SIAM J. APPL. MATH.
Vol. 31, No. 1, July 1976

POLYNOMIAL FACTORIZATION VIA THE RICCATI EQUATION*

DAVID J. CLEMENTS AND BRIAN D. O. ANDERSON†

Abstract. The paper tackles the problem of factoring a real polynomial $f(z)$, nonzero on $|z|=1$, as the product of two polynomials $u(z)$ and $v(z)$ with zeros in $|z|<1$ and $|z|>1$. The individual zeros of $f(z)$ are not found. Riccati difference equations are shown to provide a tool for executing the factorization.

1. Introduction. The following problem arises in a certain design method for telephone channel equalizers [1]. Suppose $f(z)$ is a real polynomial with $f(z)\neq 0$ for any z on the unit circle, $|z|=1$. Find real polynomials $u(z)$ and $v(z)$, $u(z)$ with all zeros inside the unit circle and $v(z)$ with all zeros outside the unit circle, such that $f(z)=u(z)v(z)$. Because of the context in which this problem occurs, where z can be thought of as a transform variable associated with discrete-time signals, we shall call this problem the *discrete-time factorization problem.* Presumably there are other situations in which this problem arises, but we are currently unaware of them.

Via the bilinear transformation $z=s+1/(s-1)$, we can set up a corresponding continuous-time factorization problem (again, the terminology is of an electrical engineering nature). Suppose $f(s)$ is a real polynomial with $f(j\omega)\neq 0$ for any real ω; the problem is to find real polynomials $u(s)$ and $v(s)$, $u(s)$ with all zeros in $\mathrm{Re}\,[s]<0$, $v(s)$ with all zeros in $\mathrm{Re}\,[s]>0$, such that $f(s)=u(s)v(s)$.

In principle, the solution to either factorization problem can be obtained by computing the zeros of $f(\cdot)$ and then taking $u(\cdot)$ to be the polynomial with zeros comprising the particular zeros of $f(\cdot)$ in $|z|<1$ or $\mathrm{Re}\,[s]<0$, as the case may be.

However, in the light of alternative and more efficient procedures being available for a restricted version of the problem, described below, we are led to a search for an extension of these alternative procedures to the unrestricted problem.

The restricted problem is that of *spectral factorization of a polynomial.* The spectral factorization problem in discrete time is a discrete-time factorization problem where $f(z)$ is required to be self-inversive, i.e., if $f(z_0)=0$, then $f(z_0^{-1})=0$; with $f(z)$ of degree $2n$, one then has $f(z)=\pm u(z)z^n u(z^{-1})$, i.e., $v(z)=\pm z^n u(z^{-1})$ in the factorization. In the continuous-time version of the problem, $f(s)$ is required to be even, i.e., $f(s)=f(-s)$; then the factorization is of the form $f(s)=\pm u(s)u(-s)$. (The reason for the nomenclature "spectral factorization" lies in the fact that this sort of factorization arises in the study of rational power spectra, [2], [3].)

* Received by the editors August 9, 1974, and in revised form April 15, 1975.

† Department of Electrical Engineering, University of Newcastle, New South Wales, 2308, Australia. Much of the work reported in this paper was carried out in the Department of Computer and Information Science, University of Massachusetts, Amherst, Massachusetts. This work was supported by the Australian Research Grants Committee, the Australian Radio Research Board and the National Science Foundation under Grant GJ35759.

Recently, it has been shown that a computational procedure for solving the discrete-time spectral factorization problem can be based on use of Riccati difference equations [4], while it has been known for a longer time that the continuous-time problem could be tackled using Riccati differential equations [5], [6]. (The independent variable in both cases is time, so the discrete-time character of difference equations and the continuous-time character of differential equations provides further justification for the terminology discrete-time and continuous-time problem.) The contribution of this paper is to extend the applicability of Riccati equation methods to the unrestricted factorization problems.

Let us note some of the difficulties that arise in attempting the extension. In the spectral factorization problem, the matrices which solve the Riccati equations are square and symmetric, while in the unrestricted problem, in general they are not even square, let alone symmetric. Of itself, this does not contribute a difficulty, but one major difficulty does arise out of this: it is necessary to examine steady state ($t \to \infty$) limits of solutions of the Riccati equations, and the symmetry property (and consequent partial ordering of the matrices) in the spectral factorization problem is apparently crucial to establishing existence of the limits. Second, it also seems crucial in establishing the existence of limits in the spectral factorization problem to have an interpretation of the equation solutions in terms of a variational problem. (The variational problem, by its nature, allows conclusions about the ordering of some quantities to be drawn and this ordering is used in establishing limit existence.) In the general factorization problem, there is no underlying variational problem.

What all this means is that, though the statements of the results for the unrestricted factorization problem look similar to statements of the results for the spectral factorization problem, the proofs tend to be dissimilar.

As further evidence of the dissimilarity, we can observe that matrix versions of the spectral factorization results are known [4], [6], whereas we have to this point been unable to obtain results for a matrix version of the general factorization problem. Moreover, we can establish positively that the methods for the scalar problem cannot be varied in a straightforward manner to tackle the matrix problem, in contrast to the Riccati equation method for the spectral factorization problem.

There appears to be little other work examining the problem of this paper. Amongst such work, we note, however, that of Bauer [7], [8] and Roberts [9]. The paper [7] is concerned with the determination of polynomial zeros. In [8], this problem is first blended in with the problem of this paper by using ideas of [7] (in that an algorithm is given which obtains, simultaneously and iteratively, the k zeros s_i of an nth-degree polynomial with largest modulus, and the kth-degree polynomial of which the s_i are zeros). This algorithm is then varied to consider just the obtaining of the kth-degree polynomial. The entire set of algorithms are outgrowths of the Bernoulli method of determining polynomial zeros [10]. For the algorithm to work, fulfillment of a certain starting condition is required. We are unable to give computational comparisons between the methods of this paper and those of [9], except to a very limited extent. By special choice of certain free parameters, the algorithm of [9] is shown to become one of determining an L-U

factorization [10] of an infinite banded matrix. The same turns out to be true of one of our algorithms.

The work of Roberts [9] is related to, rather than directly concerned with, the problem of this paper, and it would appear to be much less computationally appealing. Roberts' idea is to start with a square matrix A and obtain its "sign", another matrix B with the property that the eigenvalues of AB are those of A with positive real part, or the negatives of those of A with negative real parts. (It is assumed that A has no pure imaginary eigenvalues.) An iterative procedure to determine sign A involves using $Z_{r+1} = \frac{1}{2}(Z_r + Z_r^{-1})$ for $r = 0, 1, 2, \cdots$, with Z_0 of dimension the same as A. Clearly, the computational burden is sizable. The connection with the material of this paper is obtained when one notes that given A and its characteristic polynomial $a(s)$, computation of AB and then its characteristic polynomial effectively solves the continuous-time factorization problem for $a(s)$.

An outline of the contents of the paper is as follows. In § 2, we solve the continuous-time problem with the aid of a Riccati differential equation. (As a computational procedure, use of the differential equation would not be as effective as an initial conversion via bilinear transformation to a discrete-time problem, solution of the discrete-time problem with the aid of a Riccati difference equation, and conversion of the solution back to a continuous-time solution, again via the bilinear transformation. The reason is that the difference equation is generally far more amenable to computer solution than the differential equation.) The continuous-time results are seemingly easier to understand than the discrete-time results, which are covered in § 3. In § 4, we show that a specialized version of the discrete-time results amounts to providing an L-U factorization of an infinite banded matrix. The situation is analogous to that for spectral factorization, where the Riccati difference equation method has been shown [4] to specialize an L-U factorization method of Bauer, [8], [11]. In § 5, initially we discuss methods of determining the zero patterns of polynomials in a finite number of rational operations. Then we indicate why the results in this paper fail to extend directly to the matrix case and present examples of nonsymmetric Riccati equations with finite escape times. Computational aspects of the proposed algorithm are discussed, and the section is rounded off with some miscellaneous minor points.

2. Continuous-time problem.

2.1. The factorization procedure. Let $m(s)$ be a real, monic (unity leading coefficient) polynomial of degree n, with n_+ zeros in $\mathrm{Re}[s] < 0$, n_- zeros in $\mathrm{Re}[s] > 0$ (both n_+ and n_- including multiplicities), and $m(j\omega) \neq 0$ for any real ω. Thus $n = n_+ + n_-$. The determination of n_+ and the checking of the condition $m(j\omega) \neq 0$ for any real ω will be discussed in § 5. Choose arbitrary real monic polynomials $f_+(s)$ and $f_-(s)$ of degrees n_+ and n_-, respectively, such that $f_+(s)$ and $f_-(-s)$ are coprime (have no common factor of positive degree). Further, assume $m(s)$ and $f_+(s)f_-(-s)$ are coprime.[1]

[1] There is no necessity for $f_+(s)$ and $f_-(s)$ to have all zeros in $\mathrm{Re}[s] < 0$, but of course they may have this property.

Find the unique polynomials $h_+(s)$ and $h_-(s)$ of degrees at most n_+-1 and n_--1, respectively, such that

$$\frac{m(s)}{f_+(s)f_-(-s)}=1+\frac{h_+(s)}{f_+(s)}+\frac{h_-(-s)}{f_-(-s)}. \tag{2.1}$$

(Note that this can be done in a finite number of rational operations.)

Recall from linear systems theory [12] that a realization of dimension r of the ratio $p(s)/q(s)$ of two real polynomials $p(s)=p_ms^{m-1}+p_{m-1}s^{m-2}+\cdots+p_1$ and $q(s)=s^m+q_ms^{m-1}+\cdots+q_1$ is a triple $\{A, b, c\}$ with A a real $r\times r$ matrix, and b and c real r-vectors such that $p(s)/q(s)=c'(sI-A)^{-1}b$; the triple is completely controllable if $\operatorname{rank}[b \quad Ab \quad \cdots \quad A^{r-1}b]=r$, or equivalently $\omega'A=\lambda\omega'$ for some (possibly complex) row vector ω' and scalar λ implies $\omega'b\neq 0$; the triple is completely observable if $\{A', c, b\}$ is completely controllable; and that one completely controllable realization of $p(s)/q(s)$ is provided by

$$A=\begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -q_1 & -q_2 & -q_3 & \cdots & -q_m \end{bmatrix}, \quad b=\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}, \quad c=\begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_m \end{bmatrix}, \tag{2.2}$$

with $p(s)$ and $q(s)$ coprime constituting a necessary and sufficient condition for complete observability. One calls a completely controllable and completely observable realization minimal.

In this or some other way, determine minimal realizations $\{A_+, b_+, c_+\}$ and $\{A_-, b_-, c_-\}$ of dimensions n_+ and n_-, respectively, for $h_+(s)/f_+(s)$ and $h_-(s)/f_-(s)$; that is, we have

$$\frac{h_+(s)}{f_+(s)}=c_+'(sI-A_+)^{-1}b_+, \qquad \frac{h_-(s)}{f_-(s)}=c_-'(sI-A_-)^{-1}b_-. \tag{2.3}$$

Now, suppose there exists an $n_-\times n_+$ real matrix P satisfying

$$PA_++A_-'P=(Pb_++c_-)(P'b_-+c_+)'. \tag{2.4}$$

Then, defining polynomials $w_+(s)$ and $w_-(s)$ by

$$\frac{w_+(s)}{f_+(s)}=1+(P'b_-+c_+)'(sI-A_+)^{-1}b_+, \tag{2.5}$$

$$\frac{w_-(s)}{f_-(s)}=1+(Pb_++c_-)'(sI-A_-)^{-1}b_-, \tag{2.6}$$

it is shown below that

$$m(s)=w_-(-s)w_+(s). \tag{2.7}$$

We later show that (2.4) has a number of solutions, each producing a factorization of $m(s)$ into polynomials $w_+(s)$ and $w_-(-s)$ according to (2.5) and (2.6). In

i.e., have all zeros in $\mathrm{Re}[s]<0$, and moreover, this P is shown to be the limiting solution of a Riccati differential equation.

