A STUDY OF MATRIX EQUATIONS

by

Eileen M. McDonald

A Doctoral Thesis
Submitted in partial fulfilment of the
requirements for the award of Doctor of Philosophy
of the Loughborough University of Technology.

December 1987.

Abstract

Matrix equations have been studied by Mathematicians for many years. Interest in them has grown due to the fact that these equations arise in many different fields such as vibration analysis, optimal control, stability theory etc.

This thesis is concerned with methods of solution of various matrix equations with particular emphasis on quadratic matrix equations. Large scale numerical techniques are not investigated but algebraic aspects of matrix equations are considered.

Many established methods are described and the solution of a matrix equation by consideration of an equivalent system of multivariable polynomial equations is investigated. Matrix equations are also solved by a method which combines the given equation with the characteristic equation of the unknown matrix.

Several iterative processes used for the solution of scalar equations are applied directly to the matrix equation. A new iterative process based on elimination methods is also described and examples given.

The solutions of the equation $X^2 = P$ are obtained by a method which derives a set of polynomial equations connecting the characteristic coefficients of X and P. It is also shown that the equation $X^2 = P$ has an infinite number of solutions if P is a derogatory matrix.

Acknowledgements

My thanks are due to my supervisor Professor C. Storey for his guidance, helpful advice and encouragement at every stage of this work.

I am also very grateful to Mrs. B. Wright for her expert typing of this thesis.

The financial support of Derbyshire County Council is also gratefully acknowledged.

Contents

CHAPTER 1

Introduction

Matrix equations have been studied by mathematicians for many years. Cayley [1858] showed that a square matrix satisfies an algebraic equation of degree equal to its own order and Sylvester [1884] published many new results concerning matrix equations. The topic has been investigated extensively since then by many mathematicians who have built on the foundation of this early work. Interest in matrix equations has grown due to the fact that they arise in many different fields such as vibration analysis, optimal control, stability theory and filtering theory. In these areas, two matrix equations are of particular importance. These are the Matrix Lyapunov equation

$$PA + A^{T}P + Q = 0 \qquad 1.1$$

and the Algebraic Matrix Riccati equation

$$PA + A^{T}P - PBR^{-1}B^{T}P + Q = 0 \quad . \qquad 1.2$$

The numerical solution of these two equations has been studied extensively in recent years.

This thesis is not concerned with the investigation of large scale numerical techniques but concentrates on algebraic aspects of matrix equations. The examples given, to illustrate various methods, have been chosen so that computation can be carried out either by hand or with the assistance of a micro computer.

Numerical methods dominate the techniques for solving matrix equations which have evolved in recent years, but Dennis, Traub &

Weber [1976] are among the minority of authors who have studied the algebra of matrix polynomials and the solution of matrix polynomial equations. [But see also the recent books

GOHBERG, I., P. LANCASTER and L. RODMAN, 'Matrix Polynomials', Academic Press, New York, 1982.

GOHBERG, I., P. LANCASTER and L. RODMAN, 'Invariant subspaces of matrices with Applications', Wiley, Interscience, 1986.]

Some methods, such as iterative processes which are used to solve polynomial equations over the complex field may also be applied directly to the matrix equation to yield matrix solutions. Many methods however cannot be applied directly. This is due to the problems which arise from the fact that the set of matrices form a ring rather than a field. The non commutativity of matrix multiplication and the existence of divisors of zero lead to complications.

Whereas the scalar quadratic equation can always be solved over the complex field by use of the quadratic formula, no such simple method exists in the matrix case. The matrix quadratic equation has several forms such as

$$AX^2 + BX + C = 0$$

$$XAX + BX + XC + D = 0$$

$$X^2A + XB + C = 0 \ .$$

In the scalar case the quadratic equation has at most two solutions. The number of solutions of the matrix quadratic equation depends upon many factors such as the size of the matrices. Ingraham [1941] showed that the unilateral matrix equation may

have an infinite number of solutions and Bell [1950] showed that the unilateral matrix equation has an infinite number of solutions if and only if there exist two distinct solutions X_1 and X_2 which are similar.

A brief description of the work contained in this thesis follows.

In Chapter 2, the solution of matrix equations by consideration of the equivalent system of polynomial equations is studied. Decision methods, multivariable resultants and elimination methods are considered. The constituent equations are solved directly for several matrix equations involving 2×2 matrices and solutions obtained in terms of the trace and determinant of the coefficient matrices.

Chapter 3 is devoted to a study of various established methods of solution with particular emphasis on methods which can be applied to the quadratic matrix equation. The methods are illustrated with examples involving 2×2 and 3×3 matrices.

