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \qquad M^+ = \frac{1}{4}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \qquad \hat{\xi}_m = M^+W^Ty = \frac{1}{2}\begin{bmatrix} 1 \\ 1 \end{bmatrix}y$$

$$\tilde{\xi}_m = MM^+\xi - \hat{\xi}_m = -\frac{1}{2}\begin{bmatrix} 1 \\ 1 \end{bmatrix}e \qquad \operatorname{cov}\tilde{\xi}_m = M^+ = \frac{1}{4}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

$$\tilde{\xi}_{mb} = (I-MM^+)\xi = \frac{1}{2}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}\xi$$

$$\min_F E\| \xi - FY \|_q^2 = \min_F \left\{ \| \xi - FW\xi \|_q^2 + E\| Fe \|_q^2 \right\} =$$

$$= \| \xi_0 \|_q^2 + \min_F \left\{ \| \xi_1 - FW\xi_1 \|_q^2 + \| F \|_q^2 \right\}$$

where $W\xi_0 = 0$, $\langle \xi_1, \xi_0 \rangle_q = \xi_1^T q \xi_0 = 0$, $q\xi_1 \in R(W^T)$,

$$\xi = \xi_0 + \xi_1.$$

E.g. $q = I$:

$$F_m = \frac{1}{2}\begin{bmatrix}1\\1\end{bmatrix} \Rightarrow \xi_1 - FW\xi_1 = 0 \qquad \| \xi_0 \|^2 = \frac{1}{2}\xi^T \begin{bmatrix}1 & -1\\ -1 & 1\end{bmatrix}\xi$$

$$\| F \| = 1/2$$

E.g. $q = \begin{bmatrix}1 & 0\\ 0 & 2\end{bmatrix}$:

$$F_q = \frac{2}{3}\begin{bmatrix}1\\1/2\end{bmatrix} \qquad \hat{\xi}_q = \frac{1}{3}\begin{bmatrix}2\\1\end{bmatrix} y \qquad \tilde{\xi}_q = -\frac{1}{3}\begin{bmatrix}2\\1\end{bmatrix} e$$

$$\tilde{\xi}_{qb} = \frac{1}{3}\begin{bmatrix}1 & -2\\ -1 & 2\end{bmatrix}\xi \qquad \| \xi_0 \|_q^2 = \frac{1}{3}\xi^T \begin{bmatrix}1 & -2\\ -2 & 4\end{bmatrix}\xi$$

$$\| F \|_q = 5/9$$

Recursive equations: The pseudo inverse $(NMN^T)^+$ can be evaluated recursively using formulas derived by Cline [3], and both $\hat{\xi}_m$ and $\hat{x}_m$ satisfy difference equations like the Kalman filters.

Theorem 3: The minimal variance linear estimate of the state x of (2.1) obtained from the minimal variance, minimal bias, linear estimate of ξ

$$\hat{x}_m(t|t) = \phi(t,t_0)N^T\hat{\xi}(t)$$

satisfies the recursion:

$$\begin{cases} \hat{x}_m(t|t) = \hat{x}_m(t|t-1) + K(t)[y(t) - \theta\hat{x}_m(t|t-1)] \\ \hat{x}_m(t+1|t) = \phi\hat{x}_m(t|t) \end{cases} \tag{4.7}$$

where

$$K = \Lambda\theta^T(\theta\Lambda\theta^T)^+\left\{I - (\theta P_m\theta^T+R_2)[A(\theta P_m\theta^T+R_2)A]^+\right\} +$$

$$+ P_m\theta^T[A(\theta P_m\theta^T+R_2)A]^+ \tag{4.8}$$

$$A = I - (\theta\Lambda\theta^T)(\theta\Lambda\theta^T)^+ \tag{4.9}$$

An alternative expression for K is

$$K = \left(\Lambda\theta^T(\theta P_m\theta^T+R_2')^{-1}\theta\Lambda\right)^+\Lambda\theta^T(\theta P_m\theta^T+R_2)^{-1} +$$

$$+ P_m\theta^T[A(\theta P_m\theta^T+R_2)A]^+ \tag{4.8a}$$

$P_m(t)$, the covariance of

$$\tilde{x}_m(t|t-1) = \phi(t,t_0)N^T[NM_tN^T(NM_tN^T)^+\xi - \hat{\xi}_m(t-1)]$$

satisfies

$$P_m(t+1) = \phi[(I-K\theta)P_m(I-K\theta)^T + KR_2K^T]\phi^T, \quad P_m(t_0) = 0 \tag{4.10}$$

The matrix Λ defined by

$$\Lambda(t) = \phi(t,t_0)N^T[I - NM_tN^T(NM_tN^T)^+]N\phi^T(t,t_0) \tag{4.11}$$

satisfies

$$\Lambda(t+1) = \phi[\Lambda - \Lambda\theta^T(\theta\Lambda\theta^T)^+\theta\Lambda]\phi^T, \quad \Lambda(t_0) = N^TN \tag{4.12}$$

Proof: Introduce some notation:

$r^T r = R_2^{-1}$, $c_t = N\phi^T(t,t_0)\theta^T r^T$

so that

$$N\lambda_t = \sum_{s=t_0}^{t-1} c_s ry(s) \qquad \tilde{M}_t = NM_tN^T = \sum_{s=t_0}^{t-1} c_s c_s^T$$

$$\hat{\xi}_m(t-1) = \tilde{M}_t^+ N\lambda_t \qquad \hat{\xi}_m(t) = [\tilde{M}_t + c_t c_t^T]^+ [N\lambda_t + c_t ry(t)]$$

Now use the pseudo inverse formula and delete indices t:

$$(\tilde{M}+cc^T)^+ = \{(Dcc^TD)^+ + [I - (c^TD)^+c^T] \cdot$$

$$\cdot [\tilde{M}^+ - \tilde{M}^+cGBGc^T\tilde{M}^+][I - c(Dc)^+]\}$$

where $D = I - \tilde{M}^+\tilde{M}$, $G = I - (Dc)^+Dc$, $B = [I+Gc^T\tilde{M}^+cG]^{-1}$

Note that D is the projection on $R(\tilde{M})^\perp$ so that $D_t c_s = 0$, $s < t$ giving $D_t N\lambda_t = 0$. Thus

$$\hat{\xi}_m(t) = [I - (c^TD)^+c^T][\tilde{M}^+ - \tilde{M}^+cGBGc^T\tilde{M}^+]N\lambda +$$

$$+ \{(Dcc^TD)^+ + [I - (c^TD)^+c^T][\tilde{M}^+ - \tilde{M}^+cGBGc^T\tilde{M}^+] \cdot$$

$$\cdot [c - c(Dc)^+c]\}ry(t)$$

but $c - c(Dc)^+c = c - c(Dc)^+Dc = cG$ and $(Dcc^TD)^+c = (Dcc^TD)^+Dc = (c^TD)^+ = Dc(c^TDc)^+$ and $[\tilde{M}^+ - \tilde{M}^+cGBGc^T\tilde{M}^+]cG = \tilde{M}^+cG[I - BGc^T\tilde{M}^+cG]G = \tilde{M}^+cGBG$. Hence

$$\hat{\xi}_m(t) = \tilde{M}^+N\lambda + \{Dc(c^TDc)^+ + [I - Dc(c^TDc)^+c^T]\tilde{M}^+cGBG\} \cdot$$

$$\cdot [ry - c^T\tilde{M}^+N\lambda]$$

Introduce K'

$$K' = \{Dc(c^TDc)^+[I - c^T\tilde{M}^+cGBG] + M^+cGBG\}r =$$

$$= \{Dc(c^TDc)^+[I - (I+c^T\tilde{M}^+c)GBG] + \tilde{M}^+cGBG\}r$$

$$\hat{\xi}_m(t) = \hat{\xi}_m(t-1) + K'(t)[y(t) - \theta\phi(t,t_0)N^T\hat{\xi}_m(t-1)] \qquad (4.13)$$

The pseudo inverse algebra collected in Appendix takes care of the correlation $R_2 = r^{-1}r^{-T}$ in the expression of K', so that from (A5)

$$K' = Dcr^{-T}(r^{-1}c^TDcr^{-T})^+\{I - (R_2+r^{-1}c^T\tilde{M}^+cr^{-T}) \cdot$$

$$\cdot [A(R_2+r^{-1}c^T\tilde{M}^+cr^{-T})A]^+\} + \tilde{M}^+cr^{-T}[A(R_2+r^{-1}c^T\tilde{M}^+cr^{-T})A] \qquad (4.14)$$

where $A = I - (r^{-1}c^T D c r^{-T})(r^{-1}c^T D c r^{-T})^+$

Now (2.4) applied to (4.13) gives (4.7) with $K(t) = \phi(t,t_0)N^T K'(t)$.

A defined by (4.11) can be written $A(t) = \phi(t,t_0)N^T D_t N \phi^T(t,t_0)$ and the covariance $P_m(t) = \phi(t,t_0)N^T \bar{M}_t^+ N\phi^T(t,t_0)$ by (4.4). This directly proves (4.8) and (4.9).

It remains to prove (4.8a) and the recursions (4.10) and (4.12). Since $R_2 > 0$ eq. (4.8a) is a rather direct consequence of Lemma 3 in Appendix. Now regard

$$I - D_{t+1} = \bar{M}_{t+1}^+ \bar{M}_{t+1} = [I - (c^T D)^+ c^T][\bar{M}^+ - \bar{M}^+ cGBGc^T\bar{M}^+]\bar{M} +$$

$$+ (c^T D)^+ c^T + [I - (c^T D)^+ c^T]\bar{M}^+ cGBGc^T =$$

$$= [I - (c^T D)^+ c^T][I-D] + (c^T D)^+ c^T -$$

$$- [I - (c^T D)^+ c^T]\bar{M}^+ cGBGc^T(I-D) - \bar{M}^+ cGBGc^T\} =$$

$$= I - D + (c^T D)^+ c^T D + [I - (c^T D)^+ c^T]\bar{M}^+ cGBGc^T D$$

But $Gc^T D = 0$ so $D_{t+1} = D - Dc(c^T Dc)^+ c^T D$. The algebra in Appendix gives

$$Dc(c^T Dc)^+ c^T D = Dc(I-G)r^{-T}(r^{-1}c^T D c r^{-T})^+ r^{-1}(I-G)c^T D =$$

$$= Dcr^{-T}(r^{-1}c^T D c r^{-T})^+ r^{-1}c^T D$$

and (4.12) follows immediately.

