BULL. AUSTRAL. MATH. SOC.
VOL. 10 (1974), 245-253.

34A40

A special property of the matrix Riccati equation

A.N. Stokes

In the domain of real symmetric matrices ordered by the positive definiteness criterion, the symmetric matrix Riccati differential equation has the unusual property of preserving the ordering of its solutions as the independent variable changes. Here is is shown that, subject to a continuity restriction, the Riccati equation is unique among comparable equations in possessing this property.

0. Introduction

The matrix Riccati equation has attracted attention recently because of its occurrence in a number of different situations. Its solutions determine solutions of the optimal linear regulator problem (Kalman [4], Athans and Falb [1]); the question of whether or not it has a solution on an interval is related to the question of the disconjugacy of a linear hamiltonian system on an interval (Reid [5], Coppel [2]), and Schumitzky [6] has demonstrated an equivalence between matrix Riccati equations and Fredholm resolvents. Recently, Fair [3] has written about continued fraction solutions of a general Riccati equation, which the author [7] has investigated from a different direction.

This paper provides a further reason why the matrix Riccati equation, at least in its symmetric form, is of special interest. Theorem 1 below

Received 13 November 1973. This work was done with the aid of a Postgraduate Research Scholarship of the Australian National University. The idea that the result embodied in the main theorem is necessary arose in a discussion with Mr W.A. Coppel, whose helpful suggestions are gratefully acknowledged.

asserts that the Riccati equation, when $n > 1$, is unique in possessing the order-preserving property defined in the next section.

1. The order-preserving property

Consider the symmetric matrix equation

$$W' = F(t, W) \tag{1}$$

where F is a symmetric $n \times n$ matrix-valued function defined for all t and all symmetric $n \times n$ matrices W, and is continuous in W for each t. The prime denotes differentiation with respect to t. Then if $n > 1$, the Riccati equation is the only such equation with the following property:

DEFINITION. (1) has the *order-preserving property* if, whenever W_1, W_2 are symmetric matrices with $W_1 \geq W_2$, for any point a there is a neighbourhood $[b, c]$, $b < a < c$, on which two solutions $W_1(t)$, $W_2(t)$ of (1), with $W_1(a) = W_1$, $W_2(a) = W_2$, exist and obey $W_1(t) \geq W_2(t)$ on $[b, c]$. (By $W_1 \geq W_2$, we mean that $W_1 - W_2$ is non-negative definite.)

2. Main theorem

THEOREM 1. *If* (1) *is the equation as defined in the previous section, having the order-preserving property, and if* $n > 1$, *then* $F(t, W)$ *must be a function which can be written in the form*

$$F(t, W) = A(t) + B(t)W + WB^*(t) + WC(t)W \tag{2}$$

where $A(t)$, $B(t)$ *and* $C(t)$ *are* $n \times n$ *matrices, and* $A(t)$, $C(t)$ *are symmetric for all* t.

Proof. Suppose (1) indeed has the order-preserving property. Let W_1, W_2 be two $n \times n$ symmetric matrices having the property that there is a vector x for which $W_1 x = W_2 x$. Then there is a symmetric matrix for which $W_3 \geq W_1$, $W_3 \geq W_2$ and $W_1 x = W_3 x = W_2 x$.

Then for any point t, there exists an interval (b, c), $b < t < c$ and two solutions $W_1(u)$, $W_3(u)$ of (1) existing on (b, c) for which

https://doi.org/10.1017/S0004972700040855 Published online by Cambridge University Press

$W_1(t) = W_1$, $W_3(t) = W_3$ and $W_3(u) \geq W_1(u)$ on (b, c) .

But $x^*W_1(t)x = x^*W_3'(t)x$, and $x^*\big(W_3'(u)-W_1(u)\big)x \geq 0$ near and on either side of t .

Therefore $x^*W_1'(t)x = x^*W_3'(t)x$; that is,

$$x^*F\big(t, W_1(t)\big)x = x^*F\big(t, W_3(t)\big)x .$$

Similarly, $x^*F\big(t, W_2(t)\big)x = x^*F\big(t, W_3(t)\big)x$. So

$$W_1x = W_2x \Rightarrow x^*F\big(t, W_1\big)x = x^*F\big(t, W_2\big)x . \tag{3}$$

Henceforth mention of t is suppressed, and we define a function g from $R^n \times R^n$ to R by

$$g(x, y) = x^*F(W)x \text{ , where } y = Wx . \tag{4}$$

The function g is well-defined if $x \neq 0$, for if there are two matrices W_1 and W_2 with $W_1x = W_2x$, then (3) implies that each gives the same value of $x^*F(W)x$.

When $n > 1$, (4) restricts F to the form of a quadratic function. The rest of the proof consists of a manipulation of (4) to demonstrate this fact.