Now let us establish (2.7). Using equations (2.5) and (2.6), we have

$$\begin{aligned}\frac{w_-(-s)w_+(s)}{f_-(-s)f_+(s)} &= [1+b_-'(-sI-A_-')^{-1}(Pb_++c_-)][1+(P'b_-+c_+)'(sI-A_+)^{-1}b_+]\\ &= 1+c_+'(sI-A_+)^{-1}b_++b_-'(-sI-A_-')^{-1}c_-\\ &\quad +b_-'(-sI-A_-')^{-1}Pb_++b_-'P(sI-A_+)^{-1}b_+\\ &\quad +b_-'(-sI-A_-')^{-1}(Pb_++C_-)(P'b_-+c_+)'(sI-A_+)^{-1}b_+\\ &= \frac{m(s)}{f_-(-s)f_+(s)}+b_-'(-sI-A_-')^{-1}[P(sI-A_+)+(-sI-A_-')P\\ &\qquad +(Pb_++c_-)(P'b_-+c_+)'](sI-A_+)^{-1}b_+\\ &= \frac{m(s)}{f_-(-s)f_+(s)},\end{aligned}$$

where the final equality follows by using (2.4).

2.2. Existence of solutions P. Rewrite (2.4) as an algebraic Riccati equation

$$P(A_+-b_+c_+')+(A_-'-c_-b_-')P-Pb_+b_-'P-c_-c_+'=0 \tag{2.8}$$

and associate with (2.8) the $n\times n$ matrix M;

$$M=\begin{bmatrix} A_+-b_+c_+' & -b_+b_-' \\ c_-c_+' & -A_-'+c_-b_-' \end{bmatrix}. \tag{2.9}$$

The relations between the solutions of (2.8) and the eigenvectors of M have been studied in detail for the "symmetric" case of $A_+=A_-$, $b_+=b_-$ and $c_+=c_-$ in [13] and [14]. Some of the results carry over to the nonsymmetric case above, as we now discuss.

Let $z_i=[x_i' \quad y_i']'$ be the n-dimensional eigenvector of M corresponding to the eigenvalue λ_i, with z_i partitioned into an n_+-dimensional vector x_i and an n_--dimensional vector y_i. When λ_i is an eigenvalue appearing in a $k\times k$ block of the Jordan form of M as

$$\begin{bmatrix} \lambda_i & 1 & & & \\ & \lambda_i & 1 & & \\ & & \ddots & \ddots & \\ & & & \ddots & \\ & & & & 1 \\ & & & & \lambda_i \end{bmatrix},$$

we define the generalized eigenvectors $z_i, \cdots, z_{i+k-1}$ as nontrivial solutions of

$$\begin{aligned}(M-\lambda_i I)z_i &= 0,\\ (M-\lambda_i I)z_{i+1} &= z_i,\\ &\vdots\\ (M-\lambda_i I)z_{i+k-1} &= z_{i+k-2},\end{aligned} \tag{2.10}$$

where z_{i+j} is called a generalized eigenvector of rank j corresponding to the eigenvalue λ_i. In case λ_i appears in several Jordan blocks, the modification is straightforward; to avoid confusion, we shall not extend our notation to cover this case. It should be clear that if $z_1, z_2, \cdots, z_n$ is a set of appropriately ordered linearly independent generalized eigenvectors of M corresponding to the eigenvalues $\lambda_1, \cdots, \lambda_n$ of M, and if we define the $n \times n$ matrix $T = [z_1 \cdots z_n]$, then $MT = TJ$, where J is the Jordan normal form of M.

We will always assume that when we consider a set of generalized eigenvectors $z_1, \cdots, z_s$, if z_i is an eigenvector of rank j corresponding to an eigenvalue λ_s, then all associated lower ranking generalized eigenvectors are also in the set $z_1, \cdots, z_s$.

Proposition 1C is a combination of results in [14] and [15], with the proof being identical to that in [14].

PROPOSITION 1C. *Let A_+, A_- be real square matrices of dimension n_+, n_-, let b_+, c_+ be real n_+-vectors, and let b_-, c_- be real n_--vectors. Then each solution P of* (2.8), *real or complex, has the form*

$$P = [y_1 \quad y_2 \quad \cdots \quad y_{n_+}][x_1 \quad x_2 \quad \cdots \quad x_{n_+}]^{-1}, \tag{2.11}$$

where $z_1, \cdots, z_{n_+}$ are generalized eigenvectors corresponding to eigenvalues $\lambda_1, \cdots, \lambda_{n_+}$ of the matrix M of (2.9), *and $z_i' = [x_i' \quad y_i']$. Conversely, given a set of generalized eigenvectors $z_1, \cdots, z_{n_+}$ of M corresponding to eigenvalues $\lambda_1, \cdots, \lambda_{n_+}$, a solution to* (2.8) *of the form* (2.11) *exists provided $[x_1 \quad \cdots \quad x_{n_+}]$ is nonsingular.*

Remark. P is invariant with respect to the order of the eigenvectors $z_1, \cdots, z_{n_+}$ in (2.11).

COROLLARY 1C [14]. *For a choice of $z_1, \cdots, z_{n_+}$ and $\lambda_1, \cdots, \lambda_{n_+}$ above such that $[x_1 \quad \cdots \quad x_{n_+}]$ is nonsingular, the eigenvalues of*

$$K_+ = A_+ - b_+(c_+' + b_-'P) \tag{2.12}$$

are $\lambda_1, \cdots, \lambda_{n_+}$, and $x_1, \cdots, x_{n_+}$ is a set of corresponding generalized eigenvectors.

The next result, also known [14], gives conditions for a solution of (2.8) to be real.

PROPOSITION 2C. *With hypotheses as in Proposition* 1C, *suppose that P is a solution of* (2.8) *determined via* (2.11). *Necessary and sufficient conditions for P to be real are*

(i) *all the eigenvectors $z_1, \cdots, z_{n_+}$ are real, or*

(ii) *if z_i of rank k corresponding to eigenvalue λ_i, $\operatorname{Im}(\lambda_i) \neq 0$ is used, then $\bar{z}_i$ of rank k corresponding to eigenvalue $\bar{\lambda}_i$ must also be included in the solution.*

The next lemma precisely characterizes the eigenvalues of M.

LEMMA 1C. *Let $m(s)$ be a real monic polynomial and suppose that for some real square matrices A_+, A_- of dimension n_+, n_-, and for some real n_+-vectors b_+, c_+ and real n_--vectors b_-, c_-,*

$$\frac{m(s)}{\det(sI-A_+)\det(-sI-A_-)} = 1 + c_+'(sI-A_+)^{-1}b_+ + c_-'(-sI-A_-)^{-1}b_-.$$

Then the eigenvalues of M as defined in (2.8) coincide with the zeros of $m(s)$.

Proof. Let I_n denote the $n \times n$ identity matrix. Then

$$\begin{aligned}
&\det(sI_n - M)\\
&= \det\begin{bmatrix}(sI_{n_+}-A_+)+b_+c_+' & b_+b_-'\\ -c_-c_+' & (sI_{n_-}+A_-')-c_-b_-'\end{bmatrix}\\
&= \det\begin{bmatrix}sI_{n_+}-A_+ & 0\\ 0 & sI_{n_-}+A_-'\end{bmatrix}\det\left[I_n+\begin{bmatrix}(sI_{n_+}-A_+)^{-1}b_+\\ (-sI_{n_-}-A_-')^{-1}c_-\end{bmatrix}\begin{bmatrix}c_+' & b_-'\end{bmatrix}\right]\\
&= \det(sI_{n_+}-A_+)\det(sI_{n_-}+A_-)\\
&\quad\cdot[1+c_+'(sI_{n_+}-A_+)^{-1}b_+ + b_-'(-sI_{n_-}-A_-')^{-1}c_-]\\
&= m(s),
\end{aligned}$$

where we have used the identity $\det(I+AB) = \det(I+BA)$ for any matrices A and B of suitable size.

Note that in Propositions 1C and 2C, no complete controllability or complete observability requirements were imposed. Nor have they been explicitly imposed in Lemma 1C; the reader may however be able to discern that if $\det(sI-A_+)$ and $m(s)$ and $\det(-sI-A_-)$ and $m(s)$ are coprime, then $\{A_+, b_+, c_+\}$ and $\{A_-, b_-, c_-\}$ are forced to be minimal (and conversely).

A second point to note is that propositions 1C and 2C hold if the vectors b_+, etc., are replaced by matrices with the same number of rows as the vectors they replace, and all with the same number of columns. A matrix version of Lemma 1C can also be obtained. However, matrix versions of later results of this section cannot be obtained.

The first main result of the section follows.

THEOREM 1C. *Let $c_+'(sI-A_+)^{-1}b_+$ and $c_-'(sI-A_-)^{-1}b_-$ be two real rational transfer functions*[2] *with $[A_+, b_+]$ and $[A_-, b_-]$ completely controllable. Let $z_1, \cdots, z_{n_+}$ be a set of independent generalized eigenvectors corresponding to the eigenvalues $\lambda_1, \cdots, \lambda_{n_+}$ of the matrix M defined in (2.9). With x_i the vector comprising the first n_+ entries of z_i, the matrix $[x_1 \quad x_2 \quad \cdots \quad x_{n_+}]$ is nonsingular.*

Proof. For convenience, order the $z_1, \cdots, z_{n_+}$ so that

$$M[z_1 \quad \cdots \quad z_{n_+}] = [z_1 \quad \cdots \quad z_{n_+}]J, \tag{2.13}$$

[2] A real rational transfer function (standard electrical engineering term) is simply a ratio of two real polynomials.

with J in Jordan normal form. Write $X=[x_1 \quad \cdots \quad x_{n_+}]$, $Y=[y_1 \quad \cdots \quad y_{n_+}]$, and expand (2.13) to give

$$A_+X-b_+c_+'X-b_+b_-'Y=XJ, \tag{2.14}$$

$$c_-c_+'X-A_-'Y+c_-b_-'Y=XJ. \tag{2.15}$$

Assume, in order to establish a contradiction, that X is singular. Then both the nullspace of X, denoted by $\mathcal{N}(X)$, and also $\mathcal{N}(X')$ have nonzero elements. For each pair of vectors $\alpha \in \mathcal{N}(X)$, $\beta \in \mathcal{N}(X')$, we have, from (2.14),

$$\beta' b_+ b_-' Y\alpha = 0. \tag{2.16}$$

Assume temporarily there exists $\beta \neq 0$, $\beta \in \mathcal{N}(X')$ such that $\beta' b_+ \neq 0$. Then since (2.16) holds for all $\alpha \in \mathcal{N}(X)$ and since $\beta' b_+$ and $b_-' Y\alpha$ are scalars, we have that

$$b_-' Y\alpha = 0 \quad \text{for all } \alpha \in \mathcal{N}(X) \tag{2.17}$$

and hence, by (2.14), that

$$XJ\alpha = 0 \quad \text{for all } \alpha \in \mathcal{N}(X).$$

That is, $\mathcal{N}(X)$ is J-invariant and therefore there exists $\alpha^* \in \mathcal{N}(X)$, $\alpha^* \neq 0$ such that $J\alpha^* = \lambda^*\alpha^*$, where λ^* is some eigenvalue of J. Using (2.15), we then obtain

$$A_-' Y\alpha^* = -\lambda^* Y\alpha^*, \tag{2.18}$$

and (2.17) gives

$$b_-' Y\alpha^* = 0. \tag{2.19}$$

Equations (2.18), (2.19) together with the fact that $Y\alpha^* \neq 0$ (if $Y\alpha^* = 0$, then $[z_1 \quad \cdots \quad z_{n_+}]\alpha^* = 0$—contradicting the linear independence of the z_i) contradict the complete controllability of $[A_-, b_-]$.