The recursion for P_m is derived from recursions for $\tilde{x}_m$ and $\tilde{\xi}_m$.

$$\tilde{\xi}_m(t) = \bar{M}_{t+1}^+ M_{t+1}\xi - \tilde{\xi}_m(t) = \bar{M}^+\bar{M}\xi + Dc(c^T Dc)^+ c^T D\xi -$$

$$- \hat{\xi}_m(t-1) - K'(t)[y(t) - \theta\phi(t,t_0)\hat{\xi}_m(t-1)] =$$

$$= \tilde{\xi}_m(t-1) - K'(t)[\theta\tilde{x}_m(t|t-1) + e(t) + \theta\phi(t,t_0)N^T D\xi] +$$

$$+ Dc(c^T Dc)^+ c^T D\xi$$

But since $Gc^T D = 0$

$$K'\theta\phi(t,t_0)N^T D = K'r^{-1}c^T D = \{Dc(c^T Dc)^+[I - c^T\bar{M}^+ cGBG] +$$

$$+ \bar{M}^+ cGBG\}c^T D = Dc(c^T Dc)^+ c^T D$$

and thus

$$\tilde{x}_m(t+1|t) = \phi[I - K(t)\theta]\tilde{x}_m(t|t-1) - \phi K(t)e(t) \qquad (4.15)$$

$$P_m(t+1) = \phi\{[I-K\theta]P_m[I-K\theta]^T + KR_2K^T\}\phi^T$$

which concludes the proof of the theorem.

□

Comments: In the considered special case Theorem 1 and Theorem 3 give the same estimate. Theorem 3 also gives the interpretation of P_m as a covariance matrix and shows that $\hat{x}_m$ is obtained from the minimal bias estimate $\hat{\xi}_m$ of the unknown initial value. It is very natural to assume that N^TN is a projection, which means that the unknown parts of the initial value are "equally unknown". For instance by direct verification of the pseudo inverse conditions it then follows that

$$N^T(NMN^T)^+N = (N^TNMN^TN)^+$$

and

$$N^TNMN^TN(N^TNMN^TN)^+ = N^TN\ MM^+N^TN$$

so that

$$N^TDN = N^T[I - NMN^T(NMN^T)^+]N = N^TN[I-MM^+]N^TN$$

$$\Lambda(t) = \phi(t,t_0)N^TN[I-M_tM_t^+]N^TN\phi^T(t,t_0) \qquad (4.16)$$

Thus N^TDN is the projection on the unobservable part of the unknown initial value. Λ is the projection transformed to $x(t)$.

In the next section the full problem with also x_0^S and v will be treated, and the interpretation of the estimate and the matrices P_m and Λ will be similar.

5. COMBINATION OF TWO FILTERS.

Now return to the original problem (2.1) and to the separation made in Section 3. It was there shown that

$$\hat{x} = \hat{x}_{\Pi} + \hat{z}$$

where $\hat{x}_{\Pi}$ was the stochastic term (3.6) and $\hat{z}$ the bias term. According to Theorem 2 z was the state of a system with only white measurement noise. Such systems were treated in Section 4, and a good minimal variance estimate of z is obtained from the minimal variance, minimal bias linear estimate of ξ.

$$\hat{z}_m(t|t) = \left\{ \prod_{s=t_0}^{t-1} (\phi - \phi K_{\Pi}(s)\theta) \right\} \hat{\xi}_m(t) = \psi(t,t_0)\hat{\xi}_m(t) \qquad (5.1)$$

which satisfies

$$\hat{z}_m(t|t) = \hat{z}_m(t|t-1) + K_z(t)[y(t) - \theta\hat{x}_{\Pi}(t|t-1) - \theta\hat{z}_m(t|t-1)] \qquad (5.2)$$

$$\hat{z}_m(t+1|t) = [\phi - \phi K_{\Pi}(t)\theta]\hat{z}_m(t|t)$$

$$K_z = \Lambda\theta^T(\theta\Lambda\theta^T)^+\left\{I - (\theta\Sigma_m\theta^T+\theta\Pi\theta^T+R_2)[A(\theta\Sigma_m\theta^T+\theta\Pi\theta^T+R_2)A]^+\right\} + \Sigma_m\theta^T[A(\theta\Sigma_m\theta^T+\theta\Pi\theta^T+R_2)A]^+ \qquad (5.3)$$

$$A = I - \theta\Lambda\theta^T(\theta\Lambda\theta^T)^+$$

$$\Lambda(t+1) = \phi(I-K_{\Pi}\theta)[\Lambda - \Lambda\theta^T(\theta\Lambda\theta^T)^+\theta\Lambda](I-K_{\Pi}\theta)^T\phi^T$$

$$\Lambda(t_0) = N^TN \qquad (5.4)$$

$$\begin{cases} \Sigma_m(t+1) = \phi(I-K_\Pi\theta)\Big\{(I-K_z\theta)\Sigma_m(I-K_z\theta)^T + \\ \qquad + K_z(R_2+\theta\Pi\theta^T)K_z^T\Big\}(I-K_\Pi\theta)^T\phi^T \\ \Sigma_m(t_0) = 0 \end{cases} \tag{5.5}$$

where Σ_m is the covariance of $\tilde{z}_m(t|t-1)$

$$\tilde{z}_m(t|t-1) = \psi(t,t_0)N^T[NM_tN^T(NM_tN^T)^+\xi - \hat{\xi}(t-1)] \tag{5.6}$$

with

$$M_t = \sum_{s=t_0}^{t-1} \psi^T(s,t_0)\theta^T\left(\theta\Pi(s)\theta^T + R_2\right)^{-1}\theta\psi(s,t_0) \tag{5.7}$$

It can now be proven that (3.8) gives the same estimate as Theorem 1:

Theorem 4: The minimal variance linear estimate of the state x of (2.1) obtained from the minimal variance, minimal bias, linear estimate of the unknown initial state ξ by

$$\hat{x}_m(t|t) = \hat{x}_\Pi(t|t) + \hat{z}_m(t|t) \tag{5.8}$$

with $\hat{x}_\Pi$ defined by (3.6) and $\hat{z}_m$ by (5.1) satisfies the recursion

$$\hat{x}_m(t|t) = \hat{x}_m(t|t-1) + K(t)[y(t) - \theta\hat{x}_m(t|t-1)] \tag{5.9}$$

$$\hat{x}_m(t+1|t) = \phi\hat{x}_m(t|t)$$

where

$$K = \Lambda\theta^T(\theta\Lambda\theta^T)^+\Big\{I - (\theta P_m\theta^T+R_2)[A(\theta P_m\theta^T+R_2)A]^+\Big\} + \\ + P_m\theta^T[A(\theta P_m\theta^T+R_2)A]^+ \tag{5.10}$$

$$A = I - (\theta\Lambda\theta^T)(\theta\Lambda\theta^T)^+ \tag{5.11}$$

An alternative expression for K is

$$K = [\Lambda\theta^T(\theta P_m\theta^T+R_2)^{-1}\theta\Lambda]^+\Lambda\theta^T(\theta P_m\theta^T+R_2)^{-1} +$$

$$+ P_m\theta^T[A(\theta P_m\theta^T+R_2)A]^+ \tag{5.10a}$$

$P_m(t)$, the covariance of $\tilde{x}_m(t|t-1)$

$$\tilde{x}_m(t|t-1) = \tilde{x}_1(t|t-1) + \tilde{z}_m(t|t-1) \tag{5.12}$$

with $\tilde{z}_m$ from (5.6) satisfies

$$P_m(t+1) = R_1 + \phi[(I-K\theta)P_m(I-K\theta)^T + KR_2K^T]\phi^T$$

$$P_m(t_0) = R_0^S \tag{5.13}$$

and Λ, the transformed projection on the unobservable part of the unknown initial value satisfies

$$\Lambda(t+1) = \phi[\Lambda - \Lambda\theta^T(\theta\Lambda\theta^T)^+\theta\Lambda]\phi^T \qquad \Lambda(t_0) = N^TN \tag{5.14}$$

Proof:

$$\hat{x}_m(t|t) = \hat{x}_\Pi(t|t) + \hat{z}_m(t|t) = \hat{x}_\Pi(t|t-1) + \hat{z}_m(t|t-1) +$$

$$+ K_\Pi[y-\theta\hat{x}_\Pi] + (I-K_\Pi\theta)K_z(y-\theta\hat{x}_\Pi-\theta\hat{z}_m) =$$

$$= \hat{x}(t|t-1) + K(t)[y(t) - \theta\hat{x}(t|t-1)]$$

with $K = K_\Pi + (I-K_\Pi\theta)K_z$. Note that $(I-K\theta) = (I-K_\Pi\theta)(I-K_z\theta)$.

The covariance P_m of $\tilde{x}_m$ is $P_m = \Pi + \Sigma_m$ so that by (5.5) and (3.5)

$$P_m(t+1) = R_1 + \phi\{(I-K_\Pi\theta)\Pi + (I-K_\Pi\theta)(I-K_z\theta)\Sigma_m(I-K_z\theta)^T \cdot$$

$$\cdot (I-K_\Pi\theta)^T + (I-K_\Pi\theta)K_z(R_2+\theta\Pi\theta^T)K_z^T(I-K_\Pi\theta)^T\}\theta =$$

$$= R_1 + \phi\{(I-K\theta)P_m(I-K\theta)^T + KR_2K^T\}\phi$$

Since $(I-K\theta)\Pi - (I-K\theta)\Pi(I-K\theta)^T + (K-K_\Pi)(R_2+\theta\Pi\theta^T)(K-K_\Pi)^T - KR_2K^T = 0$ which follows from:

$$(I-K\theta)\Pi(I-K\theta)^T = \Pi - K\theta\Pi - \Pi\theta^TK^T + K\theta\Pi\theta^TK^T$$

$$(K-K_\Pi)(R_2+\theta\Pi\theta^T)(K-K_\Pi)^T = KR_2K^T + K\theta\Pi\theta^TK^T - (K-K_\Pi) \cdot$$

$$\cdot (R_2+\theta\Pi\theta^T)K_\Pi^T - K_\Pi(R_2+\theta\Pi\theta^T)K^T =$$

$$= KR_2K^T + K\theta\Pi\theta^TK^T - (K-K_\Pi)\theta\Pi - \Pi\theta^TK^T$$

Λ defined by (5.4) also fulfils (5.14), since

$$\theta\left(\Lambda - \Lambda\theta(\theta\Lambda\theta^T)^+\theta\Lambda\right) = 0.$$