Let e_i be the unit vector whose i-th component is 1 . Then $F_{ii}(W) = e_i^*F(W)e_i = g\big(e_i, We_i\big)$. So $F_{ii}(W)$ is a function only of the coefficients W_{ij} , $j = 1, \ldots, n$. And

$$\begin{aligned} 2F_{ij}(W) &= \big(e_i+e_j\big)^*F(W)\big(e_i+e_j\big) - F_{ii}(W) - F_{jj}(W) \\ &= g\big(e_i+e_j, We_i+We_j\big) - g\big(e_i, We_i\big) - g\big(e_j, We_j\big) . \end{aligned}$$

So $F_{ij}(W)$ is a function of W_{ik} and W_{jk} only, $k = 1 \ldots n$.

The problem is now artifically restricted to a 2×2 problem, as follows. For arbitrary i, j , $i \neq j$, let $W_1 = W_{ii}$, $W_2 = W_{ij}$, $W_3 = W_{jj}$, and assume during what follows that all other coefficients of W remain fixed. Let

https://doi.org/10.1017/S0004972700040855 Published online by Cambridge University Press

$$F_1(W) = F_{ii}(W) \ , \quad F_2(W) = F_{ij}(W) \ , \quad F_3(W) = F_{jj}(W) \ .$$

Suppressing constant coefficients, and letting α be any constant, (4) implies

$$(5) \quad \alpha^2 F_1(W_1, W_2) + 2\alpha F_2(W_1, W_2, W_3) + F_3(W_2, W_3) = \\ = g(\alpha, 1; \alpha W_1 + W_2, \alpha W_2 + W_3) \ .$$

Neither side of (5) is affected by the change

$$W_1 \to W_1 + \varepsilon \ ,$$

$$W_2 \to W_2 - \alpha\varepsilon \ ,$$

$$W_3 \to W_3 + \alpha^2\varepsilon \ ,$$

for any ε . Therefore,

$$(6) \quad \alpha^2 F_1(W_1+\varepsilon, W_2-\alpha\varepsilon) + 2\alpha F_2\left(W_1+\varepsilon, W_2-\alpha\varepsilon, W_3+\alpha^2\varepsilon\right) + F_3\left(W_2-\alpha\varepsilon, W_3+\alpha^2\varepsilon\right) \\ = \alpha^2 F_1(W_1, W_2) + 2\alpha F_2(W_1, W_2, W_3) + F_3(W_2, W_3) \ .$$

This is the basic equation to be manipulated; it is rewritten by first replacing α by $-\alpha$ and ε by $-\varepsilon$, then (W_1, W_2, W_3) by $\left(W_1+\varepsilon, W_2-\alpha\varepsilon, W_3+\alpha^2\varepsilon\right)$, so:

$$(7) \quad \alpha^2 F_1(W_1, W_2-2\alpha\varepsilon) - 2\alpha F_2(W_1, W_2-2\alpha\varepsilon, W_3) + F_3(W_2-2\alpha\varepsilon, W_3) \\ = \alpha^2 F_1(W_1+\varepsilon, W_2-\alpha\varepsilon) - 2\alpha F_2\left(W_1+\varepsilon, W_2-\alpha\varepsilon, W_3+\alpha^2\varepsilon\right) + F_3\left(W_2-\alpha\varepsilon, W_3+\alpha^2\varepsilon\right) \ .$$

Adding (6) and (7):

$$(8) \quad \alpha^2\left[F_1(W_1, W_2-2\alpha\varepsilon) - F_1(W_1, W_2)\right] + F_3(W_2-2\alpha\varepsilon, W_3) - F_3(W_2, W_3) \\ + 2\alpha\left[2F_2\left(W_1+\varepsilon, W_2-\alpha\varepsilon, W_3+\alpha^2\varepsilon\right) - F_2(W_1, W_2-2\alpha\varepsilon, W_3) - F_2(W_1, W_2, W_3)\right] = 0 \ .$$

Dividing by α and letting $\alpha \to 0$:

$$(9) \quad \lim_{\alpha\to 0} \frac{F_3(W_2-2\alpha\varepsilon, W_3) - F_3(W_2, W_3)}{2\alpha\varepsilon} = -\frac{2}{\varepsilon}\left[F_2(W_1+\varepsilon, W_2, W_3) - F_2(W_1, W_2, W_3)\right] \ .$$

https://doi.org/10.1017/S0004972700040855 Published online by Cambridge University Press

Therefore, $\frac{\partial F_3}{\partial W_2}(W)$ exists for all α, ε. We abbreviate $\frac{\partial F_i}{\partial W_j}$ by F_{ij}, $i, j = 1, 2, 3$. Then

$$F_2(W_1+\varepsilon, W_2, W_3) = F_2(W_1, W_2, W_3) + \tfrac{1}{2}\varepsilon F_{32}(W_2, W_3) . \tag{10}$$

So F_2 is a linear function of W_1 for fixed W_2, W_3.