Therefore

$$\beta' b_+ = 0 \quad \text{for all } \beta \in \mathcal{N}(X'). \tag{2.20}$$

By (2.14), this implies that

$$\beta' A_+ X = 0 \quad \text{for all } \beta \in \mathcal{N}(X'). \tag{2.21}$$

Consequently, $\mathcal{N}(X')$ is A_+'-invariant and there exists $\beta^* \in \mathcal{N}(X')$, $\beta^* \neq 0$, such that $A_+'\beta^* = \mu^*\beta^*$, where μ^* is an eigenvalue of A_+'. From (2.20), we also have $b_+'\beta^* = 0$ and therefore a contradiction to the complete controllability of $[A_+, b_+]$. Hence $[x_1 \quad \cdots \quad x_{n_+}]$ is nonsingular.

As a consequence of the structure of M and the fact that $[A_+, c_+]$ is completely observable if and only if $[A_+', c_+]$ is completely controllable, we have a dual result.

COROLLARY 2C. *Let $c_+'(sI-A_+)^{-1}b_+$ and $c_-'(sI-A_-)^{-1}b_-$ be two real rational transfer functions with $[A_+, c_+]$ and $[A_-, c_-]$ completely observable. Let $z_1, z_2, \cdots, z_{n_-}$ be a set of independent generalized eigenvectors corresponding to the eigenvalues $\lambda_1, \cdots, \lambda_{n_-}$ of the matrix M defined in* (2.9). *With y_i the vector*

comprising the last n_- entries of z_i, the matrix $[y_1 \quad y_2 \quad \cdots \quad y_{n_-}]$ is nonsingular.

Although Propositions 1C and 2C have extensions to the matrix case, Theorem 1C has no immediate extension. The vector character of b_+ and b_- is crucial when concluding from (2.16) and the condition $\beta' b_+ \neq 0$ that $b'_- Y\alpha = 0$. This is the first piece of evidence that the matrix problem is nontrivially different from the scalar problem.

Now let us suppose that A_+, etc., are obtained via the procedure described in § 2.1. For each choice of eigenvalues $\lambda_1, \cdots, \lambda_{n_+}$ of M and corresponding choice of linearly independent generalized eigenvectors $z_1, \cdots, z_{n_+}$, the above theorem guarantees the existence of $[x_1 \quad \cdots \quad x_{n_+}]^{-1}$, and by Corollary 1C, $K_+ = A_+ - b_+(c'_+ + b'_- P)$ with $P = [y_1 \quad y_2 \quad \cdots \quad y_{n_+}][x_1 \quad x_2 \quad \cdots \quad x_{n_+}]^{-1}$ has $\lambda_1, \cdots, \lambda_{n_+}$ as its eigenvalues and $x_1, \cdots, x_{n_+}$ the corresponding set of generalized eigenvectors. Further, the characteristic polynomial of K_+ is

$$\begin{aligned}
\det(sI_{n_+} - K_+) &= \det[sI_{n_+} - A_+ + b_+(c'_+ + b'_- P)] \\
&= \det(sI_{n_+} - A_+) \det[I_{n_+} + (sI_{n_+} - A_+)^{-1} b_+(c'_+ + b'_- P)] \\
&= f_+(s)\{1 + (P'b_- + c_+)'(sI_{n_+} - A_+)^{-1} b_+\} \\
&= w_+(s)
\end{aligned}$$

(recalling (2.5)), and therefore the zeros of $w_+(s)$ are $\lambda_1, \cdots, \lambda_{n_+}$.

In particular, by choosing the eigenvalues $\lambda_1, \cdots, \lambda_{n_+}$ to lie in the half-plane $\mathrm{Re}[s] < 0$, we have $w_+(s)$, and therefore $w_-(s)$, Hurwitz. Further, as we shall now show, the matrix P associated with this choice of $\lambda_1, \cdots, \lambda_{n_+}$ may be obtained as the limiting solution of a Riccati differential equation.

2.3. A limiting solution. The following specialization of a result well known in the study of certain conjugacy problems [16] connects the solution of the Riccati differential equation

$$\dot{P}(t) = P(t)(A_+ - b_+c'_+) + (A'_- - c_-b'_-)P(t) - P(t)b_+b'_-P(t) - c_-c'_+ \tag{2.22}$$

and the solutions of the linear differential equation

$$\dot{Z}(t) = -MZ(t), \tag{2.23}$$

where $Z(t)$ has dimension $n \times n_+$: partition a solution of (2.23) via $Z(t) = [X'(t) \quad Y'(t)]'$, where $X(t)$ is $n_+ \times n_+$; then if $X(t)$ is nonsingular on some interval $[\alpha, \beta]$, $P(t) = Y(t)X^{-1}(t)$ is a solution of (2.22) on $[\alpha, \beta]$.

In the present context, we shall show that with $Z(0) = [I_{n_+} \quad 0]'$, $X(t)$ is nonsingular on $[t_1, \infty)$ for some suitably large t_1. Then $P(t) = Y(t)X^{-1}(t)$ will exist as a solution of (2.22) on $[t_1, \infty)$, and we can show that $\lim_{t\to\infty} P(t) = P$ exists and satisfies (2.8), the steady state version of (2.22). Finally, we shall show that this particular solution of (2.8) ensures that $w_+(s)$ defined by (2.5) is Hurwitz.

To effect this program, we begin by writing the solution of (2.23) explicitly in terms of the eigenvalues and generalized eigenvectors of M. Choose a nonsingular T such that

$$T^{-1}MT = \begin{bmatrix} \Lambda_+ & 0 \\ 0 & \Lambda_- \end{bmatrix},$$

where Λ_+ and Λ_- are the Jordan normal forms corresponding to the eigenvalues of M lying in the half-planes $\mathrm{Re}\,[s]<0$ and $\mathrm{Re}\,[s]>0$, respectively. Partition T as

$$T=\begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix}.$$

Then it is straightforward to show that

$$Z(t)=\begin{bmatrix} X(t) \\ Y(t) \end{bmatrix}=\begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix}\begin{bmatrix} e^{-\Lambda_+ t} & 0 \\ 0 & e^{-\Lambda_- t} \end{bmatrix}\begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix}^{-1}\begin{bmatrix} I \\ 0 \end{bmatrix}.$$

Since T is nonsingular, the columns of $\begin{bmatrix} T_{11} \\ T_{21} \end{bmatrix}$, which are generalized eigenvectors of M, determine a nonsingular T_{11} by Theorem 1C, and Corollary 2C shows that T_{22} is nonsingular. Then, because $\det T=\det (T_{11}-T_{12}T_{22}^{-1}T_{21})\det T_{22}$ by a standard formula, it is apparent that $T_{11}-T_{12}T_{22}^{-1}T_{21}$ is nonsingular. One can check then that

$$\begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix}^{-1}\begin{bmatrix} I \\ 0 \end{bmatrix}=\begin{bmatrix} (T_{11}-T_{12}T_{22}^{-1}T_{21})^{-1} \\ -T_{22}^{-1}T_{21}(T_{11}-T_{12}T_{22}^{-1}T_{21})^{-1} \end{bmatrix},$$

and so

$$\begin{aligned} X(t) &= T_{11}\,e^{-\Lambda_+ t}(T_{11}-T_{12}T_{22}^{-1}T_{21})^{-1}-T_{12}\,e^{-\Lambda_- t}T_{22}^{-1}T_{21}(T_{11}-T_{12}T_{22}^{-1}T_{21})^{-1} \\ &= [T_{11}-T_{12}\,e^{-\Lambda_- t}T_{22}^{-1}T_{21}\,e^{\Lambda_+ t}]\,e^{-\Lambda_+ t}(T_{11}-T_{12}T_{22}^{-1}T_{21})^{-1} \end{aligned} \tag{2.24}$$

and

$$Y(t)=[T_{21}-T_{22}\,e^{-\Lambda_- t}T_{22}^{-1}T_{21}\,e^{\Lambda_+ t}]\,e^{-\Lambda_+ t}(T_{11}-T_{12}T_{22}^{-1}T_{21})^{-1}. \tag{2.25}$$

As $t\to+\infty$, $e^{-\Lambda_- t}\to 0$ and $e^{\Lambda_+ t}\to 0$ in view of the restrictions on the real parts of the entries of Λ_- and Λ_+. Then $T_{11}-T_{12}\,e^{-\Lambda_- t}T_{22}^{-1}T_{21}\,e^{\Lambda_+ t}$ approaches a nonsingular matrix and is therefore nonsingular for suitably large t (in view of the continuity of the determinant as a function of its elements).

Next, as we know, $P(t)=Y(t)X^{-1}(t)$ satisfies (2.22), and so

$$P(t)=[T_{21}-T_{22}\,e^{-\Lambda_- t}T_{22}^{-1}T_{21}\,e^{\Lambda_+ t}][T_{11}-T_{12}\,e^{-\Lambda_- t}T_{22}^{-1}T_{21}\,e^{\Lambda_+ t}]^{-1}$$

on $[t_1,\infty)$ for some suitably large t_1 satisfies (2.22). Letting $t\to\infty$, we see that

$$\lim_{t\to\infty} P(t)=T_{21}T_{11}^{-1} \tag{2.26}$$

exists. By Proposition 1C, this limiting matrix satisfies the limiting or steady state Riccati equation (2.8) and is associated with precisely these eigenvalues of M with negative real parts. In summary, we have proved the following.

THEOREM 2C. *Under the same hypotheses as Theorem* 1C *and the assumption that* $[A_+, c_+]$ *and* $[A_-, c_-]$ *are completely observable, let P be that solution of* (2.8) *(whose existence is guaranteed by Proposition* 1C *and Theorem* 2C*) derived using those generalized eigenvectors of the matrix M of* (2.9) *with eigenvalues possessing negative real parts. Then* $P=\lim_{t\to\infty}P(t)$, *where* $P(t)$ *is the solution of* (2.22) *computed with initial condition* $P(t_1)=Y(t_1)X^{-1}(t_1)$ *and defined on* $[t_1,\infty)$, *and*

where $X(t)$ and $Y(t)$ are defined in the remarks following (2.23), *which is solved with initial condition* $Z'(0) = [I \quad 0]$.

In actual computation, $P(t)$ is a better quantity to work with than $X(t)$ and $Y(t)$ separately, roughly because the entries of $X(t)$ and $Y(t)$ consist of sums of exponential terms with possibly widely differing exponents, but in which approximation by the term with greatest exponent may cause substantial inaccuracy in $Y(t)X^{-1}(t)$; see [17] for a discussion.

For this reason, computation of $P = lim_{t\to\infty} P(t)$ could proceed along the following lines:

1. Solve (2.22) forward in time from initial condition $P(0) = 0$ until $\|P(t)\|$ reaches some bound K set in advance. Suppose this occurs at time t_α.
2. Set $Z(t_\alpha) = [I \quad P'(t_\alpha)]'$ and solve (2.23) forward in time through any possible singularity of $X(t)$. Determine a time $t_\beta > t_\alpha$ for which $\|Y(t_\beta)X^{-1}(t_\beta)\| < K$.
3. Solve (2.22) forward in time from t_β from initial condition $P(t_\beta) = Y(t_\beta)X^{-1}(t_\beta)$.
4. *Repeat the above process until* $lim_{t\to\infty} P(t) = P$ is reached.

The general idea is to use the Riccati equation except when an escape time occurs, and to use the linear equation to solve through the escape time.