In Chapter 4 a method is described which makes use of the characteristic equation of the solution matrix. Since the unknown matrix X satisfies simultaneously the given equation and also its own characteristic equation, elimination methods may be used to eliminate powers of X higher than 1 until finally a linear equation in X is obtained, from which the solution may be found. Possibilities for the characteristic equation of the unknown matrix X are obtained by using the fact that if the equation is the unilateral equation $\sum_{o}^{n} A_i X^{n-i} = 0$ then the characteristic polynomial of a solution X is a factor of the scalar polynomial $\det\left[\sum_{o}^{n} A_i \lambda^{n-i}\right]$.

Sylvester [1884] suggested that solutions may be obtained in this way but the development of the technique is new.

Chapter 5 considers the application of iterative methods to the solution of matrix equations. Several iterative processes used for the solution of scalar equations are applied directly to the matrix equation. A new iterative process based on the elimination method of Chapter 4 is also described and examples given.

Chapter 6 deals with the solution of the equation $X^2 = P$, that is the problem of finding the square root of a given matrix P. A new method is described which derives a set of polynomial equations connecting the characteristic coefficients of X and P and obtains the solution matrix X by solving this set of equations. It is also shown that the equation $X^2 = P$ has an infinite number of solutions if P is a derogatory matrix.

As stated earlier, current work on matrix equations focuses on numerical methods and in particular on methods of solution of equations 1.1 and 1.2. [A useful survey of numerical methods for these equations up to 1973/74 can be found in the Report by Hewer and Nazarott, "A survey of numerical methods for the solution of algebraic Riccati equations", Michelson Laboratory, Naval Weapons Centre, 1974.] More recently Davis [1981] has described a method which is based on the Newton iterative method. Solutions of 1.1 and 1.2 have also been obtained by use of the matrix sign function (Popeea and Lupas [1976]) and a recent publication by Charlier and Van Dooren [1987] combines symmetric factorization techniques with the matrix sign algorithm to obtain solutions.

It is possible that the elimination methods described in

Chapters 4 - 6 might be numerically viable and that computerized versions could be used to solve matrix equations for large order matrices. This however is a subject for further research.

CHAPTER 2

The Solution of Matrix Equations by Consideration of the Equivalent System of Polynomial Equations

2.1 INTRODUCTION.

Any matrix equation involving m×m matrices is equivalent to a system of polynomial equations where the unknowns $x_{11}x_{12}x_{13},\ldots,x_{mm}$ are the elements of the required matrix X. Though there are many established methods of solution for systems of polynomial equations, no attempt appears to have been made to apply these methods to the constituent equations arising from a matrix equation. An attempt to do this is made in this chapter.

The constituent equations have a special form and are not completely general polynomial equations, for example, in the case of the matrix equation $X^2 = P$ where $X = \begin{pmatrix} x_1 & x_2 \\ x_3 & x_4 \end{pmatrix}$ $P = \begin{pmatrix} p_1 & p_2 \\ p_3 & p_4 \end{pmatrix}$ the equivalent system of equations is

(1) $x_1^2 + x_2x_3 = p_1$

(2) $x_1x_2 + x_2x_4 = p_2$

(3) $x_1x_3 + x_3x_4 = p_3$

(4) $x_4^2 + x_2x_3 = p_4$.

It can be seen that each individual equation involves only 3 unknowns, that no equation involves x_2^2 or x_3^2 and that both x_2 and x_3 can be expressed as a rational function of x_1 and x_4. Similar

special forms can be found in systems arising from matrix equations involving matrices of higher order. It is possible that the special form of the equations could be used to find solutions.

Iterative methods for the solution of systems of equations can obviously be applied to the constituent equations but the description of these is left until Chapter 5.

Most algebraic methods of solution for systems of polynomial equations involve the systematic reduction of the unknowns one by one until the system is reduced to a single equation in one unknown. The solvability of the final equation then indicates the solvability of the system of polynomial equations. These methods use the fact that the reduction to a single equation can be carried out in a finite number of operations. The finiteness of the number of operations may be illusory from a practical point of view however since the number of operations at each reduction is very large and depends on the number and degree of the polynomials.

In section 2.2 decision methods described by Seidenberg [1954] using ideas developed by Tarski [1951] are considered and in section 2.3 the solution of systems of equations by use of multivariable resultants is described.

In section 2.4 elimination methods are applied and in 2.5 the direct solution of the constituent equations is used for 2×2 matrix equations.

2.2 DECISION METHODS DUE TO TARSKI AND SEIDENBERG.

Given a set of multivariable polynomials $f_i(x_1x_2,\ldots,x_n)$ $i = 1,\ldots,n$ the problem is to decide whether there is a set of real values $(x_1x_2,\ldots,x_n)$ for which $f_i(x_1x_2,\ldots,x_n) = 0$ for $i = 1,\ldots,n$.