In (8) let $\alpha \to \infty$ and $\varepsilon = \frac{t}{\alpha^2}$ for some constant t. Then dividing by α:

$$\lim_{\alpha\varepsilon\to 0} \left[\frac{t}{2\alpha\varepsilon}\left(F_1(W_1, W_2-2\alpha\varepsilon)-F_1(W_1, W_2)\right)\right] + 2\left[F_2(W_1, W_2, W_3+t)-F_2(W_1, W_2, W_3)\right] = 0 . \tag{11}$$

Therefore, $\frac{\partial F_1}{\partial W_2} = F_{12}(W_1, W_2)$ exists and

$$tF_{12}(W) = 2\left[F_2(W_1, W_2, W_3+t)-F_2(W_1, W_2, W_3)\right] , \tag{12}$$

that is, F_2 is also a linear function of W_3 for fixed W_1, W_2.

Rewriting (6), replacing α by $-\alpha$, (W_1, W_2, W_3) by $\left(W_1+\varepsilon, W_2-\alpha\varepsilon, W_3+\alpha^2\varepsilon\right)$, then

$$\alpha^2 F_1(W_1+2\varepsilon, W_2) - 2\alpha F_2\left(W_1+2\varepsilon, W_2, W_3+2\alpha^2\varepsilon\right) + F_3\left(W_2, W_3+2\alpha^2\varepsilon\right)$$
$$= \alpha^2 F_1(W_1+\varepsilon, W_2-\alpha\varepsilon) - 2\alpha F_2\left(W_1+\varepsilon, W_2-\alpha\varepsilon, W_3+\alpha^2\varepsilon\right) + F_3\left(W_2-\alpha\varepsilon, W_3+\alpha^2\varepsilon\right)$$

and adding (6) to this equation,

$$\alpha^2\left[F_1(W_1+2\varepsilon, W_2)-F_1(W_1, W_2)\right] + F_3\left(W_2, W_3+2\alpha^2\varepsilon\right) - F_3(W_2, W_3) + 2\alpha\left[2F_2\left(W_1+\varepsilon, W_2-\alpha\varepsilon, W_3+\alpha^2\varepsilon\right)-F_2\left(W_1+2\varepsilon, W_2, W_3+2\alpha^2\varepsilon\right)-F_2(W_1, W_2, W_3)\right] = 0 . \tag{13}$$

F_2 is a linear function of W_1, so $F_1(W_1+2\varepsilon, W_2) - F_1(W_1, W_2)$ is also linear in W_1, from (13). Suppressing W_2 for the time being,

https://doi.org/10.1017/S0004972700040855 Published online by Cambridge University Press

$$[F_1(W_1+3\varepsilon)-F_1(W_1+2\varepsilon)] - [F_1(W_1+2\varepsilon)-F_1(W_1+\varepsilon)]$$
$$= [F_1(W_1+2\varepsilon)-F_1(W_1+\varepsilon)] - [F_1(W_1+\varepsilon)-F_1(W_1)] ,$$

that is,

$$F_1(W_1+3\varepsilon) - 3F_1(W_1+2\varepsilon) + 3F_1(W_1+\varepsilon) - F_1(W_1) = 0 . \tag{14}$$

Given any three values of F_1 , say $F_1(1)$, $F_1(0)$, $F_1(-1)$, then (14) can be used to determine values at all integer points, and the consequent equation,

$$8F_1(W_1+\varepsilon) = 3F_1(W_1+2\varepsilon) + 6F_1(W_1) - F_1(W_1-2\varepsilon) ,$$

all $(m+\frac{1}{2})$ values of F_1 (m any integer) and so on, giving values at any argument of the form $\frac{p}{2^q}$, p, q any integers. The latter points are dense in the continuum, so F_1 is determined by (14), given any three values. So F_1 must be quadratic in W_1 .

Similarly F_3 is a quadratic function of W_3 .

We return now to n dimensions, and the original notation for coefficients of F and W . Whenever a vector x has no zero components, then for any $y \in R^n$,

$$g(x, y) = x^*F(W)x ,$$

where $W_{ii} = \frac{y_i}{x_i}$, $i = 1 \ldots n$, and $W_{ij} = 0$ if $i \neq j$. Then

$$g(x, y) = \sum_{i=1}^{n} \sum_{j=1}^{n} x_i x_j F_{ij}(W) .$$

In this sum, in the cases when $i = j$, F_{ij} is a quadratic function of $W_{ii} = \frac{y_i}{x_i}$ and independent of all other variables, so $x_i^2 F_{ii}(W)$ is a homogeneous quadratic form in x_i, y_i .