Notice that $Z(t_\alpha) = [I \quad P'(t_\alpha)]'$ will not be the same as $\exp(-Mt_\alpha)[I \quad 0]'$, viz., the $Z(t_\alpha)$ which would result from $Z(0) = [I \quad 0]'$. The question then arises as to whether the $P(t)$ for $t \geqq t_1$ defined in Theorem 2C agrees with that defined above. Call $\bar{Z}(t)$ the solution of (2.23) with $\bar{Z}(0) = [I \quad 0]'$. Then $Z(t_\alpha) = \bar{Z}(t_\alpha)X^{-1}(t_\alpha)$; so $X(t_\beta) = \bar{X}(t_\beta)X^{-1}(t_\alpha)$ and $Y(t_\beta) = \bar{Y}(t_\beta)X^{-1}(t_\alpha)$ and $Y(t_\beta)X^{-1}(t_\beta) = \bar{Y}(t_\beta)\bar{X}^{-1}(t_\beta)$. This shows that no difficulty arises through working with different trajectories of (2.23) (provided they are determined as described above) in the sense that $P(t)$ is uniquely defined for all t as $Y(t)X^{-1}(t)$, except where $X(t)$ is singular, despite the nonuniqueness of $X(t)$ and $Y(t)$.

Let us say that (2.22) is *solved in the generalized sense* on $[0, \infty)$ if we avoid solving through escape times by use of (2.23). Theorem 2C can then be rephrased as follows.

THEOREM 2'C. *Under the same hypotheses as Theorem* 2C, *let P be that solution of* (2.8) *derived using those generalized eigenvectors of the matrix M of* (2.9) *with eigenvalues possessing negative real parts. Then* $P = \lim_{t\to\infty} P(t)$, *where* $P(t)$ *is the solution in the generalized sense of* (2.22) *with initial condition* $P(0) = 0$.

In the spectral factorization problem, it is known that the Riccati equation has no escape times in $[0, \infty)$, and so the use of the linear equation is avoided [6]. This is not, in general, true for the unrestricted problems, as an example in § 5 shows. In § 5, we also explore other points of computational interest, including numerical stability, the effect of change of initial conditions, and rate of convergence.

3. Discrete-time factorization. The results in many ways parallel the continuous time results, and we shall follow the layout of § 2 in this section in order to make the parallel as clear as possible.

3.1. The factorization procedure. Let $m(z)$ be a real polynomial of degree n with n_- zeros outside the unit circle $|z| = 1$, n_+ inside, and no zeros on the unit

circle. (Determination of n_+ and n_- and checking the no-zero-on-the-unit-circle condition are discussed in § 5.) Recalling that the goal is to factor $m(z)$ as $u(z)v(z)$, where $u(\cdot)$ has all zeros inside the unit circle and $v(\cdot)$ all zeros outside the unit circle, it is evident that without loss of generality, we may assume $m(0) \neq 0$.

Now let $f_+(z)$ and $f_-(z)$ be real, monic polynomials of degrees n_+ and n_-, respectively such that $f_+(z)$ and $z^{n_-}f_-(z^{-1})$ are coprime. Further, assume that $m(z)$ and $f_+(z)z^{n_-}f_-(z^{-1})$ are also coprime. Then determine the unique polynomials $h_+(z)$ and $h_-(z)$ of degrees at most $n_+ - 1$ and $n_- - 1$, respectively, such that

$$\frac{m(z)}{f_+(z)z^{n_-}f_-(z^{-1})} = r + \frac{h_+(z)}{f_+(z)} + \frac{h_-(z^{-1})}{f_-(z^{-1})}. \tag{3.1}$$

In contrast to the continuous-time problem, r may equal zero. For the moment, assume $r \neq 0$ and normalize (3.1) to

$$\frac{m(z)}{f_+(z)z^{n_-}f_-(z^{-1})} = 1 + \frac{h_+(z)}{f_+(z)} + \frac{h_-(z^{-1})}{f_-(z^{-1})}. \tag{3.2}$$

Then, as for the continuous-time case, choose minimal realizations $\{A_+, b_+, c_+\}$ and $\{A_-, b_-, c_-\}$ satisfying

$$\frac{h_+(z)}{f_+(z)} = c_+'(zI - A_+)^{-1}b_+, \qquad \frac{h_-(z)}{f_-(z)} = c_-'(zI - A_-)^{-1}b_-. \tag{3.3}$$

Suppose there exists an $n_- \times n_+$ matrix P satisfying

$$-P + A_-'PA_+ = (A_-'Pb_+ + c_-)(1 + b_-'Pb_+)^{-1}(b_-'PA_+ + c_+'). \tag{3.4}$$

Then we may define $w_+(z)$ and $w_-(z)$ by

$$\frac{w_+(z)}{f_+(z)} = 1 + \alpha^{-1}(b_-'PA_+ + c_+')(zI - A_+)^{-1}b_+, \tag{3.5}$$

$$\frac{w_-(z)}{f_-(z)} = \alpha + (A_-'PB_+ + c_-)'(zI - A_-)^{-1}b, \tag{3.6}$$

where $\alpha = 1 + b_-'Pb_+$. Noting the identity

$$-P + A_-'PA_+ = (z^{-1}I - A_-')P(zI - A_+) - (z^{-1}I - A_-')Pz - z^{-1}P(zI - A_+), \tag{3.7}$$

we can show that

$$m(z) = z^{n_-}w_-(z^{-1})w_+(z). \tag{3.8}$$

Remarks similar to those made following (2.7) again hold.

3.2. Existence of solutions *P*. With $\bar{A}_+ = A_+ - b_+c_+'$ and $\bar{A}_- = A_- - b_-c_-'$, it can be shown that (3.4) is equivalent to

$$0 = P - \bar{A}_-'P\bar{A}_+ + \bar{A}_-'Pb_+\alpha^{-1}b_-'P\bar{A}_+ + c_-c_+'. \tag{3.9}$$

As in [18], we can associate an $n \times n$ matrix M with (3.9). Define

$$M = \begin{bmatrix} \bar{A}_+ - b_+b_-'\bar{A}_-'^{-1}c_-c_+' & -b_+b_-'\bar{A}_-'^{-1} \\ \bar{A}_-'^{-1}c_-c_+' & \bar{A}_-'^{-1} \end{bmatrix}, \tag{3.10}$$

where we assume, for the moment, that $\bar{A}_-'^{-1}$ exists.

We now state without proof (see [18]) results analogous to Propositions 1C and 2C, and Corollary 1C.

PROPOSITION 1D. *Let A_+, A_- be real square matrices of dimension n_+, n_-, let b_+, c_+ be real n_+-vectors, and let b_-, c_- be real n_--vectors. Assume that $A_- - b_-c_-'$ is nonsingular. Then each solution P of* (3.9), *real or complex, has the form*

$$P=[y_1 \quad y_2 \quad \cdots \quad y_{n_+}][x_1 \quad x_2 \quad \cdots \quad x_{n_+}]^{-1}, \tag{3.11}$$

where $z_1, \cdots, z_{n_+}$ are generalized eigenvectors[3] *corresponding to eigenvalues $\lambda_1, \cdots, \lambda_{n_+}$ of the matrix M of* (3.10), *and $z_i=[x_i' \quad y_i']'$. Conversely, given a set of generalized eigenvectors $z_1, \cdots, z_{n_+}$ of M corresponding to eigenvalues $\lambda_1, \cdots, \lambda_{n_+}$, a solution to* (3.9) *of the form* (3.11) *exists provided $[x_1 \quad \cdots \quad x_{n_+}]$ is nonsingular.*

COROLLARY 1D. *For a choice of $z_1, \cdots, z_{n_+}$ and $\lambda_1, \cdots, \lambda_{n_+}$ above, such that $[x_1 \quad \cdots \quad x_{n_+}]$ is nonsingular, the eigenvalues of*

$$K_+ = \bar{A}_+ - b_+b_-'\bar{A}_-'^{-1}(P+c_-c_+') \tag{3.12}$$

are $\lambda_1, \cdots, \lambda_{n_+}$, and $x_1, \cdots, x_{n_+}$ is a set of corresponding generalized eigenvectors.

PROPOSITION 2D. *With hypotheses as in Proposition* 1D, *suppose that P is a solution of* (3.9) *determined by* (3.11). *Necessary and sufficient conditions for P to be real are:*

(i) *all the eigenvectors $z_1, \cdots, z_{n_+}$ are real, or*

(ii) *if z_i of rank k corresponding to eigenvalue λ_i, $\operatorname{Im}(\lambda_i) \neq 0$, is used, then $\bar{z}_i$ of rank k corresponding to $\bar{\lambda}_i$ must also be included in the solution.*

As a comparison with the continuous-time results will show, the continuous-time definition of P and M are replaced by discrete-time definitions, an extra hypothesis is required, and the matrix K_+ is defined differently.

Lemma 1D which follows also requires an additional hypothesis over Lemma 1C, and of course different quantities arise.

LEMMA 1D. *Let $m(z)$ be a real polynomial such that for some real square matrices A_+, A_- of dimension n_+, n_- and for some real n_+-vectors b_+, c_+ and real n_--vectors b_-, c_-, the following equality holds:*

$$\frac{m(z)}{(z^{n_-})\det(zI-A_+)\det(z^{-1}I-A_-)} = 1+c_+'(zI-A_+)^{-1}b_+ + c_-'(z^{-1}I-A_-)^{-1}b_-.$$

Suppose also that $A_- - b_-c_-'$ is nonsingular. Then the eigenvalues of M as defined in (3.10) *coincide with the zeros of $m(z)$.*

Proof. By direct calculation, one has

$$zI_n - M = \begin{bmatrix} I_{n_+} & -b_+b_-' \\ 0 & I_{n_-} \end{bmatrix} \begin{bmatrix} I_{n_+} & 0 \\ 0 & -z\bar{A}_-'^{-1} \end{bmatrix} \begin{bmatrix} zI_{n_+}-A_+ & 0 \\ 0 & z^{-1}I_{n_-}-A_-' \end{bmatrix}$$
$$\cdot \left(I_n + \begin{bmatrix} (zI_{n_+}-A_+)^{-1}b_+z \\ (z^{-1}I_{n_-}-A_-')^{-1}c_- \end{bmatrix} [z^{-1}c_+' \quad b_-'] \right).$$

[3] The same conventions governing selection of generalized eigenvector sets apply here as in § 2.

Taking determinants yields

$$\det(zI_n - M) = z^{n-}\det[-\bar{A}'_-]^{-1}\det[zI-A_+]\det[z^{-1}I-A_-] \cdot (1+c'_+(zI-A_+)^{-1}b_+ + c'_-(z^{-1}I-A_-)^{-1}b_-).$$

The result is immediate.

Finally, in analogy to Theorem 1C, we have the following.

THEOREM 1D. *Let $c'_+(zI-A_+)^{-1}b_+$ and $c'_-(zI-A_-)^{-1}b_-$ be two real rational transfer functions with $[A_+, b_+]$ and $[A_-, b_-]$ completely controllable and with $A_- - b_-c'_-$ nonsingular. Let $z_1, \cdots, z_{n_+}$ be a set of independent generalized eigenvectors corresponding to the eigenvalues $\lambda_1, \cdots, \lambda_{n_+}$ of the matrix M defined in* (3.10). *With x_i the vector comprising the first n_+ entries of z_i, the matrix $[x_1 \quad x_2 \quad \cdots \quad x_{n_+}]$ is nonsingular.*

A proof can be constructed easily using that of Theorem 1C as a guide.

Likewise, there is an obvious discrete-time equivalent of Corollary 2C.

Now let us suppose that A_+, etc., are obtained via the procedure described in § 3.1. Let us also note that K_+ as defined in Corollary 1D can be written as $K_+ = A_+ - b_+\alpha^{-1}(b'_-PA_+ + c'_+)$ with the aid of (3.4). Then it is easy to show that the eigenvalues of K_+ are the zeros of $w_+(z)$ in (3.5). Therefore, for each choice of eigenvalues $\lambda_1, \cdots, \lambda_{n_+}$ of M and suitable eigenvectors $z_1, \cdots, z_{n_+}$, a solution, P of (3.4) exists and the factor $w_+(z)$ of $m(z)$ has $\lambda_1, \cdots, \lambda_{n_+}$ as its zeros.