Tarski [1951] showed that it is possible to determine, in a finite number of steps, a finite number of systems of equations, inequalities and inequations in the coefficients of the system $f_i(x_1x_2,\ldots,x_n) = 0$ such that the system $f_i = 0$ will have a solution in $\mathbb{R}$ if and only if every equation, inequation and inequality of one of the derived systems is satisfied by the coefficients. This is shown by using a generalized form of Sturm's Theorem. The type of result which may be obtained is illustrated in the case of the following equation in a single variable. Consider the 'reduced' quartic equation

$$x^4 + q\,x^2 + r\,x + s = 0 \quad .$$

Here it can be shown that this equation has a real root if and only if one of the following alternatives is satisfied

I $\quad d \leq 0$

II $\quad d > 0,\ q < 0,\ L > 0$

III $\quad d = 0,\ r \neq 0$

IV $\quad d = 0,\ r = 0,\ q \leq 0$

where $$d = 4\left(4s + \frac{q^2}{3}\right)^3 - 27\left(\frac{8}{3}qs - r^2 - \frac{2q^3}{27}\right)^2$$

and $$L = 8qs - 2q^3 - 9r^2 \quad .$$

This method can be extended to sets of equations in several

variables as follows. The system of equations $f_i(x_1x_2,\ldots,x_n) = 0$ $i = 1,\ldots,n$ can be replaced by a single equation $\Sigma f_i^2(x_1x_2,\ldots,x_n) = 0$. This equation can then be treated as an equation in the variable x_n with coefficients which are polynomials in $(x_1x_2,\ldots,x_{n-1})$. By use of the parameterized version of Sturm's theorem, the conditions for this equation to have a real root can be found and this set of conditions will be a set of polynomial equations, inequations and inequalities which are polynomials in $(x_1x_2,\ldots,x_{n-1})$. This set can then be replaced by a single equation since any set of equations $g_i(x_1x_2,\ldots,x_{n-1}) = 0$ can be replaced by a single equation $\Sigma g_i^2 = 0$ and a finite set of inequations $g_i \neq 0$ can be replaced by a single one $\pi g_i \neq 0$ and an inequation $g \neq 0$ is equivalent to $g^2 > 0$. Moreover the sign > 0 can also be eliminated as the condition $h > 0$ is equivalent to the condition $z^2h = 1$ for some z.

Using these reductions a single equation in n-1 variables can be obtained and the process repeated again by considering this as an equation in x_{n-1} with coefficients which are polynomials in $(x_1x_2,\ldots,x_{n-2})$. Hence the variables are eliminated one by one until a single equation in x_1 is obtained, the solvability of which can be decided by use of Sturm's theorem.

Seidenberg[1954] arrives at the same result - that the solvability of a system of equations can be decided in a finite number of steps, but he does not use Sturm's theorem to eliminate each variable. Having replaced the system of equations by a single equation by setting $\Sigma f_i^2 = 0$ Seidenberg uses the fact that if there is a vector $(a_1a_2,\ldots,a_n)$ which is a solution of $f(x_1x_2,\ldots,x_n) = 0$ then there is a vector of smallest absolute value. The problem is

therefore to minimize $\sum_{1}^{n} x_i^2$ subject to the constraint that $f(x_1,x_2,\ldots,x_n) = 0$. This can be done using Lagrange multipliers by setting $\omega = \sum_{1}^{n} x_i^2 - \lambda f(x_1x_2,\ldots,x_n)$.

Then the minimum value of ω corresponds to the minimum value of $\sum_{1}^{n} x_i^2$. Hence we set $\frac{\partial \omega}{\partial x_i} = 0$, $i = 1,2,\ldots,n$ and by eliminating λ we obtain a second equation $g(x_1x_2,\ldots,x_n) = 0$. The variable x_n is then eliminated by forming the resultant with respect to x_n of the polynomials $f(x_1x_2,\ldots,x_n)$ and $g(x_1x_2,\ldots,x_n)$. If $h(x_1x_2,\ldots,x_{n-1}) = \mathrm{Res}(f,g)$ then there is a common solution of $f(x_1,\ldots,x_n) = 0$ and $g(x_1,\ldots,x_n) = 0$ if and only if $h(x_1x_2,\ldots,x_{n-1}) = 0$.

We have therefore obtained an equation in (n-1) variables. The process is repeated and the variables eliminated one by one until an equation in the single variable x_1 is obtained. The solvability of the final equation can then be decided by Sturm's Theorem and hence the solvability of the original system of equations can be decided.