And if $i \neq j$, F_{ij} is a function of W_{ii}, W_{ij} only, and is linear

https://doi.org/10.1017/S0004972700040855 Published online by Cambridge University Press

in each taken independently. So again $x_i x_j F_{ij}\left(\frac{y_i}{x_i}, \frac{y_j}{x_j}\right)$ is a homogeneous quadratic form in x_i, x_j, y_i, y_j .

So $g(x, y)$ is a homogeneous quadratic function in x and y, unless some coefficient of x is zero. But $g(x, y)$ is also continuous, (except when $x = 0$) so is a homogeneous quadratic form everywhere. Although it is not defined by (4) when $x = 0$, the domain of definition can be extended to include such points. So

$$g(x, y) = (x^*, y^*)\begin{pmatrix} A & B \\ B^* & C \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}$$

where A, B, C are $n \times n$ matrices, $A = A^*$, $C = C^*$. Therefore

$$\begin{aligned} g(x, Wx) &= x^*Ax + x^*BWx + x^*WB^*x + x^*WCWx \\ &= x^*F(W)x . \end{aligned}$$

So $F(W) = A + BW + WB^* + WCW$.

To get this result, a fixed value of t was used. The coefficients A, B, C will generally be functions of t. They need not be continuous, but the order-preserving property, as stated, will impose some limitations on their behaviour. If $F(t, W)$ is assumed continuous in t, then $A(t)$, $B(t)$, $C(t)$ are continuous also. This can be shown by considering special W values (for example, $A(t) = F(t, 0)$).

3. Remarks

The converse statement, that the Riccati equation has the order-preserving property, is important in the theory of disconjugacy for self-adjoint linear systems and, in control theory, in the theory of the linear regulator problem. It can be established by straight-forward manipulations, as in Reid [5], or by less special argument (Coppel [2]) which can be seen as an application of general arguments about differential inequalities in finite-dimensional spaces where order relationships are specified by a cone of positive vectors, and no particular form of cone is specified (Stokes [7]).

Taking the latter view, the Riccati equation has a special relationship with the cone of vectors in a $\frac{1}{2}n(n+1)$ dimensional space

https://doi.org/10.1017/S0004972700040855 Published online by Cambridge University Press

corresponding to the set of non-negative definite symmetric matrices. Szarski [8] shows a similar relationship between the orthant cone of n-dimensional vectors with non-negative components and a trivial system of n equations each in one variable, with no inter-relations. Circular cones in n-space (leading to a Lorentz-type ordering) are also associated with a system of equations quadratic in the dependent variables (Stokes [7]). Here, preservation of order under a transformation corresponds in Minkowski space-time to preserving the physical property of observability or attainability of one point from another.

The order-preserving property, as we have stated it, is highly restrictive, but it results from the combination of two unidirectional order-preserving properties. The requirement that order relations among solutions of (1) be preserved as t increases, for example, imposes on on $F(t, W)$ a kind of quasi-monotonicity condition, in the sense of Walter [9]. This is less restrictive and is fulfilled, for example, if $F(t, W) = W^3$.

References

[1] Michael Athans, Peter L. Falb, *Optimal control. An introduction to the theory and its applications* (McGraw Hill, New York; Toronto, Ontario; London; 1966).

[2] W.A. Coppel, *Disconjugacy* (Lecture Notes in Mathematics, 220. Springer-Verlag, Berlin, Heidelberg, New York, 1971).

[3] Wyman Fair, "Continued fraction solution to the Riccati equation in a Banach algebra", *J. Math. Anal. Appl.* 39 (1972), 318-323.

[4] R.E. Kalman, "Contributions to the theory of optimal control", *Bol. Soc. Mat. Mexicana* (2) 5 (1960), 102-119.

[5] William T. Reid, *Riccati differential equations* (Mathematics in Science and Engineering, 86. Academic Press, New York, London, 1972).

[6] Alan Schumitzky. "On the equivalence between matrix Riccati equations and Fredholm resolvents", *J. Comput. System Sci.* 2 (1968), 76-87.

[7] A.N. Stokes, "Differential inequalities and the matrix Riccati equation", (PhD Thesis, Australian National University, Canberra, 1972. See also, the abstract, *Bull. Austral. Math. Soc.* 9 (1973), 315-317).

[8] Jacek Szarski, *Differential inequalities* (Monografie Matematyczne, 43. PWN - Polish Scientific Publishers, Warszawa, 1965).

[9] Wolfgang Walter, *Differential- und Integral-Ungleichungen und ihre Anwendung bei Abschätzungs- und Eindeutigkeitsproblemen* (Springer Tracts in Natural Philosophy, Ergebnisse der angewandten Mathematik, 2. Springer-Verlag, Berlin, Göttingen, Heidelberg, New York, 1964).

Division of Environmental Mechanics,
CSIRO,
Canberra, ACT.