Moreover, as we now show, the matrix P corresponding to the n_+ eigenvalues of M lying inside the unit circle can be obtained as the limiting solution of a Riccati difference equation.

3.3. A limiting solution. By Lemma 1D and the assumption $m(0) \neq 0$, M^{-1} exists. This yields the following lemma.

LEMMA 2D. *With quantities as defined previously, consider the linear difference equation*

$$Z_{k+1} = M^{-1}Z_k, \qquad k = 0, 1, \cdots, \tag{3.13}$$

where Z_k has dimension $n \times n_+$ and is partitioned via $Z_k = [X'_k \quad Y'_k]'$, where X_k is $n_+ \times n_+$. Consider also the nonlinear difference equation

$$P_{k+1} = \bar{A}'_-P_k\bar{A}_+ - \bar{A}'_-P_kb_+\alpha_k^{-1}b'_-P_k\bar{A}_+ - c_-c'_+, \tag{3.14}$$

where $\alpha_k = 1 + b'_-P_kb_+$. Then if X_k is nonsingular on some interval $[k_0, k_1]$, $P_k = Y_kX_k^{-1}$ satisfies (3.14) *on $[k_0, k_1]$.*

Proof. Note first that $\bar{A}_+$ is nonsingular, for using the form of M in (3.10), we have

$$M = \begin{bmatrix} I_{n_+} & -b_+b'_- \\ 0 & I_{n_-} \end{bmatrix}\begin{bmatrix} \bar{A}_+ & 0 \\ \bar{A}'^{-1}_- c_-c'_+ & \bar{A}'^{-1}_- \end{bmatrix}$$

and $\det M = \det \bar{A}_+(\det \bar{A}'^{-1}_-)$; M is nonsingular since $m(0) \neq 0$. Now a straightforward calculation verifies that

$$M^{-1} = \begin{bmatrix} \bar{A}_+^{-1} & \bar{A}_+^{-1}b_+b'_- \\ -c_-c'_+\bar{A}_+^{-1} & \bar{A}'_- - c_-c'_+\bar{A}_+^{-1}b_+b'_- \end{bmatrix},$$

and so

$$(3.15)\quad X_{k+1}=\bar{A}_+^{-1}(X_k+b_+b_-'Y_k),\qquad Y_{k+1}=\bar{A}_-'Y_k-c_-c_+'\bar{A}_+^{-1}(X_k+b_+b_-'Y_k).$$

Then

$$\begin{aligned}P_{k+1}=Y_{k+1}X_{k+1}^{-1}&=\bar{A}_-'Y_k(X_k+b_+b_-'Y_k)^{-1}\bar{A}_+-c_-c_+'\\&=\bar{A}_-'Y_kX_k^{-1}(1+b_+b_-'Y_kX_k^{-1})^{-1}\bar{A}_+-c_-c_+',\end{aligned}$$

and the result follows since, as may be checked,

$$P_k-P_kb_+(1+b_-'P_kb_+)^{-1}b_-'P_k=P_k(1+b_+b_-'P_k)^{-1}.$$

Notice in the above proof that nonsingularity of X_k and X_{k+1} forces (see (3.15)) nonsingularity of $I+b_+b_-'P_k$. Hence $0\neq\det[I+b_+b_-'P_k]=1+b_-'P_kb_+$, and α_k is guaranteed to be invertible. Conversely, if X_k and α_k are invertible, so is X_{k+1}.

Following the pattern of § 2, we can write down the solution of (3.13) with initial condition $Z_0=[I_{n_+}\quad 0]'$ explicitly in terms of the eigenvalues and generalized eigenvectors of M. Suppose that

$$T^{-1}MT=\begin{bmatrix}\Lambda_+ & 0\\ 0 & \Lambda_-\end{bmatrix},\qquad T=\begin{bmatrix}T_{11} & T_{12}\\ T_{21} & T_{22}\end{bmatrix},$$

where Λ_+ and Λ_- are Jordan forms with eigenvalues inside and outside $|z|=1$, respectively. Then we have

$$\begin{aligned}X_k&=[T_{11}-T_{12}\Lambda_-^{-k}T_{22}^{-1}T_{21}\Lambda_+^k]\Lambda_+^{-k}(T_{11}-T_{12}T_{22}^{-1}T_{21})^{-1},\\Y_k&=[T_{21}-T_{22}\Lambda_-^{-k}T_{22}^{-1}T_{21}\Lambda_+^k]\Lambda_+^{-k}(T_{11}-T_{12}T_{22}^{-1}T_{21})^{-1},\end{aligned}$$

where all inverses are well-defined. Evidently, for sufficiently large K, X_k is nonsingular and $\lim_{k\to\infty}Y_kX_k^{-1}$ exists and is equal to that particular solution of (3.9) associated with the n_+ eigenvalues of M lying in $|z|<1$.

The algorithm for finding this matrix using a combination of (3.13) and (3.14) is as follows:

1. Solve (3.14) forward in time from initial condition $P_0=0$ (note that $\alpha_0=1$). Stop if and when $\alpha_k=0$, say at $k=k_\alpha$.
2. Set $Z_{k_\alpha}=[I\quad P_{k_\alpha}']'$ and use (3.13) to generate $z_{k_\alpha+1}, z_{k_\alpha+2},\cdots$ at each stage checking existence of X_k^{-1}.
3. If for $k=k_\beta>k_\alpha$, X_k^{-1} exists, set $P_{k_\beta}=Y_{k_\beta}X_{k_\beta}^{-1}$ and use (3.14) again for $k\geqq k_\beta$.
4. Repeat the process until $\lim_{k\to\infty}P_k=P$ is reached.

The theory guarantees that for some suitably large k_1, $\alpha_k\neq 0$ for all $k\geqq k_1$, and (3.14) may be used on $[k_1,\infty)$.

The parallel of Theorem 2'C is as follows:

THEOREM 2D. *Under the same hypotheses as Theorem* 1D *and the assumption that* $[A_+, c_+]$ *and* $[A_-, c_-]$ *are completely observable, let* P *be that solution of* (3.9) *derived using those generalized eigenvectors of* M *of* (3.10) *with eigenvalues in* $|z|<1$. *Then* $P=\lim_{k\to\infty}P_k$, *where* P_k *is the solution in the generalized sense of* (3.14).

Notice that (3.9), satisfied by P, is a steady state version of (3.14), satisfied by P_k, as one expects.

3.4. Discussion of assumptions. In §§ 3.1, 3.2 and 3.3, we have assumed that $\bar{A}_- = A_- - b_- c_-'$ is nonsingular and that the constant r in (3.1) is nonzero. We shall show that these assumptions are in no way limiting. Toward this end, we prove the following lemma.

LEMMA 3D. *With notation as defined earlier,*

(i) $\bar{A}_+$ *and* A_+ *have no common eigenvalues,*

(ii) $\bar{A}_-$ *and* A_- *have no common eigenvalues.*

Proof. Clearly,

$$\det[zI_{n_+} - \bar{A}_+] = \det[zI_{n_+} - A_+]\left\{1 + \frac{h_+(z)}{f_+(z)}\right\} = f_+(z) + h_+(z).$$

By the minimality of $\{A_+, b_+, c_+\}$, $h_+(z)$ and $f_+(z)$ are coprime and therefore $\det[zI_{n_+} - \bar{A}_+]$ and $f_+(z) = \det[zI_{n_+} - A_+]$ are coprime.

Similarly, $\det[zI_{n_-} - \bar{A}_-]$ and $f_-(z) = \det[zI_{n_-} - A_-]$ are coprime.

By Lemma 3D, $\bar{A}_+$ and $\bar{A}_-$ are guaranteed nonsingular by choosing A_+ and A_- singular; or equivalently, choosing $f_+(z)$ and $f_-(z)$ such that $f_+(0) = 0$, $f_-(0) = 0$.

Second, as we now show, the assumption $r \neq 0$ is no restriction, as we can always choose $f_+(z)$ and $f_-(z)$ such that $r \neq 0$.

Suppose that

$$m(z) = z^n + m_{n-1}z^{n-1} + \cdots + m_{n_+ + 1}z^{n_+ + 1} + m_{n_+}z^{n_+} + \cdots + m_1 z + m_0,$$

and set $f_+(z) = z^{n_+}$ and $f_-(z) = z^{n_- - 1}(z + a)$. Then it follows that $r = 0$ if and only if

$$\frac{m(z)}{z^{n_+}(1 + za)} = \frac{a(z)}{z^{n_+}} + \frac{zb(z)}{1 + za}, \tag{3.16}$$

where $a(z)$ is a polynomial of degree $\leqq n_+ - 1$ and $b(z)$ is a polynomial of degree $\leqq n_- - 1$.

Thus from (3.16), $r = 0$ iff

$$z^n + m_{n-1}z^{n-1} + \cdots + m_{n_+ + 1}z^{n_+ + 1} = z^{n_+ + 1}b(z)$$

and

$$m_{n_+}z^{n_+} + \cdots + m_1 z + m_0 = a(z)(1 + za).$$

It is therefore clear that choosing any a^{-1} not a zero of $m_{n_+}z^{n_+} + \cdots + m_1 z + m_0$ *will guarantee that* $r \neq 0$.

In point of fact, random selection of all the coefficients of $f_+(z)$ and $f_-(z)$ will almost always produce an r which is nonzero.

Examples and other remarks will be presented in § 5.

4. Nonsymmetric *L-U* factorization. In this section we relate the factorization problem of the previous section to the problem of finding an L-U factorization of a nonsymmetric banded matrix. This was also done by Bauer for a quite different algorithm, [8]. For a discussion of the corresponding symmetric problem see [4] and [11]. The significance of the material of this section is that it outlines what is possibly the best computational approach of all to solving the discrete-time problem, with, however, one slight catch discussed further below.

Suppose that by normalization we can write

$$m(z)=c_1^- z^n+c_2^- z^{n-1}+\cdots+c_{n_-}^- z^{n+1-n_-}+z^{n-n_-}+c_{n_+}^+ z^{n-1-n_-}+\cdots+c_2^+ z+c_1^+. \tag{4.1}$$

We set $f_+(z)=z^{n_+}$ and $f_-(z)=z^{n_-}$. Then

$$\frac{h_+(z)}{f_+(z)}=c_{n_+}^+ z^{-1}+\cdots+c_1^+ z^{-n_+}=c_+'(zI-A_+)^{-1}b_+,$$

where

$$A_+=\begin{bmatrix}0&1&0&\cdots&0\\0&0&1&&0\\\vdots&&&&\vdots\\0&0&0&\cdots&1\\0&0&0&\cdots&0\end{bmatrix},\quad b_+=\begin{bmatrix}0\\0\\\vdots\\0\\1\end{bmatrix},\quad c_+=\begin{bmatrix}c_1^+\\c_2^+\\\vdots\\c_{n_+-1}^+\\c_{n_+}^+\end{bmatrix},$$

and similarly,

$$\frac{h_-(z)}{f_-(z)}=c_{n_-}^- z^{-1}+\cdots+c_1^- z^{-n_-}=c_-'(zI-A_-)^{-1}b_-,$$

with obvious definitions of A_-, b_-, c_-. It follows from (3.5) that if

$$\frac{w_+(z)}{f_+(z)}=\alpha^{-1}(\alpha+w_{n_+}^+ z^{-1}+\cdots+w_1^+ z^{-n_+}), \tag{4.2}$$

then

$$[w_1^+\ \cdots\ w_{n_+}^+]=\alpha^{-1}[c_1^+\ \ c_2^+ P_{n_-,1}\ \cdots\ c_+^+ P_{n_-,n_+-1}], \tag{4.3}$$

where $\alpha=1+P_{n_-,n_+}$.