It remains to prove that $K = K_\Pi + (I-K_\Pi\theta)K_Z$ can be evaluated by (5.10). Then (5.10a) follows like in Theorem 3. Denote $R_2 + \theta\Pi\theta^T + \theta\Sigma_m\theta^T$ by R:

$$K = K_Z + K_\Pi(I-\theta K_Z) = \Sigma_m\theta^T[ARA]^+ + \Lambda\theta^T(\theta\Lambda\theta^T)^+\{I - R[ARA]^+\} +$$

$$+ K_\Pi\{I - \theta\Sigma_m\theta^T[ARA]^+\} - K_\Pi\theta\Lambda\theta^T(\theta\Lambda\theta^T)^+[I - R(ARA)^+] =$$

$$= (\Pi+\Sigma_m)\theta^T[ARA]^+ + \Lambda\theta^T(\theta\Lambda\theta^T)^+\{I - R[ARA]^+\} +$$

$$+ K_\Pi\{I - [(\theta\Pi\theta^T+R_2) + \theta\Sigma_m\theta^T][ARA]^+\} + K_\Pi[I - (\theta\Lambda\theta^T) \cdot$$

$$\cdot (\theta\Lambda\theta^T)^+]\{I - R[ARA]^+\} - K_\Pi\{I - R(ARA)^+\} =$$

$$= P_m\theta^T[ARA]^+ + \Lambda\theta^T(\theta\Lambda\theta^T)^+\{I - R[ARA]^+\}$$

since $A\left(I - R(ARA)^+\right) = 0$ □

Comments: In order to get a correct interpretation of Λ in (5.14) and (5.4) it must be noticed that the unobservable subspace is the same for the z-system of Theorem 2 and the original system (2.1), so both $I-MM^+$ and Λ are the same! Thus if N^TN is a projection.

$$\Lambda(t) = \phi(t,t_0)N^TN(I-M_tM_t^+)N^TN\phi^T(t,t_0) \tag{5.15}$$

$$M_t = \sum_{s=t_0}^{t-1} \phi^T(s,t_0)\theta^T\theta\phi(s,t_0) \tag{5.16}$$

In the same way with $\hat{z}_m(t|t-1) = \psi(t,t_0)N^T\hat{\xi}_m(t-1)$ and with $\tilde{x}_m = \tilde{x}_1 + \tilde{z}_m$ the bias term can be written

$$E(x-\hat{x}) = \tilde{x}_{mb}(t|t-1) = x(t) - \hat{x}_m(t|t-1) - \tilde{x}_m(t|t-1) =$$

$$= z - \hat{z}_m - \tilde{z}_m = \psi(t,t_0)N^TN(I-M_tM_t^+)N^T\xi =$$

$$= \phi(t,t_0)N^TN(I-M_tM_t^+)N^T\xi \qquad (5.17)$$

so that Λ spans the bias. P_m is the covariance of $x - \hat{x}_m$ since

$$P_m = E(x-\hat{x}-\tilde{x}_{mb})^2$$

6. LINEAR STOCHASTIC CONTROL.

An application of the above filter results is the control of a linear stochastic system for which a quadratic loss function:

$$J = E \sum_{t=t_0}^{N-1} \left\{ x^T(t) Q_1 x(t) + u^T(t) Q_2 u(t) \right\} + x^T(N) Q_0 x(N) \quad (6.1)$$

should be minimized with respect to $u(t_0),\ldots,u(N)$ subject to the constraint

$$\begin{cases} x(t+1) = \phi x(t) + \Gamma u(t) + v(t), \quad x(t_0) = x_0 = x_0^S + N^T \xi \\ \\ y(t) = \theta x(t) + R(t) \end{cases} \quad (6.2)$$

with v, e, x_0 defined in Section 2. The expectation in (6.1) is taken with respect to the introduced statistics v, e and x_0^S. The choice of $u(t)$ is restricted to linear maps of Y_{t-1}.

Rewrite J as in [2] using S and L defined by

$$L(t) = \left(Q_2 + \Gamma^T S(t+1)\Gamma\right)^{-1} \Gamma^T S(t+1)\phi \quad (6.3)$$

$$S(t) = \phi^T S(t+1)\phi - \phi^T S(t+1)\Gamma L(t) + Q_1, \quad S(N) = Q_0 \quad (6.4)$$

$$J = E x_0^T S(t_0) x_0 + E \sum_{t=t_0}^{N-1} \left\{ v^T(t) S(t+1) v(t) + \left(u(t) + L(t)x(t)\right)^T \cdot \right.$$

$$\left. \cdot \left(Q_2 + \Gamma^T S(t+1)\Gamma\right)\left(u(t) + L(t)x(t)\right) \right\} =$$

$$= \xi^T N S(t_0) N^T \xi + \operatorname{tr} R_0^S S(t_0) + \sum_{t=t_0}^{N-1} \operatorname{tr} R_1(t) S(t+1) + E \sum_{t=t_0}^{N-1} (u+Lx)^T \cdot$$

$$\cdot (Q_2 + \Gamma^T S \Gamma)(u+Lx) \quad (6.5)$$

and the minimal T is

$$T^0 = E\| \tilde{x}_q + \tilde{x}_{qb} \|_q^2 = \| \tilde{x}_{qb} \|_q^2 + \mathrm{tr}(q \cdot \mathrm{cov}\, \tilde{x}_q) =$$

$$= \| \tilde{x}_{qb} \|_q^2 + \mathrm{tr}\left\{P_q L^T (Q_2 + \Gamma^T S \Gamma) L\right\}$$

Now $\tilde{x}_q(t|t-1)$ and also $\tilde{x}_{qb}(t|t-1)$ do not change with different $u(s)$, $s < t$, restricted to linear functions of Y_{s-1}, a fact that is fairly easy to show using for instance innovations. Thus the last sum in (6.5) is minimized, starting with the term T_{N-1}, by u from (6.9). The minimal J is thus

$$J^0 = \xi^T N S(t_0) N^T \xi + \mathrm{tr}\, R_0^S S(t_0) + \sum_{t=t_0}^{N-1} \mathrm{tr}\, R_1(t) S(t+1) +$$

$$+ \sum_{t=t_0}^{N-1} T_t^0$$

This concludes the proof of the following theorem.

Theorem 5: The loss function (6.1) for the system (6.2) is minimized for worst possible ξ by the input

$$u(t) = - L(t) \hat{x}_q(t|t-1) \qquad (6.10)$$

in the class of linear functions of Y_{t-1}, (6.7). L(t) is defined by (6.3), (6.4), and $\hat{x}_q$ is the minimal variance, minimal bias linear estimate of x(t) with the error measured by the matrix q in (6.8). The minimal loss, which depends on the unknown constant ξ, is

$$J^0 = \xi^T N S(t_0) N^T \xi + \operatorname{tr} R_0^S S(t_0) + \sum_{t=t_0}^{N-1} \operatorname{tr} R_1(t) S(t+1) +$$

$$+ \sum_{t=t_0}^{N-1} \operatorname{tr} P_q L^T (Q_2 + \Gamma^T S \Gamma) L + \sum_{t=t_0}^{N-1} \| \tilde{x}_{q_b} \|_q^2 \qquad (6.11)$$

Comments: The bias terms $\tilde{x}_{qb}$ will be zero as soon as ξ is observable. The restriction (6.7) on u is rather natural since all the random variables have zero mean value, but it can be argued that it implies that possible values of ξ are assumed to be centered around zero. If a bias term is allowed in u, it is no longer meaningful to minimize the bias of $\hat{x}$. The estimate $\hat{x}_m$ of Theorem 1 would be as good as any $\hat{x}_q$, since the estimates will become unbiased at the same time. Since q is time varying there seems to be no hope to obtain simple general recursive formulas for $\hat{x}_q$.

$$J^0 = \xi^T N S(t_0) N^T \xi + \operatorname{tr} R_0^S S(t_0) + \sum_{t=t_0}^{N-1} \operatorname{tr} R_1(t) S(t+1) +$$

$$+ \sum_{t=t_0}^{N-1} \operatorname{tr} P_q L^T (Q_2 + r^T S r) L + \sum_{t=t_0}^{N-1} \| \tilde{x}_{q_b} \|_q^2 \qquad (6.11)$$

Comments: The bias terms $\tilde{x}_{qb}$ will be zero as soon as ξ is observable. The restriction (6.7) on u is rather natural since all the random variables have zero mean value, but it can be argued that it implies that possible values of ξ are assumed to be centered around zero. If a bias term is allowed in u, it is no longer meaningful to minimize the bias of $\hat{x}$. The estimate $\hat{x}_m$ of Theorem 1 would be as good as any $\hat{x}_q$, since the estimates will become unbiased at the same time. Since q is time varying there seems to be no hope to obtain simple general recursive formulas for $\hat{x}_q$.

7. CONTINUOUS TIME, DUALITY.

In the continuous time case it is not possible to obtain recursive estimates by letting the covariance increase like in Section 2 or by some pseudo inverse formulas like in Sections 4 and 5. The minimal unbiased estimate will be obtained by duality. Consider the system

$$\begin{cases} \dot{x} = Ax + v \\ y = Cx + e \end{cases} \qquad x(t_0) = x_0 = x_0^S + x_0^N \qquad (7.1)$$

where v and e are uncorrelated white noise with covariances $R_1\delta(t)$ and $R_2\delta(t)$, $R_2 > 0$. x_0^S is uncorrelated with e and v, has zero mean value and covariance R_0^S. The only thing known about x_0^N is that it is restricted to a subspace spanned by the full column rank rectangular matrix N^T:

$$x_0^N = N^T\xi, \quad \xi \text{ arbitrary}$$

It is a celebrated fact that the filter problem is the dual of an optimization problem [2, 5, 9] and that was used for the proof of the continuous time filter problem in [2]. There is also a well-known duality between observability and controllability, i.e. reconstruction or unbiased estimates and fixed end-point problems [2, 6, 9].

These two dualities will here be shown to combine.

Consider the minimal variance unbiased estimate for the system (7.1). It is convenient first to estimate an arbitrary linear combination of x(t), say $a^Tx(t)$. Thus the variance

$$V = E[a^T x(t) - a^T \hat{x}(t)]^2 \tag{7.2}$$

should be minimized for linear functions of Y_t, say

$$a^T \hat{x}(t) = - \int_0^t u^T(s) y(s) ds \tag{7.3}$$

under the constraint of being unbiased

$$E a^T \hat{x}(t) = a^T E x(t) \tag{7.4}$$

The notation $a^T \hat{x}(t)$ will be justified later. There is really an $\hat{x}(t)$ such that $a^T \hat{x}(t)$ is the best estimate for all a.

Theorem 6: The unbiased filter estimate problem for (7.1) is dual to the restricted end-point optimization problem, with the loss function

$$V = z^T(t_0) R_0^S z(t_0) + \int_{t_0}^t [z^T(s) R_1 z(s) + u^T(s) R_2 u(s)] ds \tag{7.5}$$

and the constraints

$$\begin{cases} -\dot{z} = A^T z + C^T u, & z(t) = a \qquad (7.6) \\ N z(t_0) = 0 \qquad (7.7) \end{cases}$$

Proof: Use z defined by (7.6), $a^T \hat{x}$ from (7.3) and the system equation (7.1) in the same way as in [2] to rewrite the estimation error

$$a^T x(t) - a^T \hat{x}(t) = z^T(t_0) x(t_0) + \int_{t_0}^t z^T(s) v(s) ds + \int_{t_0}^t u^T(s) e(s) ds$$

In order to fulfil (7.4)

$$E z^T(t_0)x(t_0) = z^T(t_0)N^T\xi = \big(Nz(t_0)\big)^T\xi = 0$$

for all ξ, which is equivalent to (7.7). V can thus be expressed as

$$V = E[a^Tx(t) - a^T\hat{x}(t)]^2 =$$

$$= z^T(t_0)R_0^Sz(t_0) + \int_{t_0}^{t} z^T(s)R_1z(s) + u^T(s)R_2u(s)\,ds$$

using the covariances of x_0^S, v and e.

□

The solution of the optimization is well-known, cf. [6, 9], but requires controllability of the restricted end-state, i.e. observability of $N^T\xi$.

$$u(s) = -R_2^{-1}C\big[\Pi(s)z(s) + \psi(s,t_0)N^T\big(NM(t_0,t)N^T\big)^{-1}N\psi(t,t_0)a\big] \tag{7.8}$$

$$\dot{\Pi} = A\Pi + \Pi A^T + R_1 - \Pi C^TR_2^{-1}C\Pi,\ \Pi(t_0) = R_0^S \tag{7.9}$$

$$\frac{d}{dt}\psi(t,s) = \big(A - \Pi(t)C^TR_2^{-1}C\big)\psi(t,s),\ \psi(s,s) = I \tag{7.10}$$

$$M(t_0,t) = \int_{t_0}^{t}\psi^T(s,t_0)C^TR_2^{-1}C\psi(s,t_0)ds \tag{7.11}$$

M is an observability Gramian for (7.1). Now solve (7.6), (7.8) for z giving $\hat{x}$ independent of a:

$$z(s) = \Big[\psi^T(t,s) - M(s,t)\psi(t,t_0)N^T\big(NM(t_0,t)N^T\big)^{-1}N\psi^T(t,t_0)\Big]a$$

$$\hat{x}(t) = \int_0^t\Big[\psi(t,s)\Pi(s) - \psi(t,t_0)N^T\big(NM(t_0,t)N^T\big)^{-1}\cdot$$

$$\cdot N\psi^T(t,t_0)M(s,t)\Pi(s) + \psi(t,t_0)N^T\big(NM(t_0,t)N^T\big)^{-1}\cdot$$

$$\cdot N\psi^T(t,t_0)\Big]C^TR_2^{-1}y(s)ds \tag{7.12}$$

Note that the minimal V is

$$a^T[\Pi(t) + \psi(t,t_0)N^T(NM(t_0,t)N^T)^{-1}N\psi^T(t,t_0)]a =$$

$$= a^T[\Pi(t) + \Sigma(t)]a$$

defining the error covariance $P(t) = \Pi(t) + \Sigma(t)$.

It is now possible to state the theorem

Theorem 7: The best linear unbiased state estimate for (7.1) is

$$\hat{x}(t|t) = \hat{x}_\Pi(t|t) + \psi(t,t_0)N^T(NM(t_0,t)N^T)^{-1}N\lambda(t_0,t) \qquad (7.13)$$

with ψ, M and Π from (3.9), (3.10) and (3.8) and $\hat{x}_\Pi$ and λ from

$$\frac{d}{dt}\hat{x}_\Pi(t|t) = A\hat{x}_\Pi(t|t) + \Pi(t)C^TR_2^{-1}[y(t) - C\hat{x}_\Pi(t|t)],$$

$$\hat{x}_\Pi(t_0|t_0) = 0 \qquad (7.14)$$

$$\begin{cases} -\frac{d}{ds}\lambda(s,t) = (A - \Pi(s)C^TR_2^{-1}C)\lambda(s,t) + \\ \qquad + C^TR_2^{-1}[y(s) - C\hat{x}_\Pi(s|s)] \\ \lambda(t,t) = 0 \end{cases} \qquad (7.15)$$

The error covariance is $P(t) = \Pi(t) + \Sigma(t)$.

Proof: Integrate (7.14) and (7.15)

$$\hat{x}_\Pi(t|t) = \int_{t_0}^{t} \psi(t,s)\Pi(s)C^TR_2^{-1}y(s)ds$$

$$\lambda(s,t) = \int_s^t \psi^T(q,s)C^T R_2^{-1}[y(q) - C\hat{x}_\Pi(q|q)]dq$$

and combine (7.12) with

$$\int_{t_0}^t \psi^T(s,t_0)M(s,t)\Pi(s)C^T R_2^{-1} y(s)ds =$$

$$= \int_{t_0}^t \psi^T(q,t_0)C^T R_2^{-1} C\left[\int_{t_0}^q \psi(q,s)\Pi(s)C^T R_2^{-1} y(s)ds\right]dq \qquad \square$$

Example 7.1: $A = 0$, $R_1 = \sigma^2$, $R_2 = 1$, $x(0) = \xi$ unknown, $C = 1$.

$$\dot{\Pi} = \sigma^2 - \Pi^2, \quad \Pi(0) = 0 \quad \Rightarrow \quad \Pi(t) = \sigma \tanh \sigma t$$

$$\psi(t,s) = \exp\left\{\int_s^t -\sigma \tanh \sigma q \, dq\right\} = \cosh \sigma s / \cosh \sigma t$$

$$\Sigma(t) = \frac{1}{\cosh^2 \sigma t} \Big/ \int_0^t \frac{ds}{\cosh^2 \sigma s} = \sigma/(\sinh \sigma t \cdot \cosh \sigma t) \quad t > 0$$

$$P(t) = \Pi(t) + \Sigma(t) = \sigma \coth \sigma t \quad t > 0$$

$$\hat{x}_\Pi(t|t) = \sigma \int_0^t \frac{\sinh \sigma s}{\cosh \sigma s} y(s)ds$$

$$\hat{x}(t|t) = \hat{x}_\Pi(t|t) + \frac{\sigma}{\sinh \sigma t} \int_0^t \frac{y(s) - \hat{x}_\Pi(s|s)}{\cosh \sigma s} ds =$$

$$= \frac{\sigma}{\sinh \sigma t} \int_0^t \cosh \sigma s \, y(s)ds \quad t > 0$$

This could be obtained analytically using P(t)

$$\hat{x}(t|t) = \int_0^t \exp\left[-\sigma \int_s^t \coth \sigma q dq\right] \sigma \coth \sigma s y(s) ds =$$

$$= \frac{\sigma}{\sinh \sigma t} \int_0^t \cosh \sigma s y(s) ds$$

but to use a numerical algorithm for the Riccati equation started with a large variance would lead to enormous difficulties.

Example 7.2: $A = -a$, $C = 1$, $R_1 = 0$, $R_2 = 1$

$$\Pi(t) = 0, \quad \psi(t,s) = \phi(t,s) = \exp\{-a(t-s)\}$$

$$P(t) = \Sigma(t) = e^{-2at} / \int_0^t e^{-2as} ds = 2a\, e^{-2at}/(1-e^{-2at}) \quad t > 0$$

$$\hat{x}(t|t) = 2a\, e^{-at} \int_0^t e^{-as} y(s) ds/(1-e^{-2at}) \quad t > 0$$

The initial estimate in both the examples is

$$\lim_{t \to 0^+} \hat{x}(t|t) = y(0)$$

which has infinite covariance.

Comments: The best unbiased estimate obtained by Theorem 7 is a sum of two estimates, similar to the discrete time case. The last term consists of a transformed smoothing estimate of the initial constant ξ.

It must be emphasized that it is not possible to obtain the estimate from a recursive filter. From (7.13) and the examples it can be seen what happens at the initial point. The estimate may very well exist but with infinite variance. For time invariant systems the whole state becomes observable after an infinitely short time, so even if a pseudo inverse is used instead of P, it is infinite and the gain required in the differential equation at $t = t_0$ is infinite. The discontinuity in $\hat{x}$ and P must be calculated in some other way, preferably from Theorem 7, if a Kalman filter is to be started at $t = t_0^+$.

However, any observability index is very small at t_0^+ and P is very large. It is suggested that the Kalman filter should not be started until a little later for numerical reasons. In fact, $\lambda(t_0,t)$ of (7.15) can be obtained by integration in the forward direction:

$$\frac{d}{dt}\lambda(t_0,t) = \psi^T(t,t_0)C^TR_2^{-1}[y(t) - C\hat{x}_{II}(t|t)],$$

$$\lambda(t_0,t_0) = 0 \qquad (7.16)$$

The differential equation is not asymptotically stable, but it does not matter since it is only needed a short time. Thus $\hat{x}$ can be obtained by (7.13) from $\hat{x}_{II}$ and λ, until P(t) is small enough to start the Kalman filter.

It should also be noted, that it is impossible to get <u>any</u> unbiased state estimate with a constant gain Kalman filter in the continuous time case. "Dead-beat" filters must have time varying gain. Compare the dual problem with dead-beat controllers.

8. CONCLUSIONS.

It has been shown how the Kalman filter should be started if part of the initial state is totally unknown. The formulas can be obtained formally by letting the covariance of that part go to infinity. The common way of starting a Kalman filter with a very large covariance is thus almost optimal, but the numerical properties of such a solution are bad.

The optimal discrete time solution uses two "Riccati equations", one to keep track of the bias from the unknown parts of the initial state, and one for the error covariance.

The interpretation of the estimate is provided by a separation into two filters. Before it is possible to obtain an unbiased estimate, the obtained estimate minimizes the mean square error of the unknown initial value in the Euclidian norm. The linear stochastic regulator problem is, however, shown to require an estimate that minimizes the error in another norm. The minimum of the loss function will contain additional terms containing the unknown initial value.