Now consider the matrix Riccati equation, equivalent to (3.14),

$$P_{k+1}=A_-' P_k A_+-(A_-' P_k b_+ + c_-)\alpha_k^{-1}(b_-' P_k A_+ + c_+'). \tag{4.4}$$

Write the time index on P as an argument in parentheses, and use subscripts to denote particular entries; guided by (4.2) and (4.3), define

$$[w_1^+(k)\ \cdots\ w_{n_+}^+(k)]=\alpha(k)^{-1}[c_1^+\ \ c_2^+ P_{n_-,1}(k)\ \cdots\ c_{n_+}^+ P_{n_-,n_+-1}(k)], \tag{4.5a}$$

and let us consider the limit of the $w_i^+(k)$, $i=1,\cdots,n_+$, as $k\to\infty$. We shall seek to express the $P_{n_-,j}(k)$, $j=1,\cdots,n_+-1$, in terms of the $w_i^+(k)$ for $l<k$, to obtain thereby a recursion equation for the $w_i^+(k)$. In fact, we need to also introduce quantities $w_i^-(k)$, $i=1,\cdots,n_-$, and obtain recursive equations with the $w_i^+(k)$ and $w_j^-(k)$ linked.

If

$$\frac{w_-(z)}{f_-(z)}=\alpha+w_{n_-}^- z^{-1}+\cdots+w_1^- z^{-n_-},$$

we have

$$[w_1^- \quad \cdots \quad w_{n_-}^-] = [c_1^- \quad c_2^- + P_{1,n_+} \quad \cdots \quad c_{n_-}^- + P_{n_- - 1, n_+}]. \tag{4.5b}$$

The $w_i^-(k)$ are defined obviously.

We make two important observations, both verifiable by direct calculation. First,

$$(A_-' P_k b_+ + c_-)\alpha(k)^{-1}(b_-' P_k A_+ + c_+') = \begin{bmatrix} w_1^-(k) \\ \vdots \\ w_{n_-}^-(k) \end{bmatrix} [w_1^+(k) \quad \cdots \quad w_{n_+}^+(k)], \tag{4.6}$$

and second,

$$A_-' P_k A_+ = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & P_{11}(k) & & P_{1,n_+-1}(k) \\ \vdots & & & \vdots \\ 0 & P_{n_- - 1,1}(k) & \cdots & P_{n_- - 1, n_+ - 1}(k) \end{bmatrix}. \tag{4.7}$$

Assuming with no loss of generality that $n_- \leqq n_+$, we obtain from these observations and the Riccati equation (4.4)

$$\begin{aligned} P_{k+1} &= \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & P_{11}(k) & & P_{1,n_+-1}(k) \\ \vdots & \vdots & & \vdots \\ 0 & P_{n_- - 1,1}(k) & \cdots & P_{n_- - 1, n_+ - 1}(k) \end{bmatrix} - \begin{bmatrix} w_1^-(k) \\ w_2^-(k) \\ \vdots \\ w_{n_-}^-(k) \end{bmatrix} [w_1^+(k) \quad w_2^+(k) \cdots w_{n_+}^+(k)] \\ &= \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & & 0 \\ 0 & 0 & P_{11}(k-1) & & P_{1,n_+-2}(k-1) \\ \vdots & & & & \vdots \\ 0 & 0 & P_{n_- - 2,1}(k-1) & \cdots & P_{n_- - 2, n_+ - 2}(k-1) \end{bmatrix} - \begin{bmatrix} 0 \\ w_1^-(k-1) \\ w_2^-(k-1) \\ \vdots \\ w_{n_- - 1}^-(k-1) \end{bmatrix} \begin{bmatrix} 0 \\ w_1^+(k-1) \\ w_2^+(k-1) \\ \vdots \\ w_{n_+ - 1}^+(k-1) \end{bmatrix}' \\ &\quad - \begin{bmatrix} w_1^-(k) \\ w_2^-(k) \\ w_3^-(k) \\ \vdots \\ w_{n_-}^-(k) \end{bmatrix} \begin{bmatrix} w_1^+(k) \\ w_2^+(k) \\ w_3^+(k) \\ \vdots \\ w_{n_+}^+(k) \end{bmatrix}' \end{aligned}$$

$$= -\begin{bmatrix} 0 \\ \vdots \\ 0 \\ w_1^-(k-n_-+1) \end{bmatrix}\begin{bmatrix} 0 \\ \vdots \\ w_1^+(k-n_-+1) \\ \vdots \\ w_{n_+-n_-+1}^+(k-n_-+1) \end{bmatrix}' \cdots - \begin{bmatrix} 0 \\ w_1^-(k-1) \\ \vdots \\ w_{n_--1}^-(k-1) \end{bmatrix}\begin{bmatrix} 0 \\ w_1^+(k-1) \\ \vdots \\ w_{n_+-1}^+(k-1) \end{bmatrix}'$$

$$- \begin{bmatrix} w_1^-(k) \\ w_2^-(k) \\ \vdots \\ w_{n_-}^-(k) \end{bmatrix}\begin{bmatrix} w_1^+(k) \\ w_2^+(k) \\ \vdots \\ w_{n_+}^+(k) \end{bmatrix}'$$

This equation implies

$$P_{n_-,n_+}(k+1) = -\sum_{i=0}^{n_--1} w_{n_--i}^-(k-i)w_{n_+-i}^+(k-i),$$

$$P_{n_-,j}(k+1) = -\sum_{i=0}^{j-1} w_{n_--i}^-(k-i)w_{j-i}^+(k-i) \quad \text{for } j<n_+, \tag{4.8}$$

$$P_{j,n_+}(k+1) = -\sum_{i=0}^{j-1} w_{j-i}^-(k-i)w_{n_+-i}^+(k-i) \quad \text{for } j<n_-.$$

Finally, equations (4.8) inserted into (4.5) define the recursion for the $w_i^+(k)$ and $w_i^-(k)$.

Provided the quantity $\alpha(k)=1+P_{n_-,n_+}(k)$ (expressible in terms of the $w_i^+(\cdot)$ and $w_i^-(\cdot)$ via (4.8)) is never zero, a recursion using solely the $w_i^+(\cdot)$ and $w_i^-(\cdot)$ is a very appealing one for solving the factorization problem, since each step in the recursion requires the computation of only n_++n_-+1 quantities, rather than $n_+\times n_-$ as when the full Riccati equation is solved. We argue in the next section that $\alpha(k)$ is almost never zero, so that one will almost always be safe using this sort of recursion. This is not to say, though, that numerical difficulties may not be encountered when $|\alpha(k)|$ is small.

Now we shall reinterpret this idea as one of L-U factorization. This reinterpretation may make possible the easy application of standard programs to the factorization problem. Consider the following banded matrix (where, for the moment, the φ_{ij} have no special significance)

$$\Phi = \begin{bmatrix} \varphi_{11} & \varphi_{12} & \varphi_{13} & 0 & 0 & 0 & \cdots \\ \varphi_{21} & \varphi_{22} & \varphi_{23} & \varphi_{24} & 0 & 0 & \cdots \\ \varphi_{31} & \varphi_{32} & \varphi_{33} & \varphi_{34} & \varphi_{35} & 0 & \cdots \\ \varphi_{41} & \varphi_{42} & \varphi_{43} & \varphi_{44} & \varphi_{45} & \varphi_{46} & \cdots \\ 0 & \varphi_{52} & \varphi_{53} & \varphi_{54} & \varphi_{55} & \varphi_{56} & \cdots \\ 0 & 0 & \varphi_{63} & \varphi_{64} & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & & & \end{bmatrix},$$

and factor it as a product of a lower triangular matrix L and an upper triangular matrix U as follows.

$$LU = \begin{bmatrix} l_{11} & 0 & 0 & 0 & \cdots \\ l_{21} & l_{22} & 0 & 0 & \cdots \\ l_{31} & l_{32} & l_{33} & 0 & \cdots \\ l_{41} & l_{42} & l_{43} & l_{44} & \cdots \\ 0 & l_{52} & l_{53} & \cdots & \cdots \\ \vdots & \vdots & \vdots & & \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{13} & 0 & 0 & \cdots \\ 0 & u_{22} & u_{23} & u_{24} & 0 & \cdots \\ 0 & 0 & u_{33} & u_{34} & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & & \end{bmatrix}.$$

To find L and U, the calculations commence as below.

$$\begin{aligned} l_{11}u_{11} &= \varphi_{11} && \text{choose } l_{11} = \varphi_{11},\ u_{11} = 1, \\ u_{12} &= l_{11}^{-1}\varphi_{12} && \text{assuming } \varphi_{11} \neq 0, \\ u_{13} &= l_{11}^{-1}\varphi_{13}, \\ l_{21} &= \varphi_{21}, \\ l_{31} &= \varphi_{31}, \\ l_{41} &= \varphi_{41}; \\ l_{21}u_{12} + l_{22}u_{22} &= \varphi_{22}, && \text{choose } l_{22} = \varphi_{22} - l_{21}u_{12},\ u_{22} = 1, \\ u_{23} &= l_{22}^{-1}(\varphi_{23} - l_{21}u_{13}) && \text{assuming } \varphi_{22} - l_{21}u_{12} \neq 0. \end{aligned}$$

In general, we have the formulas, for all k,

$$\begin{aligned} l_{kk} &= \varphi_{kk} - l_{k,k-1}u_{k-1,k} - l_{k,k-2}u_{k-2,k} - \cdots \\ &= \varphi_{kk} - \sum_{i=1}^{k-1} l_{k,k-i}u_{k-i,k}, \\ u_{kk} &= 1, \\ l_{k+j,k} &= l_{kk}^{-1}\left(\varphi_{k+j,k} - \sum_{i=1}^{k-1} l_{k+j,k-i}u_{k-i,k}\right) \quad \text{for } j > 0, \end{aligned} \tag{4.9}$$

assuming $l_{kk} \neq 0$ for $j > 0$,

$$u_{k,k+j} = \varphi_{k,k+j} - \sum_{i=1}^{k-1} l_{k,k-i}u_{k-i,k+j}.$$

Now suppose Φ is of the following form:

$$\begin{bmatrix} \vdots & & \vdots & & & & & & & & & & \\ 0 & \cdots & 0 & c_1^+ & c_2^+ & \cdots & c_{n_+}^+ & 1 & c_n^- & \cdots & c_2^- & c_1^- & 0 & \cdots \\ 0 & \cdots & 0 & 0 & c_1^+ & \cdots & \cdots & c_{n_+}^+ & 1^- & c_{n-}^- & \cdots & c_2^- & c_1^- & 0 & \cdots \\ \vdots & & \vdots & & & & & & & & & & & & \end{bmatrix}, \tag{4.10}$$

so that

$$\varphi_{kk} = 1 \qquad \text{for all } k,$$
$$\varphi_{k,k+j} = c^{-}_{n_-+1-j}, \qquad j = 1, \cdots, n_-, \quad \text{all } k,$$
$$\varphi_{k+j,k} = c^{+}_{n_++1-j}, \qquad j = 1, \cdots, n_+, \quad \text{all } k.$$

The key point is the observation that on setting

$$l_{kk} = \alpha(k) \qquad \text{for all } k,$$
$$l_{k+j,k} = w^{+}_{n_-+1-j}(k), \qquad j = 1, \cdots, n_-, \quad \text{all } k,$$
$$u_{k,k+j} = w^{-}_{n_-+1-j}(k), \qquad j = 1, \cdots, n_+, \quad \text{all } k,$$

the equations (4.9) become identical to the recursion for the $w_i^{\pm}(k)$ defined by the equations (4.8), (4.5a) and (4.5b).