The continuous time case is more complicated. The whole system becomes observable at once. After an initial discontinuity in estimate and covariance a usual Kalman filter could be started. This, however, implies almost infinite Kalman gain and poor numerical properties. It is suggested that the estimate is calculated using separation into two estimates. The "noise term" satisfies a simple filter of "Kalman structure", while the "bias term" should be calculated from a recursively updated smoothing estimate of the unknown initial state. When the error covariance has become reasonable, the usual Kalman filter should be started.

APPENDIX

Although pseudo inverses are conceptually simple some of the necessary algebra tends to obscure the ideas. Some lemmas will be shown here to simplify the proof of Theorem 3. A good general reference is [1].

Lemma 1: Let $\bar{M}$ be a symmetric matrix and c a rectangular matrix, let D and G be the projections $D = I - \bar{M}^{+}\bar{M}$, $G = I - (Dc)^{+}Dc$, and let r be an invertible matrix, then the projection $A = I - r^{-1}c^{T}Dcr^{-T}(r^{-1}c^{T}Dcr^{-T})^{+}$ can be obtained by

$$A = r^{T}Gr(r^{T}Gr)^{+} \tag{A.1}$$

Proof: If the matrix X spans the subspace orthogonal to that spanned by $c^{T}D$, i.e. $Gx = x$, then $r^{T}x \perp r^{-1}c^{T}D$. $R(A) = R(r^{T}x) = R(r^{T}G)$ so that $r^{T}Gr(r^{T}Gr)^{+}$ is the unique orthogonal projection A. □

Lemma 2: With r, D, G, c, $\bar{M}$ and A as above and with $B = [I+Gc^{T}\bar{M}^{+}cG]^{-1}$

$$r^{T}GBGr = [A(VV^{T}+R_{2})A]^{+} \tag{A.2}$$

if $r^{T}r = R_{2}^{-1}$ and $r^{-1}c^{T}\bar{M}^{+}cr^{-T} = VV^{T}$

Proof:

$$r^{T}GBGr = r^{T}Gr[R_{2}^{-1} + r^{T}GrVV^{T}r^{T}Gr]^{-1}r^{T}Gr =$$

$$= r^{T}Gr[R_{2} - R_{2}r^{T}GrV(I+V^{T}r^{T}GrR_{2}r^{T}GrV)^{-1}V^{T}r^{T}GrR_{2}]r^{T}Gr =$$

$$= r^{T}Gr - r^{T}GrV(I+V^{T}r^{T}GrV)^{-1}V^{T}r^{T}Gr$$

by the inverse lemma and the fact that G is a projection. With $F = r^T G r$, the right hand side of (A.2) gives using Lemma 1 and the pseudo inverse lemma:

$$[A(VV^T+R_2)A]^+ = [F^+FVV^TFF^+ + F^+FF^+]^+ =$$

$$= F - FF^+FV(I+V^TFF^+FF^+FV)^{-1}V^TFF^+F =$$

$$= F - FV(I+V^TFV)^{-1}V^TF$$

Note the simple form of the pseudo inverse lemma since R_2 invertible and thus $AV \in R(AR_2)$.

□

Lemma 3: With c, D, G, A and r as above

$$G = r^{-T}Ar^{-1}(r^{-T}Ar^{-1})^+ = r^{-T}(Ar^{-1}r^{-T}A)^+r^{-1} \qquad (A.3)$$

$$(c^TDc)^+ = (I-G)r^{-T}(r^{-1}c^TDcr^{-T})^+r^{-1}(I-G) \qquad (A.4)$$

Proof: The first equality of (A.3) is the inversion of Lemma 1, and the second follows from $Ar^{-1}(r^{-T}AAr^{-1})^+ = (r^{-T}A)^+ = (Ar^{-1}r^{-T}A)^+Ar^{-1}$. Direct verification of the Moore Penrose conditions, $B = A^+$ if AB and BA symmetric and ABA = A, BAB = B, can be used to show (A.4):

$$c^TDc(I-G)r^{-T}(r^{-1}c^TDcr^{-T})^+r^{-1}(I-G) =$$

$$= r(r^{-1}c^TDcr^{-T})(r^{-1}c^TDcr^{-T})^+r^{-1}(I-G) =$$

$$= r(I-A)r^{-1}(I-G) = I - G - rr^T(r^{-T}Ar^{-1})(I-G) = I - G$$

which is symmetric and $(I-G)c^TDc = c^TDc$, $(I-G)r^{-T} \cdot (r^{-1}c^TDcr^{-T})^+r^{-1}(I-G)^2 = (I-G)r^{-T}(r^{-1}c^TDcr^{-T})^+r^{-1}(I-G)$.

□

Lemma 4: The matrix K' defined by (4.13) can be expressed as

$$K' = \left\{Dc(c^TDc)^+[I - (I+c^T\bar{M}^+c)GBG] + \bar{M}^+cGBG\right\}r =$$

$$= Dcr^{-T}(r^{-1}c^TDcr^{-T})^+\left\{I - (R_2+VV^T)[A(R_2+VV^T)A]^+\right\} +$$

$$+ \bar{M}^+cr^{-T}[A(R_2+VV^T)A]^+ \qquad (A.5)$$

in quantities defined above.

Proof: Since DcG = 0, (A.4) and (A.2) give

$$Dc(c^TDc)^+[I - (I+c^T\bar{M}^+c)GBG]r =$$

$$= Dcr^{-T}(r^{-1}c^TDcr^{-T})^+(I-r^{-1}Gr)\{I - (R_2+VV^T)[A(R_2+VV^T)A]^+\}$$

From (A.3) $r^{-1}Gr = (AR_2A)^+$, and since R_2 has full rank $AV \in R(AR_2)$ so that

$$(AR_2A)^+\{I - (R_2+VV^T)[AR_2A+AVV^TA]^+\} =$$

$$= (AR_2A)^+\{I - A(R_2+VV^T)A[AR_2A+AVV^TA]^+\} = 0$$

which proves (A.5).

□

REFERENCES.

[1] Albert, A.: Regression and the Moore-Penrose Pseudo Inverse, Academic Press, New York, 1972.

[2] Åström, K.J.: Introduction to Stochastic Control Theory, Academic Press, New York, 1970.

[3] Cline, R.E.: Representation for the Generalized Inverse of Sums of Matrices. Siam. J. Numer. Anal., Serie B, 2, 99 - 114 (1965).

[4] Hagander, P.: The Use of Operator Factorization for Linear Control and Estimation, Automatica, 9, 623 - 631 (1973).

[5] Kalman, R.E.: New Methods in Wiener Filtering Theory, in Proceedings of First Symp. on Eng. Appl. of Random Function Theory and Probability, J.L. Bogdanoff and F. Kozin (eds.), Wiley, New York, 1963.

[6] Kalman, R.E., Ho, Y.-C, Narendra, K.S.: Controllability of Linear Dynamical Systems, Contributions to Differential Equations, 1, 189-213 (1961).

[7] Luenberger, D.G.: Optimization by Vector Space Methods, John Wiley, New York, 1969.

[8] Rao, C.R., Mitra, S.K.: Generalized Inverse of Matrices and its Applications, John Wiley, New York, 1971.

[9] Sorenson, H.W.: Controllability and Observability of Linear, Stochastic, Time-Discrete Control Systems, in Advances in Control Systems, C.T. Leondes (ed.), Academic Press, New York, 1968.

Numerical Solution of $A^TS + SA + Q = 0$†

PER HAGANDER
Division of Automatic Control, Lund Institute of Technology, Lund, Sweden

Communicated by K.-J. Åström

ABSTRACT

A survey of techniques to solve $A^TS + SA + Q = 0$ is presented, and nine algorithms are coded and tested on a batch of examples. Which algorithm to be recommended depends mainly on the order of the system.

1. INTRODUCTION

In recent time [2,4,5,6,9,13,14,19,20,21] great attention has been drawn to the equation

$$A^TS + SA + Q = 0, \tag{1}$$

solved for S with Q symmetric of order $n \times n$ and thus also the solution S.

This equation plays a central role in the theory of stability for linear continuous systems. It also arises in pole assignment [11], in sensitivity analysis [2d], and when evaluating loss functions in optimal control and covariance matrices in filtering and estimation for continuous systems.

The more general equation

$$A^TS + SA + Q_1 - SBQ_2^{-1}B^TS = 0, \tag{2}$$

appearing in spectral factorization [1], filtering and optimal control, has been solved by iteration of (1) [10].

Another generalization of (1) used in, e.g., network theory,

$$A^TS + SB + Q = 0, \tag{3}$$

is possible to solve by slight modifications of some of the methods indicated below.

† This work has been supported by the Swedish Board of Technical Development under contract 69-631/U489.

The equation corresponding to (1) for discrete time systems,

$$\phi^T S\phi + \tilde{Q} = S, \tag{4}$$

has somewhat different properties and use has been made of transformations between the two equations.

In many of the applications above it is necessary to solve (1) many times. The equation is therefore worth severe numerical interest. It is my intention to survey possible methods and to present algorithms. These algorithms are coded and tested on a set of examples and their accuracy and computing time are compared. All programming was done in FORTRAN for Univac 1108. The algorithms are grouped in three sections:

(1) *Direct Methods.* Solution of a large system of linear equations by general methods.
(2) *Transformation methods.* Use of the structure can be made by similarity transformation of the A-matrix to some canonical form (Jordan or diagonal form, companion form, Schwarz' form). Some algorithms use the same technique without explicitly performing the transformations.
(3) *Iterative methods.* The basic idea for these methods is that equation (1) is transformed to equation (4) either by sampling or by introducing a bilinear transformation. The equation

$$S_{k+1} = \phi^T S_k \phi + \tilde{Q} \tag{5}$$

is then iterated to stationarity by an accelerating formula.

2. COMPUTATIONAL AND PROGRAMMING ASPECTS

2.1. *Direct Methods*

2.1.1. Equation (1) has $n(n+1)/2$ unknown variables. By organizing S and Q as vectors the system is rewritten as common linear equations

$$\mathscr{A}s = q \tag{6}$$

and this can be solved by general methods, like Gauss elimination. $\mathscr{A}$ can be formed from A by use of either logical operations [5] or an indexing matrix [4] or vector. The indexing vector form is found slightly more efficient.

If λ_i are the eigenvalues of A, then the eigenvalues of $\mathscr{A}$ are sums $\lambda_i + \lambda_j$. This implies that (6) certainly has a unique solution if A is stable. If A is unstable, $\mathscr{A}$ might be ill-conditioned or singular. The original equation (1) is, however, then also ill-conditioned or singular. The equation can be solved for different Q-matrices (q-vectors) with little extra effort without inversion of $\mathscr{A}$ [8]. This can, for instance, be valuable when improving a solution.

Algorithm 1 utilizes these ideas. The programming is easily done. The main disadvantages are that the memory requirement is $(n(n+1)/2)^2$ cells, and that the number of multiplicative operations for large n is of the order $n^6/24$.

2.2 *Transformation Methods*

2.2.1. By transforming A and A^T to Jordan form it is possible to express the solution S in the eigenvalues and eigenvectors of A and A^T [12]. This is greatly simplified when A is diagonalizable:

THEOREM 1 [12]. *Let U and V be the matrices that diagonalize A and A^T:*

$$A = U^{-1}DU, \qquad A^T = V^{-1}DV, \qquad D = \operatorname{diag}\{\lambda_1, \ldots, \lambda_n\},$$

and let $\hat{Q} = VQU^{-1}$ and $\hat{s}_{ij} = -q_{ij}/(\lambda_i + \lambda_j)$. Then

$$S = V^{-1}\hat{S}U. \tag{7}$$

This theorem is used in algorithm 2, thus requiring complex arithmetic. The main drawback is, however, the eigenvalue eigenvector calculation for nonsymmetric A. This is done in an up-to-date QR-algorithm with inverse iteration, but close eigenvalues lead to overwhelming problems. No advantage of symmetry can be taken. The approach is out of the question in other cases than when eigenvalues and eigenvectors are already obtained or wanted.

2.2.2. Eigenvectors can also be used in another way applicable even to the quadratic equation (2) [15, 18].

THEOREM 2 [18]. *If*

$$\begin{bmatrix} b_1 \\ c_1 \end{bmatrix}, \ldots, \begin{bmatrix} b_n \\ c_n \end{bmatrix}$$

are the n eigenvectors corresponding to the eigenvalues with negative real part of the $2n \times 2n$ matrix

$$\begin{bmatrix} A & -BQ_2^{-1}B^T \\ -Q_1 & -A^T \end{bmatrix} \tag{8}$$

then the solution of (2) *is*

$$S = [c_1, \ldots, c_n][b_1, \ldots, b_n]^{-1}. \tag{9}$$

The computational effort is simplified by the observation:

COROLLARY. *Let*

$$\begin{bmatrix} f_i \\ g_i \end{bmatrix} = \begin{bmatrix} b_i \\ c_i \end{bmatrix}$$

for real eigenvalues λ_i *and*

$$\begin{bmatrix} f_i \\ g_i \end{bmatrix} = \mathrm{Re}\begin{bmatrix} b_i \\ c_i \end{bmatrix} + \mathrm{Im}\begin{bmatrix} b_i \\ c_i \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} f_{i+1} \\ g_{i+1} \end{bmatrix} = \mathrm{Re}\begin{bmatrix} b_i \\ c_i \end{bmatrix} - \mathrm{Im}\begin{bmatrix} b_i \\ c_i \end{bmatrix}$$

for pairs of complex eigenvalues λ_i, λ_{i+1}, $(\lambda_{i+1} = \bar{\lambda}_i)$, *then*

$$S = [g_1, \ldots, g_n]\,[f_1, \ldots, f_n]^{-1}. \tag{10}$$

Algorithm 3 is based on this corollary. Real arithmetic can be used but otherwise the remarks for algorithm 2 still hold. A new Q needs a recomputation of the whole eigenvalue problem. The method is only of theoretical interest for equation (1) but is reasonable for equation (2).

2.2.3. The companion form and its transformations lead to interesting algebraic manipulations and probably also to the fewest operations for large dimensions. One method, emanating from Nekolny and Benes [16], deals with the transfer function (G) of the system $S(A^T,B,C)$. The covariance of the output of the system for white noise input is

$$V = CSC^T \tag{11}$$

if S is the solution of (1) with

$$Q = BB^T. \tag{12}$$

Åström [23] has described this for single input, single output systems and gives recursive formulas essentially using the Routh algorithm. These can be extended to the multivariable case and used for solution of (1).

This is done in algorithm 4 including decomposition of Q [8], and computation of $G(s)$ for $S(A^T,B,I)$ by a Leverrier algorithm [7].

Full advantage of symmetry is taken, but different Q matrices are difficult to handle.

2.2.4. Other authors have used the companion form to obtain an explicit solution without performing the transformation. Smith [21a] developed an expression in powers of A, and Müller [14] used the matrices A_i from the Leverrier algorithm and achieved a nice formula:

THEOREM 3 [14]. *Let*

$$\begin{aligned} a_k &= -\frac{1}{k}\,\mathrm{tr}\,AA_{k-1}, & a_0 &= 1,\ A_0 = I, \\ A_k &= AA_{k-1} + a_kI, & k &= 1, \ldots, \end{aligned} \tag{13}$$

and let c be the solution of

$$Hc = \begin{bmatrix} 1/2 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{14}$$

where H is the Hurwitz matrix, then

$$S = \sum_{j=0}^{n-1} c_{j+1} \sum_{i=0}^{2j} (-1)^i A_i{}^T Q A_{2j-i}. \tag{15}$$

Note that $A_i = 0$ for $i \geqslant n$, because of the Cayley-Hamilton theorem. From a numerical point of view the formula is not a satisfactory solution.

Jameson [9,20] developed and tested a procedure closely related to these Leverrier computations:

THEOREM 4 [9]. *Let L and G be defined by*

$$G = A^n - a_1 A^{n-1} + \cdots + (-1)^n a_n I, \tag{16}$$

$$L = C_n + a_1 C_{n-1} + \cdots + a_{n-1} C_1, \tag{17}$$

where

$$C_1 = Q,$$

$$C_k = A^T C_{k-1} + QA^{k-1}, \qquad k = 2, \ldots, n. \tag{18}$$

Then

$$S = G^{-1} L^T. \tag{19}$$

Algorithm 5 solves (1) by this theorem. It is sensitive to round-off errors, Jameson used triple precision accumulation in his tests, and the main part of the computation must be redone for new Q:s.

2.2.5. By using the Danilevskii algorithm [7], Molinari [13] reduced some of the difficulties with the companion form transformation.

THEOREM 5 [13]. *Let T transform A to the companion form* $\hat{A}$,

$$TAT^{-1} = \hat{A},$$

and let

$$\hat{Q} = T^{-T} A T^{-1} \qquad \text{and} \qquad \hat{S} = T^{-T} S T^{-1}.$$

Define the vector b by

$$b_i = \begin{cases} \frac{1}{2} \sum_{k=1}^{2i-1} (-1)^{k+1} q_{2i-k,k}, & i = 1, \ldots, \left[\frac{n+1}{2}\right], \\ \frac{1}{2} \sum_{k=1}^{2n+1-2i} (-1)^{k+1} q_{n+1-k,2i-n-1+k}, & i = \left[\frac{n+1}{2}\right] + 1, \ldots, n, \end{cases} \tag{20}$$

and x as the solution of

$$Hx = b.$$

H is the Hurwitz matrix, and $[n]$ *denotes the integer part of n. Then the first column of* $\hat{S}$ *is obtained by*

$$\hat{s}_{i1} = (-1)^{i+1} x_i \tag{21}$$

and the other $n-1$ *columns by the recursion for* $j = 1, \ldots, n-1$:

$$\begin{aligned} \hat{s}_{i,j+1} &= -q_{ij} + p_j s_{i1} + p_i s_{j1} - s_{i+1,j} \qquad i = 1, \ldots, n-1 \\ \hat{s}_{n,j+1} &= -q_{nj} + p_j s_{n1} + p_n s_{j1}. \end{aligned} \tag{22}$$

Algorithm 6 is based on this work, which is an improvement of the foregoing. A general companion form might be used, not only the canonical forms corresponding to the transfer function, and this decreases the computation necessary and increases the pivoting possibilities. New Q-matrices can be solved with reasonable effort and an improvement routine can be applied inside the companion form transformation.

2.2.6. The Schwarz and Routh [2a, 2c, 17, 19] canonical forms have been used in order to formalize the above algebra. The transformations are usually done via the companion form, and show the same difficulties. The solution is simple only for diagonal Q and a few other special cases.

2.3. *Iterative Methods*

2.3.1 The solution of (1) can for stable A be written as [3]:

$$S = \int_0^\infty e^{A^T t} Q e^{At} dt. \tag{23}$$

Davison and Man [6] integrated (23) by the simple Euler approximation and obtained

$$\begin{aligned} S_0 &= 0, \\ S_{k+1} &= \phi^T S_k \phi + \tilde{Q}, \end{aligned} \tag{24}$$

where $\phi = \exp\{A \cdot h\}$ and $\tilde{Q} = h \cdot Q$, or accelerated

$$S_0 = \tilde{Q},$$
$$S_{k+1} = (\phi^T)^{2^k} S_k \phi^{2^k} + S_k, \tag{25}$$

and with $\|S_{k+1} - S_k\| < 10^{-6}$ as a possible stopping condition.

Algorithm 7 is used to test the above formula. The properties specified in [6] can be verified but one important fact is lacking: S_k does *not* converge to the correct solution of (1) if h is not chosen *very* small. The Euler approximation must be valid. On the other hand, the representation of ϕ will then contain less information and large round-off errors will result. The iteration must be redone for new Q:*s*. Molinari [13] refers to 7 as the commonly preferred algorithm.

2.3.2. It is also possible to view (25) as the solution of

$$-dS/dt = A^TS + SA + Q, \tag{26}$$

which can be integrated either by Runge Kutta or other conventional methods or by using the linearity for a fundamental matrix approach:

$$\Sigma(t) = \exp t \cdot \begin{bmatrix} -A & 0 \\ Q & A^T \end{bmatrix} = \begin{bmatrix} (\Sigma_{22}^T)^{-1} & 0 \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}, \tag{27}$$

$$S(t) = [\Sigma_{21}(t) + \Sigma_{22}(t)\, S(0)\, \Sigma_{22}^T(t)]. \tag{28}$$

Define $\phi = \Sigma_{22}^T(h)$ and $\tilde{Q} = \Sigma_{21}(h)\, \Sigma_{22}^T(h)$, then (28) can be rewritten as (24).

Algorithm 8 is based on the accelerated version (25). ϕ is computed by series expansion with 7 terms and automatic scaling. $\tilde{Q}$ is obtained by the iteration:

$$T_1 = Qh, \qquad T_{k+1} = \frac{h}{k+1}\{(T_k A) + (T_k A)^T\}, \tag{29}$$
$$Q_1 = T_1, \qquad Q_{k+1} = Q_k + T_{k+1},$$

which is obtained by series expansion of

$$\tilde{Q} = \int_0^h e^{A^T t}\, Q\, e^{At}\, dt.$$

No scaling is performed and the number of terms T_k is maximized to 35. The stopping condition for (29) is

$$\|T_k\|/\|Q_k\| < 10^{-7}. \tag{30}$$

Both the $\tilde{Q}$ computation and the iteration must be redone for new Q:*s*. Better methods for the $\tilde{Q}$ computation might exist.