As noted earlier, this method will break down in any situation where the associated Riccati equation (4.4) breaks down by having $\alpha_k = 0$. Repair of the breakdown via the linear equation seems more straightforward for the Riccati equation than the L-U factorization approach.

5. Discussion and examples.

5.1. Location of zeros. An essential preliminary to the factorization method presented in the preceding sections is (i) the determination of n_+ and n_- for both the discrete-time and continuous-time problems, and (ii) the checking for zeros on the unit circle and imaginary axis, both in a finite number of rational operations. There are many possible approaches to executing these tasks; here we will outline one method for each of the continuous-time and discrete-time problems.

Let $m(s)$ be a real polynomial. To check that $m(j\omega) \neq 0$ for any real ω, define real polynomials $f_1(\bar{\omega})$, $f_2(\bar{\omega})$ by

$$m(j\omega) = f_1(\bar{\omega}) + j\omega f_2(\bar{\omega}), \qquad \bar{\omega} = \omega^2.$$

Then $m(j\omega) = 0$ for some real ω if and only if $f(\bar{\omega}) = f_1^2(\bar{\omega}) + \bar{\omega} f_2^2(\bar{\omega}) = 0$ for some $\bar{\omega} \geqq 0$. This latter condition can be checked in a finite number of rational operations via Sturm's theorem [19, Chap. 15].

Given that $m(s)$ has no pure imaginary zeros, form the Hermite matrix H [20]. This matrix has entries quadratic in the coefficients of $m(\cdot)$, and is symmetric. Suppose that H has n_1 positive, n_2 negative and n_3 zero eigenvalues, respectively. These quantities are computable in a finite number of rational operations (as H is a real, symmetric matrix) [19, Chap. 10]. Then it is known [20] that $m(s)$ has $n_+ = n_1 + n_3/2$ zeros in $\mathrm{Re}[s] < 0$ and $n_- = n_1 + n_3/2$ zeros in $\mathrm{Re}[s] > 0$.

For the discrete-time case, we describe an iterative method involving the Schur–Cohn matrix and the following result [20]: for any real, self-inversive polynomial $g(z)$, both $g(z)$ and $g'(z) = d(g(z))/dz$ have the same number of zeros outside the unit circle and $g'(z)$ is not self-inversive.

Given a real $m(z)$, one can form the Schur–Cohn matrix M and determine n_1, n_2 and n_3, the numbers of positive, negative and zero eigenvalues of M, respectively. Then [20] $m(z)$ has

(i) n_1 zeros inside the unit circle, not the reciprocal of any zero outside the unit circle,

(ii) n_2 zeros outside the unit circle, not the reciprocal of any zero inside the unit circle,

(iii) n_3 zeros either on the unit circle or reciprocal to each other.

Let $m_r(z)$ be a polynomial of degree $n^{(r)}$, with $n_1^{(r)}$, $n_2^{(r)}$ and $n_3^{(r)}$ describing the eigenvalue distribution of the Schur–Cohn matrix M_r. Then the self-inversive part of $m_r(z)$, given by the greatest common divisor of $m_r(z)$ and $z^{n^{(r)}}m_r(z^{-1})$ and denoted by $g_r(z)$, has degree $n_3^{(r)}$ and is computable in a finite number of rational operations, e.g., by Euclids algorithm. Set $m_{r+1}(z) = g_r'(z)$. Then, with $m(z) = m_1(z)$ as the initial polynomial, the passage from $m_r(z)$ to $m_{r+1}(z)$ above is the rth stage of an iterative scheme which can identify the number of zeros of $m(z)$ outside the unit circle. Since $n_3^{(r+1)} \leqq n_3^{(r)} - 2$ at each stage (because $m_{r+1}(z)$ has degree less than that of $g_r(z)$, which is n_r, and is not self-inversive), the procedure will terminate in at most $n_3^{(0)}/2$ steps, i.e., $n_3^{(R)} = 0$ for some $R \leqq n_3^{(0)}/2$.

Finally, the number of zeros of $m_r(z)$ outside the unit circle $= n_2^{(r)}$ + number of zeros of $g_r(z)$ outside the unit circle $= n_2^{(r)}$ + number of zeros of $m_{r+1}(z)$ outside the unit circle, and thus we have $n_- = \sum_{i=0}^{R} n_2^{(i)}$.

5.2. Matrix generalization. As mentioned previously, the scalar symmetric factorization problem ($m(s) = \pm m(-s)$) has a relatively straightforward generalization to the matrix case, in terms of both solution existence and solution computability via the Riccati equation. In contrast, the Riccati equation treatment of the nonsymmetric scalar problem does not appear to generalize. We have earlier indicated precisely *where* the Riccati equation approach to factorization fails to extend: here we shall view the issue from another angle, which suggests *why* it fails to extend.

In the scalar case, we used the Riccati procedure to factor a real rational function $m(s)/\psi(s)$ as a product $[m_+(s)/\psi_+(s)][m_-(s)/\psi_-(s)]$ under the assumptions that $m(s)$ and $\psi(s)$ are nonzero for all $s = j\omega$, ω real, and $\lim_{s\to\infty} m(s)/\psi(s)$ is finite and nonzero; moreover, m_+/ψ_+ and m_-/ψ_- were analytic, together with their inverses,[4] throughout $\mathrm{Re}[s] \geqq 0$ and $\mathrm{Re}[s] \leqq 0$, respectively, including $s = \infty$.

The natural matrix generalization would be to obtain a factorization

$$\frac{M(s)}{\psi(s)} = \frac{M_+(s)}{\psi_+(s)} \frac{M_-(s)}{\psi_-(s)},$$

where $M(s)$, $M_+(s)$ and $M_-(s)$ are $n \times n$ real polynomial matrices, nonsingular, respectively, for $s = j\omega$, ω real, for $\mathrm{Re}[s] \geqq 0$ and for $\mathrm{Re}[s] \leqq 0$; $\psi(s)$, $\psi_+(s)$ and $\psi_-(s)$ are polynomials, nonzero in the same regions, and M/ψ, M_+/ψ_+ and M_-/ψ_- are analytic together with their inverses at $s = \infty$.

It is, however, a standard result that such factorizations are *not* necessarily possible. According to Gohberg and Krein [21], a real rational matrix $H(s)$,

[4] Strictly, this is not quite the case; here, we are implying that $\psi_+(s)$ must have all zeros in $\mathrm{Re}[s] < 0$, whereas this restriction was not made earlier. Its introduction here is helpful however in explaining the main point.

nonsingular for all $s = j\omega$ including $\omega = \infty$, possesses a *left standard factorization*

$$H(s) = L_+(s)D(s)L_-(s)$$

with the following properties: $L_+(s)$ and $L_-(s)$ are nonsingular together with their inverses in $\mathrm{Re}\,[s] \geqq 0$ and $\mathrm{Re}\,[s] \leqq 0$, respectively, including $s = \infty$, and

$$D(s) = \mathrm{diag}\left[\left(\frac{s-1}{s+1}\right)^{\alpha_1}, \cdots, \left(\frac{s-1}{s+1}\right)^{\alpha_n}\right],$$

where the integers α_i are positive, negative or zero, satisfy $\alpha_1 \geqq \alpha_2 \cdots \geqq \alpha_n$ and are uniquely determined by $H(s)$. The desired matrix generalization of the scalar factorization result therefore corresponds to the case $\alpha_i = 0$ for all i. That there are factorizations with α_i nonzero is easily checked by example:

$$H(s) = \begin{bmatrix} 0 & \dfrac{s-1}{s+1} \\ \dfrac{s+1}{s-1} & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \dfrac{s-1}{s+1} & 0 \\ 0 & \left(\dfrac{s-1}{s+1}\right)^{-1} \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Despite the possible nonexistence of the desired sort of factorization for matrices with rational entries, polynomial factorization can be achieved via an alternative procedure. Suppose $M(s)$ is a polynomial matrix and nonsingular for all $s = j\omega$. Let $M(s)$ be factored (using a finite number of rational operations) to exhibit the Smith canonical form [19]; thus

$$M(s) = P(s)B(s)Q(s),$$

where $P(s)$ and $Q(s)$ are polynomial matrices with constant determinants and $B(s)$ is a diagonal matrix of polynomials, nonsingular for all $s = j\omega$, ω real. Using the *scalar* polynomial factorization procedure, we can obtain $B(s) = B_+(s)B_-(s)$, with $B_+(s)$ and $B_-(s)$ diagonal matrices of polynomials nonzero in $\mathrm{Re}\,[s] \geqq 0$ and $\mathrm{Re}\,[s] \leqq 0$. Then with $M_+ = PB_+$, $M_- = B_-Q$, a factorizing of $M(s)$ is defined as $M_+(s)M_-(s)$, with $M_+(s)$ and $M_-(s)$ nonsingular in $\mathrm{Re}\,[s] \geqq 0$ and $\mathrm{Re}\,[s] \leqq 0$, respectively. It is easily established incidentally that $M_+(s)$ and $M_-(s)$ are unique to within multiplication by a polynomial matrix with constant determinant.

An open question still remaining is whether there might be some way of passing from a polynomial $M(s)$ to a rational $H(s)$ which possesses a left standard factorization with $D(s) = I$, whether also this factorization could be obtained using a Riccati equation technique and could in this way be used to provide a factorization $M(s) = M_+(s)M_-(s)$ with $M_+(s)$ and $M_-(s)$ nonsingular in $\mathrm{Re}\,[s] \geqq 0$ and $\mathrm{Re}\,[s] \leqq 0$, respectively.

5.3. Finite escape times. Those familiar with the occurrence of Riccati equations in the symmetric factorization problem, where no escape times occur, might conjecture that no escape times need occur in the nonsymmetric factorization problem. The following two examples show, however, that both the Riccati differential equation defined in (2.22) and the Riccati difference equation defined in (3.14) may have finite escape times.

Example 1. *Continuous-time case.* The polynomial $m(s) = s^3+s^2+(\alpha^2-1)s-\alpha^2-1$, where $\alpha>0$, has zeros at $-1\pm j\alpha$ and 1. Then, with $f_+(s)=(s+1)(s+2)$ and $f_-(s)=-s-2$, set

$$A_+=\begin{bmatrix}0 & 1\\ -2 & -3\end{bmatrix},\quad b_+=\begin{bmatrix}0\\1\end{bmatrix},\quad c_+=\frac{1}{12}\begin{bmatrix}7\alpha^2-9\\ -(\alpha^2+9)\end{bmatrix},$$

$$A_-=-2,\qquad b_-=1,\qquad c_-=-\frac{1}{12}(\alpha^2+9),$$

the notation being that defined in § 2. In terms of the notation of § 2, the matrix $P(t)$ will fail to exist at t if the matrix $X(t)$ is singular. Denoting the determinant of $X(t)$ by $D(t)$ to within an inessential scaling constant, a long calculation shows that

$$D(t)=2\beta(\beta+4)(\beta+9)\,e^{2t}+2\beta(\beta+4)(\beta-1)\cos\alpha t+\tfrac{1}{3}\alpha\beta(\beta-11)(\beta+4)\sin\alpha t,$$

with $\beta=\alpha^2$.

For large α, the coefficient of e^{2t} is $O(\beta^3)$, while that of $\sin\alpha t$ is $O(\alpha\beta^3)$, and thus it is clear that for α sufficient large, $D(t)$ takes both positive and negative values for t in some neighborhood of the origin. Hence $D(t)$ is zero for some value of $t>0$, and therefore $P(t)$ has a finite escape time. Notice also that as $t\to\infty$, the term $2\beta(\beta+4)(\beta+9)\,e^{2t}$ dominates, guaranteeing that $D(t)$ is nonzero for suitable large t.

Example 2. *Discrete-time case.* The polynomial $m(z)=3(1+\alpha^2)z^3-(19+7\alpha^2)z^2+(33+5\alpha^2)z-(9-\alpha^2)$, has zeros at $\frac{1}{3}$, $[(3+\alpha^2)\pm j2\alpha]/(1+\alpha^2)$. With $f_+(z)=z^2$, $f_-(z)=z$, set

$$A_+=0,\qquad b_+=1,\qquad c_+=-(9+\alpha^2),$$

$$A_-=\begin{bmatrix}0 & 1\\ 0 & 0\end{bmatrix},\quad b_-=\begin{bmatrix}0\\1\end{bmatrix},\quad c_-=\begin{bmatrix}3(1+\alpha)^2)\\ -(19+7\alpha^2)\end{bmatrix},$$

the notation being that defined in § 3.

It can then be shown that the discrete-time Riccati equation has no escape times provided $X_k(\alpha)\neq 0$ for any $k=0,1,2,\cdots$, where

$$X_k(\alpha)=1-\left(\frac{r(\alpha)}{3}\right)^k\left\{\frac{5\alpha^2+17\alpha}{9\alpha^2+81}\cos k\theta(\alpha)+\frac{\alpha^4+\alpha^2+12}{9\alpha^3+81\alpha}\sin k\theta(\alpha)\right\},$$

$$r^2(\alpha)=\frac{1+\alpha^2}{9+\alpha^2},$$

$$\tan\theta(\alpha)=\frac{3+\alpha^2}{2\alpha},\quad\text{with } |\theta(\alpha)|<\pi/2.$$

For α sufficiently large,

$$X_k(\alpha)\sim 1-\frac{5}{9}\left(-\frac{1}{9}\right)^{k/2}\qquad\text{for } k=0,2,4,\cdots,$$

$$X_k(\alpha)\sim 1-\left(\frac{1}{3}\right)^k(-1)^{(k-1)/2}\frac{\alpha}{9}\qquad\text{for } k=1,3,5,\cdots. \tag{5.1}$$

Thus, for each $k = 1, 5, 9, \cdots$, there is at least one value of α for which $X_k(\alpha) = 0$. Figure 1 shows a plot of $X_k(\alpha)$ versus α for $k = 0, 1, 2, 3, 4, 5$, confirming the asymptotic behavior (5.1). However, Figure 1 suggests that there is only one value of α for which $X_k(\alpha) = 0$ for each $k = 1, 5, 9, \cdots$, i.e., $X_k(\alpha) = 0$ only for *isolated* values of α.

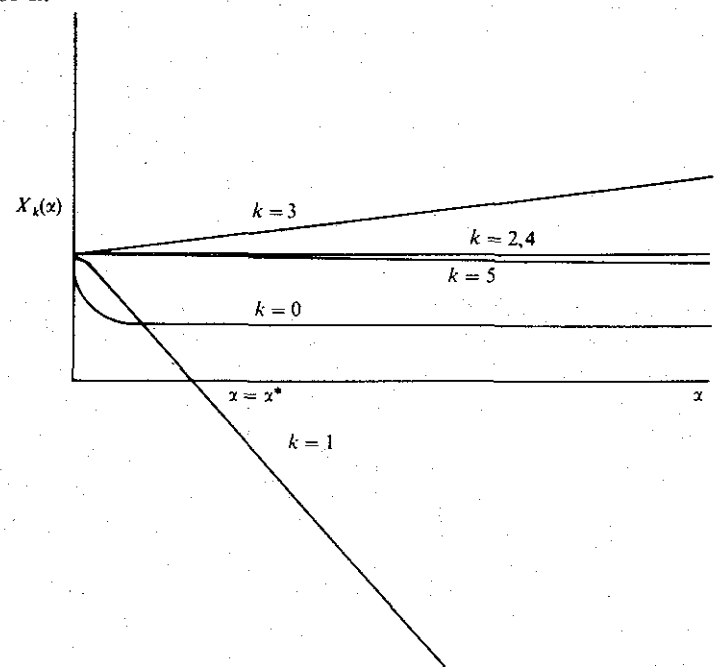

FIG. 1. *Plot of $X_k(\alpha)$ versus α for values of $k = 0, 1, 2, 3, 4, 5$. A point of intersection with the α-axis, e.g., $(\alpha^*, 0)$, indicates that for $\alpha = \alpha^*$, the associated discrete-time Riccati equation will have a finite escape time, viz., $k = 1$ for $\alpha = \alpha^*$.*

It would therefore appear that there is a basic difference between the probability of occurrence of finite escape times of the Riccati differential equation and the probability of occurrence of finite escape times of the Riccati difference equation, given that the quantities $A_+, b_+, c_+, A_-, b_-, c_-$ are chosen at random. It might be conjectured that the discrete-time Riccati equation has no finite escape times with probability 1, a fact which has some significance in terms of justifying the use of the discrete-time Riccati equation as a computational tool.

5.4. Computational aspects. For the continuous-time case, $X(t)$, $Y(t)$ as defined in (2.24), (2.25) were derived subject to the initial condition $X(0) = I$, $Y(0) = 0$. A similar derivation shows that, for the initial condition $X(0) = A$, $Y(0) = B$, $P(t)$ exists for large t and converges to $T_{21}T_{11}^{-1}$ as $t \to \infty$, provided $A - T_{12}T_{22}^{-1}B$ is nonsingular. Similar comments hold for the discrete-time case. This would suggest that computationally, the discrete-time algorithm has a degree of numerical stability.

The rate of convergence in the continuous and discrete time cases to the steady-state solution of the Riccati equation is easily seen to be exponential. For example, in the discrete time case, the material of § 3 shows, using the notation of

that section, that

$$P_k = [T_{21} - T_{22}\Lambda_-^{-k}T_{22}^{-1}T_{21}\Lambda_+^k][T_{11} - T_{12}\Lambda_-^{-k}T_{22}^{-1}T_{21}\Lambda_+^k]^{-1}.$$

The rate of exponential convergence is evidently determined by the distance between the unit circle and the set of zeros of $m(z)$.

Finally, we remark that the algorithm described in § 3 has been implemented on a digital computer. Various zero patterns for the polynomial $m(z)$were tried, and in all cases tested, there were no escape times of the discrete-time Riccati equation, the algorithm converged to the correct value and at a rate close to that determined above. Typically, for zero patterns of $m(z)$ with no zero within 0.25 of the unit circle boundary, the number of iterations required for convergence to 5 significant figures was less than 30.

5.5. Choice of the polynomials $f_+(s), f_-(s), f_+(z), f_-(z)$. The initial choice of $f_+(s)$, $f_-(s)$, etc., has been restricted only by conditions of coprimeness and the like, except in § 4 where specific $f_+(z)$ and $f_-(z)$ were adopted. One might ask what advantage could be gained by making special choices of these polynomials.

First, it should be noted that the rate of convergence of the various algorithms is unaffected, the rate being determined by the zeros of $m(\cdot)$. However, one would expect escape times of the Riccati differential equation and the occurrence of numerical difficulties caused by very large values P_k in the discrete-time equation to be linked with the choice of $f_+(s)$, etc. It does not however seem possible to explain helpful features of the linkage. Again, one might expect difficulties if, for example, a zero of $f_+(s)$ was close to a zero of $m(s)$.

More fundamentally, one might hope for a simplification of the theory in the case of a specially chosen $f_+(s)$, etc. Indeed, this is what is exhibited in § 4 for the discrete-time problem; possibly an analogous simplification could be found for the continuous-time problem.

5.6. Minor points concerning the steady state Riccati equation. In lieu of (2.4), one can obtain results from the "dual" equation

$$A_+Q + QA_-' = (Qc_- + B_+)(c_+'Q + B_-'). \tag{5.2}$$

The calculations paralleling (2.5) and the following equations are straightforward, and will not be performed here. Similar remarks hold for the discrete-time problem.

Secondly, in, for example, [6], ordering properties of solutions of the spectral factorization equivalent of (2.4) are discussed. Using the partial ordering defined for symmetric matrices by the nonnegative definite property, a partial ordering on the solutions can be shown to reflect a partial ordering on sets of eigenvalues of the matrix M. A complete parallel is not possible here, although indirectly one can conceive of a partial ordering on the solutions of (2.4) induced by the eigenvalue set of M used in defining P, but with the partial ordering possessing no other apparent significance.

REFERENCES

[1] G. SALOMONSSON, *An equalizer with feedback filter*, Ericsson Techniques, 28 (1972), pp. 57–101.

[2] E. WONG AND J. B. THOMAS, *On the multidimensional prediction and filtering problem and the factorization of spectral matrices*, J. Franklin Inst., 272 (1961), pp. 87–99.

[3] D. C. YOULA, *On the factorization of rational matrices*, IRE Trans. Information Theory, IT-7 (1961), pp. 172–189.

[4] B. D. O. ANDERSON, K. L. HITZ AND N. DIEM, *Recursive algorithms for spectral factorization*, IEEE Trans. Circuit Theory, CAS-21 (1974), pp. 742–750.

[5] B. L. HO AND R. E. KALMAN, *Spectral factorization using the Riccati equation*, Proc. 4th Allerton Conf. on Circuit and System Theory, 1966, pp. 388–399.

[6] B. D. O. ANDERSON AND S. VONGPANITLERD, *Network Analysis and Synthesis—A Modern Systems Theory Approach*, Prentice-Hall, Englewood Cliffs, N.J., 1973.

[7] F. L. BAUER, *Beiträge zur Entwicklung numerischer Verfahren für programmgesteuerte Rechenanlagen. I. Quadratisch konvergente Durchführung der Bernoulli–Jacobischen Methode zur Nullstellenbestimmung von Polynomen*, Sitz. Ber. Bayer. Akad. Wiss., (1954), pp. 275–303.

[8] ———, *Beiträge zur Entwicklung numerischer Verfahren für programmgesteuerte Rechenanlagen. II. Direkte Faktorisierung eines Polynoms*, Ibid., (1956), pp. 163–203.

[9] J. D. ROBERTS, *Linear model reduction and solution of the algebraic Riccati equation by use of the sign function*, Rep. TR13, Dept. of Engrg., Cambridge Univ., 1971.

[10] A. RALSTON, *A First Course in Numerical Analysis*, McGraw-Hill, New York, 1965.

[11] F. L. BAUER, *Ein direktes Interationsverfahren zur Hurwitz–Zerlegung eines Polynoms*, Arch. Elek. Übertr., 9 (1955), pp. 285–290.

[12] R. W. BROCKETT, *Finite Dimensional Linear Systems*, John Wiley, New York, 1970.

[13] V. KUCERA, *A contribution to matrix quadratic equations*, IEEE Trans. Automatic Control, AC-17 (1972), pp. 844–847.

[14] K. MARTENSSON, *On the matrix Riccati equation*, Information Sci., 3 (1971), pp. 17–49.

[15] J. E. POTTER, *Matrix quadratic solutions*, this Journal, 14 (1966), pp. 496–501.

[16] W. T. REID, *A matrix differential equation of the Riccati type*, Amer. J. Math., 68 (1946), pp. 237–246.

[17] D. R. VAUGHAN, *A negative exponential solution to the matrix Riccati equation*, IEEE Trans. Automatic Control, AC-14 (1969), pp. 72–75.

[18] V. KUCERA, *The discrete Riccati equation of optimal control*, Kybernetika, 8 (1972), pp. 430–447.

[19] F. R. GANTMACHER, *The Theory of Matrices*, vols. I and II, Chelsea, New York, 1959.

[20] M. MARDEN, *Geometry of Polynomials*, American Mathematical Society, Providence, R. I., 1966.

[21] I. C. GOHBERG AND M. G. KREIN, *Systems of integral equations on a half line with kernels depending on the difference of arguments*, Amer. Math. Soc. Transl., 14 (1960), pp. 217–287.