2.3.3. The two methods just described can be viewed as transforming equation (1) to equation (4) by

$$A \rightarrow \phi = \exp(Ah). \tag{31}$$

Another possibility for going from (1) to (4) by introducing a one-to-one transformation mapping the left half plane on to the unit circle is the bilinear transform [2b, 19, 21b, 22]:

$$A \rightarrow \phi = -(A + aI)(A - aI)^{-1}. \tag{32}$$

Q is then transformed to

$$\tilde{Q} = (A^T - aI)^{-1} Q(A - aI)^{-1}/2a.$$

Algorithm 9 uses the acceleration formula to solve (4) obtained in this way [21b].

The convergence rate of 8 and 9 depends on the choice of a and h and on the spread of the eigenvalues of the matrix A. The eigenvalues of ϕ, z_i, are, respectively,

$$z_i = e^{h\lambda_i} \qquad \text{and} \qquad z_i = \frac{a + \lambda_i}{a - \lambda_i}.$$

The absolute largest eigenvalue, $\lambda_{\max}$, times h is limiting for the convergence of the $\tilde{Q}$ computation in 8, and the absolute smallest one, $\lambda_{\min}$, times h determines the convergence of (25). This implies that smaller h means better $\tilde{Q}$ but more iterations of (25).

In algorithm 9 the choice

$$a = \sqrt{\lambda_{\min} \cdot \lambda_{\max}}$$

minimizes $\max_i |z_i|$, for real λ_i, thus leading to best convergence of (25).

The operations involved in 9 are simpler than in 8, and rough calculations with only real eigenvalues indicate that 9 manages a far larger spread in the A-eigenvalues, i.e., more ill-conditioned problems. A bad choice of a seems to be less critical than a bad choice of h.

3. THE NUMERICAL TEST

3.1. *Test Examples*

The algorithms are all tested for 17 different A matrices ranging in order from 2 to 10. The sample contains both stable and unstable A, as well as matrices with close eigenvalues and ill-conditioned matrices with a large eigenvalue spread. Some of the matrices were used in [2b, 5, 13, or 20] but

most of them are standard examples for matrix inversion and eigenvalue calculation. A few of the algorithms are also tested for matrices of order 20.

Six different easily generated Q matrices are used for each A matrix. Some additional testing is done with Q matrices designed to give simple integer valued S matrices. The test batch is listed in the appendix (Section 5).

3.2. *Numerical Results*

The results of the test are summarized in Table 1. The accuracy is measured by comparing the solution with the solution of algorithm 1 in double precision.

For $n = 20$ the accuracy was evaluated as the error obtained when the computed S was substituted into (1). The accuracy of algorithm 1 in double precision is also estimated in this way.

When the two ways of measuring were compared for the low order systems, no great difference was found. Call the first method "accuracy in S" and the second "accuracy in Q." For ill-conditioned A matrices some divergence could happen in the test batch mainly so that the "accuracy in S" was one or two digits better than "the accuracy in Q." For the worst A matrices the difference could be even larger. On the other hand, for the specially designed Q matrices giving simple integer valued S matrices the "accuracy in S" was often worse than the "accuracy in Q." When these differences occur they most often do for all algorithms at the same time.

For fixed matrix order the results show considerable variation depending on the actual A and Q. The figures in Table 1 represent an average for the test batch. Matrices giving failure exit are not included in the average.

It is not possible to draw general conclusions about new test examples. The only ninth-order A matrix tested is, for instance, very simple, resulting in better accuracy than for the eighth-order average.

Generally it can be said that equation (1) is difficult for very large systems, especially if A is ill-conditioned, that is, mostly if A has a large spread in the eigenvalues. It should also be stated that different Q matrices can "hide" these difficulties to varying extent.

3.3. *Discussion of the Results*

The methods described are tested as general-purpose algorithms and as such there only remain two, algorithms 1 and 9.

The eigenvalue algorithms are neither accurate nor fast, and often fail if two eigenvalues are close. Algorithm 3 is out of the question although the computing time could be almost halved. Instead of computing all eigenvalues and eigenvectors of the $2n \times 2n$ matrix, it is sufficient to compute only the n eigenvectors corresponding to the stable eigenvalues. Algorithm 2 is probably the best of all methods if the eigenvalues and eigenvectors of A are known or useful in the future analysis.

TABLE 1

RESULT OF THE TEST

Order of the example	Algorithm									
	Direct methods		Eigenvalue methods		Companion form methods			Iterative methods		
	1 double prec.	1	2	3	4	5	6	7	8	9
2×2	2 (17)	1 (7)	13 (7)	20 (7)	2 (7)	2 (8)	1 (6)	7 (2)	10 (6)	3 (7)
3×3	5 (16)	3 (6)	35 (6)	70 (6)	6 (7)	7 (5)	3 (5)	20 (2)	25 (5)	10 (6)
4×4	12 (16)	8 (6)	80 (5)	140 (5)	13 (5)	17 (4)	5 (4)	40 (2)	50 (4)	25 (6)
6×6	62 (16)	42 (6)	200 (4)	400 (3)	50 (1)	75 (2)	14 (1)	150 (2)	170 (4)	70 (6)
8×8	245 (14)	161 (5)	350 (2)	750 (1)	155 (2)	230 (1)	27 (0)	300 (2)	300 (4)	200 (6)
9×9	400 (17)	290 (7)	325 (6)	950 (1)	240 (6)	360 (6)	—	260 (2)	290 (6)	140 (7)
10×10	700 (10)	500 (2)	650 (0)	1400 (0)	340 (1)	530 (1)	50 (–)	500 (2)	500 (3)	210 (4)
20×20	—	—	4500 (4)	—	6000 (0)	8000 (0)	340 (–)	—	5000 (5)	2000 (5)

The first number represents execution time in milliseconds. The numbers in parentheses give the relative accuracy specified in significant digits.

Algorithm 5 has average properties, rather slow for small orders, proportional to n^4 for large orders, memory requirement proportional to n^2, and with bad accuracy for the difficult large-order problems. Algorithm 1 is better for small orders and algorithm 9 for large orders.

Algorithm 4 has properties similar to those of algorithm 5. It has, however, some advantages. If Q is not full rank the calculations are considerably easier, also if $G(s)$ is known or otherwise wanted. Moreover if $V = CSC^T$, and not S, is the quantity desired, no other method should be used, especially if C is just a vector.

Algorithm 4 needs positive semidefinite Q and stable A.

Algorithm 6 needs long code but small internal storage. Although there are pivoting possibilities the accuracy achieved is too bad. Double precision would make it possible to solve equation (1) for higher orders, but even in double precision large systems are impossible to handle. The difficulties arise from the companion form representation. The execution time is by far the shortest, proportional to n^3. Noncyclic matrices, like the ninth-order example, are not possible to transform to companion form, and failure exit of algorithm 6 results.

Algorithm 9 is always better than 7 and 8 both in accuracy and computing time. The figures presented are obtained for good values of the parameters a and h, respectively. It is found that, with a minimum of *a priori* knowledge of the system, both a and h can be estimated sufficiently well to give only a slight increase in execution time and round-off errors. All the iterative methods give error indication for unstable A matrices.

Algorithm 1 is the best and easiest for small systems, and no free parameter a or h has to be chosen. For large systems, however, the execution time is proportional to n^6 and the internal storage proportional to n^4. It was not even possible to test for $n = 20$ on the big Univac 1108 machine with more than $40k$ words available memory.

4. RECOMMENDATIONS

The simplest and best method for small orders is the direct solution 1. For large orders, say more than six or seven, other methods supersede it, for instance, iteration method 9. The fastest algorithm is Molinari's (6), which, however, is too sensitive to round off errors.

ACKNOWLEDGMENTS

The author would like to thank Professor K. J. Åström for his valuable comments and suggestions, and Mrs. G. Christensen, who typed the manuscripts.

5. APPENDIX

The A matrices of the test batch (eigenvalues below) were:

2×2

$$\begin{bmatrix} -3 & 0 \\ 0 & -2 \end{bmatrix}, \quad \begin{bmatrix} -2 & -3 \\ -5 & -10 \end{bmatrix}, \quad \begin{bmatrix} -1 & +2 \\ 0 & -2 \end{bmatrix}.$$

$(-3, -2)$ $(-11.56, -0.44)$ $(-2, -1)$

3×3

$$\begin{bmatrix} -0.1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & -2.26 & -0.2 \end{bmatrix}, \quad \begin{bmatrix} -1 & 0 & -3 \\ -3 & -3 & 4 \\ 0 & 0 & -2 \end{bmatrix}, \quad \begin{bmatrix} -20 & 10 & 10 \\ -18 & 17 & 22 \\ 13 & -13 & -17 \end{bmatrix}.$$

$(-0.1 \pm 1.5i, -0.01)$ $(-3, -2, -1)$ $(-1, -0.5 \pm 0.87i)$

4×4

$$\begin{bmatrix} -10 & -7 & -8 & -7 \\ -7 & -5 & -6 & -5 \\ -8 & -6 & -10 & -9 \\ -7 & -5 & -9 & -10 \end{bmatrix}, \quad \begin{bmatrix} -8 & 1 & 0 & 0 \\ -19 & 0 & 1 & 0 \\ -22 & 0 & 0 & 1 \\ -10 & 0 & 0 & 0 \end{bmatrix}, \quad \begin{bmatrix} -4 & 5 & 0 & -3 \\ 0 & -4 & 3 & 5 \\ -5 & 3 & -4 & 0 \\ -3 & 0 & -5 & -4 \end{bmatrix}.$$

$(-30.3, -3.86, -0.84, -0.010)$ $(-5, -1 \pm i, -1)$ $(-12, -2, -1 \pm 5i)$

6×6

$$\begin{bmatrix} -0.1 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -2 & 10 & 10 & 5 \\ 0 & 3 & 0 & -3 & 1 & 0 \\ 7 & 2 & 0 & 0 & -10 & 0 \\ 32 & 15 & 0 & 0 & 100 & -50 \end{bmatrix}, \quad \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & -1 \\ -1 & 1 & 1 & 0 & 0 & 1 \\ 1 & -1 & 1 & 1 & 0 & -1 \\ -1 & 1 & -1 & 1 & +1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 \end{bmatrix}.$$

$(-50, -10, -3, -2, -1, -0.1)$ $(-3.03, +1.31 \pm 1.20i, +1.47 \pm 0.35i, +1.48)$

8×8

$$\begin{bmatrix} -1 & -5 & 3 & 7 & -9 & -2 & -8 & 0 \\ 20 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -2 & 2 & 1 & -3 & -100 & -0.3 \\ 0 & 0 & 0 & -5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -10 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & -0.1 & 0 & 0 \\ 0 & 0 & 0 & -2 & 0 & 0 & -0.01 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -0.5 \end{bmatrix},$$

$(-10, -5, -2, -1 \pm i, -0.5, -0.1, -0.01)$

$$\begin{bmatrix} -0.021516 & -0.021516 & 0 & 0 & -0.001138 & 0.662 & 0 & 0 \\ 0.132 & -0.1469 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -0.4241 & 0 & 0 & 0 & 0 & 0.5561 \\ 0 & 0 & -0.516 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2.7073 & 0 & -0.4995 & 0 & 0 & 0 \\ 0 & 0 & 0.5166 & 0 & 0 & -1.834 & 0.1207 & 0 \\ 0 & 0 & 0.516 & 0 & 0 & -1.332 & 0 & 0 \\ 0 & 0 & -0.2346 & 0.0909 & 0 & 0 & 0 & -0.4546 \end{bmatrix},$$

$(-1.7, -0.50, -0.39 \pm 0.30i, -0.12, -0.11, -0.09, -0.05)$

$$\begin{bmatrix} -0.15365 & 0.0040173 & 0.17786 & -0.99009 & 0.075158 & 0 & 0 & 0 \\ 1.2482 & -2.8543 & 0 & 1.4324 & 0.72689 & 4.0383 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.56788 & -0.27685 & 0 & -0.28366 & -2.0496 & -0.13886 & 0 & 0 \\ 0 & 0 & 0 & 0 & -10 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -20 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -3 & -2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}.$$

$(-20, -10, -2.79, -2, -1, -0.27 \pm 0.89i, +0.0336)$

9×9

$$\begin{bmatrix} -1.6667 & 0 & 1.3333 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.16667 & 0 & -1.3333 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -4.667 & 0 & 1.3333 \\ 0 & 0 & 0 & 0 & 0 & 0 & -2.5 & -6 & 5 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.1667 & 0 & -4.3333 \end{bmatrix}.$$

$(-6, -5, -4, -3, -3, -3, -2, -2, -1)$

10×10

$$\begin{bmatrix} -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & -2 & -2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & -2 & -3 & -3 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & -2 & -3 & -4 & -4 & 0 & 0 & 0 & 0 & 0 \\ -1 & -2 & -3 & -4 & -5 & -5 & 0 & 0 & 0 & 0 \\ -1 & -2 & -3 & -4 & -5 & -6 & -6 & 0 & 0 & 0 \\ -1 & -2 & -3 & -4 & -5 & -6 & -7 & -7 & 0 & 0 \\ -1 & -2 & -3 & -4 & -5 & -6 & -7 & -8 & -8 & 0 \\ -1 & -2 & -3 & -4 & -5 & -6 & -7 & -8 & -9 & -9 \\ -1 & -2 & -3 & -4 & -5 & -6 & -7 & -8 & -9 & -10 \end{bmatrix},$$

$(-25.58, -14.76, -8.05, -3.89, -1.62, -0.62, -0.26, -0.12, -0.065, -0.042)$

$$\begin{bmatrix}
-1 & -1 & -1 & 2 & -1 & 1 & -2 & 2 & -4 & 3 \\
1 & -2 & -3 & 4 & -2 & 2 & -4 & 4 & -8 & 6 \\
1 & 0 & -5 & 5 & -3 & 3 & -6 & 6 & -12 & 9 \\
1 & 0 & -3 & 4 & -4 & 4 & -8 & 8 & -16 & 12 \\
1 & 0 & -3 & 6 & -5 & 4 & -10 & 10 & -20 & 15 \\
1 & 0 & -3 & 6 & -2 & 2 & -12 & 12 & -24 & 18 \\
1 & 0 & -3 & 6 & -2 & 5 & -15 & 13 & -28 & 21 \\
1 & 0 & -3 & 6 & -2 & 5 & -12 & 11 & -32 & 24 \\
1 & 0 & -3 & 6 & -2 & 5 & -12 & 14 & -37 & 26 \\
1 & 0 & -3 & 6 & -2 & 5 & -12 & 14 & -36 & 25
\end{bmatrix}.$$

$(-3, -3, -3, -3, -2, -2, -2, -2, -2, -1)$

20×20

$$\begin{aligned}
a_{ij} &= -2, && i = j, \\
&= 1, && |i-j| = 1, \\
&= 0, && \text{otherwise.}
\end{aligned}$$

Eigenvalues: $\lambda_i = -2\left(1 - \cos\dfrac{\pi \cdot i}{n+1}\right)$.

20×20

$$\begin{aligned}
a_{ij} &= -1, && j = i+1, \\
&= -1.001, && j = i, \\
&= -(0.001)^{i-j+1}, && j < i, \\
&= 0, && \text{otherwise.}
\end{aligned}$$

Eigenvalues: $\lambda_i = -1, \quad i = 1, \ldots, [n/2],$

$$= -1 - 0.004 \cdot \cos^2 \frac{\pi(i - [n/2])}{n+2}, \quad i = [n/2]+1, \ldots, n.$$

The six Q matrices were generated by:

(1) $$q_{ij} = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases}$$ pos. def.

(2) $$q_{ij} = \begin{cases} 2, & i = j, \\ -1, & |i-j| = 1, \\ 0, & \text{otherwise.} \end{cases}$$ pos. def.

(3) $$q_{ij} = \begin{cases} 1, & i = j, \\ 0.2, & i \neq j. \end{cases}$$ pos. def.

(4) $q_{ij} = 1$, all i, j. pos. semi. def.

(5) $q_{ij} = 0.999$, all i,j. pos. semi. def.

(6) $q_{ij} = 2 \max(i,j) - 1$. indef.

REFERENCES

1 Anderson, B. D. O., Solution of the spectral factorization problem, *IEEE Trans. Automatic Control* **12**, 410–14 (Aug. 1967).

2a Barnett, S., and Storey, C., The Liapunov matrix equation and Schwarz' form, *IEEE Trans. Automatic Control* **12**, 117–18 (Feb. 1967).

2b Barnett, S., and Storey, C., Remarks on numerical solution of the Liapunov matrix equation, *Electronics Letters* **3**, 417–18 (Sept. 1967).

2c Barnett, S., and Storey, C., Some applications of the Liapunov matrix equation, *J. Inst. Math. Appl.* **4**, 33–42 (1968).

2d Barnett, S., and Storey, C., Insensitivity of optimal linear control systems to persistent changes in parameters, *Intern. J. Control* **4**, 179–84 (Aug. 1966).

3 Bellman, R., *Introduction to Matrix Analysis*, McGraw-Hill, New York, 1960.

4 Bingulac, S. P., An alternate approach to expanding $PA + A'P = -Q$, *IEEE Trans. Automatic Control* **15**, 135–7 (Feb. 1970).

5 Chen, C. F., and Shieh, L. S., A note on expanding $PA + A^TP = -Q$, *IEEE Trans. Automatic Control* **13**, 122–3 (Feb. 1968).

6 Davison, E. J., and Man, F. T., The numerical solution of $A'Q + QA = -C$, *IEEE Trans. Automatic Control* **13**, 448–9 (Aug. 1968).

7 Faddeeva, V. N., *Computational Methods of Linear Algebra*, Moscow, 1950; Dover, New York, 1959.

8 Forsythe, G., and Moler, C. B., *Computer Solution of Linear Algebraic Systems*, Prentice-Hall, Englewood Cliffs, N.J., 1967.

9 Jameson, A., Solution of the equation $AX + XB = C$ by inversion of an $M \times M$ or $N \times N$ matrix, *SIAM J. Appl. Math.* **13**, 1020–3 (Sept. 1968).

10 Kleinman, D. L., Solution of linear regulator problem with infinite terminal time, *IEEE Trans. Automatic Control* **13**, 114–15 (Feb. 1968).

11 Luenberger, D. G., Observers for multivariable systems, *IEEE Trans. Automatic Control* **11**, 190–7 (April 1966).

12 Ma, E. C., A finite series solution of the matrix equation, $AX - XB = C$, *SIAM J. Appl. Math.* **14**, 490–5 (May 1966).

13 Molinari, B. P., Algebraic solution of matrix linear equations in control theory, *Proc. IEE* **116**, 1748–54 (Oct. 1969).

14 Müller, P., Die Berechnung von Ljapunov-Funktionen und von quadratischen Regelflächen für lineare, stetige, zeitinvariante Mehrgrössensysteme, *Regelungstechnik* **17**, 341–5 (Aug. 1969).

15 Mårtensson, K., On the Riccati Equation, Thesis for the degree of Teknologie Licentiat, Rept. 7002, Lund Institute of Technology, Division of Automatic Control. *Information Sciences* **3**, 17 (1971).

16 Nekolny, J., and Benes, J., Simultaneous control of stability and quality of adjustment application in statistical dynamics, *Proc. 1st IFAC Congress, Moscow 1960*, Vol. 2, pp. 734–744 (Butterworths, London, 1961).

17 Parks, P. C., A new proof of the Hurwitz stability criterion by the second method of Liapunov with applications to optimum transfer function, *4th JACC, Minneapolis 1963*, pp. 471–8 (AIChE, 1963).

18 Potter, J. E., Matrix quadratic solutions, *SIAM J. Appl. Math.* **14**, 496–551 (May 1966).
19 Power, H. M., Canonical form for the matrices of linear discrete systems, *Proc. IEE* **116**, 1245–52 (July 1969).
20 Rothschild, D., and Jameson, A., Comparison of four numerical algorithms for solving the Liapunov matrix equation, *Intern. J. Control* **11**, 181–98 (Feb. 1970).
21a Smith, R. A., Matrix calculations for Liapunov quadratic forms, *J. Differential Equations* **2**, 208–17 (April 1966).
21b Smith, R. A., Matrix equation $XA + BX = C$, *SIAM J. Appl. Math.* **16**, 198–201 (1968).
22 Taussky, O., Matrices C with $C^n \to 0$, *J. Algebra* **1**, 5–10 (1964).
23 Åström, K. J., *Introduction to Stochastic Control Theory*, Academic Press, New York, 1970.

Received October 1, 1970