Though this method, in theory, gives a means of deciding whether a system of equations has a solution the problems of computation which arise in carrying out the steps are very great. Seidenberg states that neither he nor Tarski have computed numbers of steps involved and wonders whether a decision machine could be constructed to carry out the process.

A simple illustration of the method is shown in the following example.

Example 2.2.1.

Consider the equation $AX = B$

where $A = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$ $X = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ $B = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$.

The matrix equation can be written as a pair of equations

$$f_1 : x_1 + x_2 - 1 = 0$$

$$f_2 : x_1 - x_2 - 2 = 0 \quad .$$

The pair of equations can be written as a single equation

$$F(x_1x_2) = (x_1 + x_2 - 1)^2 + (x_1 - x_2 - 2)^2 = 0$$

or $\quad 2x_1^2 + 2x_2^2 - 6x_1 + 2x_2 + 5 = 0 \quad .$

We now seek to minimize $x_1^2 + x_2^2$ with the constraint that $F(x_1x_2) = 0$.

Using Lagrange multipliers we obtain a second equation

$$G(x_1x_2) = 6x_2 + 2x_1 = 0 \ .$$

Let $H(x_2)$ be the resultant with respect to x_1 of F and G

then $\quad H(x_2) = \begin{vmatrix} 2 & -6 & 2x_2^2 + 2x_2 + 5 \\ 2 & 6x_2 & 0 \\ 0 & 2 & 6x_2 \end{vmatrix}$

$$= 80x_2^2 + 80x_2 + 20 \ .$$

F and G have a common solution if $H(x_2) = 0$

$\therefore \quad 20[4x_2^2 + 4x_2 + 1] = 0$

$\therefore \quad x_2 = -\frac{1}{2}$ and substituting this in F or G gives

$$x_1 = \frac{3}{2} \quad .$$

Hence the matrix solution is $X = \begin{pmatrix} \frac{3}{2} \\ -\frac{1}{2} \end{pmatrix}$

Though this method may be applied in particular numerical examples the problems of computation increase rapidly as the number of equations and number of variables increase. When general coefficient matrices are considered the equations quickly become unwieldy. Even in the simple example

$$AX = B \quad \text{where} \quad A = \begin{pmatrix} a_1 & a_2 \\ a_3 & a_4 \end{pmatrix} \quad B = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$

it is difficult to extract the condition $a_1a_4 - a_2a_3 \neq 0$ as can be seen as follows:

Example 2.2.2.

Consider $AX = B$ where $A = \begin{pmatrix} a_1 & a_2 \\ a_3 & a_4 \end{pmatrix}$ $X = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ $B = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$.

The equivalent system of polynomial equations is

$$f_1 : a_1x_1 + a_2x_2 - b_1 = 0$$

$$f_2 : a_3x_1 + a_4x_2 - b_2 = 0 \quad .$$

13

Let $F(x_1x_2) = f_1^2 + f_2^2$ and let $G(x_1x_2)$ be the polynomial obtained by seeking to minimize $x_1^2 + x_2^2$ with the constraint that $F(x_1x_2) = 0$.

Then

$$F(x_1x_2) = c_1x_2^2 + (c_2x_1 + c_3)x_2 + c_4x_1^2 + c_5x_1 + c_6$$

$$G(x_1x_2) = c_2x_2^2 + 2(c_4 - c_1)x_1x_2 + c_5x_2 + c_2x_1^2 - c_3x_1$$

where

$$c_1 = a_2^2 + a_4^2 \qquad c_2 = 2(a_1a_2 + a_3a_4) \qquad c_3 = -2(a_2b_1 + a_4b_2)$$

$$c_4 = a_1^2 + a_3^2 \qquad c_5 = -2(a_1b_1 + a_3b_2) \qquad c_6 = b_1^2 + b_2^2 \ .$$

Let $H(x_1)$ be the resultant of F and G with respect to x_2.

Then

$$H(x_1) = \begin{vmatrix} c_1 & c_2x_1+c_3 & c_4x_1^2+c_5x_1+c_6 & 0 \\ 0 & c_1 & c_2x_1+c_3 & c_4x_1^2+c_5x_1+c_6 \\ c_2 & 2(c_4-c_1)x_1+c_5 & c_2x_1^2-c_3x_1 & 0 \\ 0 & c_2 & 2(c_4-c_1)x_1+c_5 & c_2x_1^2-c_3x_1 \end{vmatrix}$$

therefore

$$H(x_1) = \left[3c_1^2c_2^2+4c_1c_4^3+4c_1^3c_4-8c_1^2c_4^2-c_2^2c_4^2-2c_1c_2^2c_4\right]x_1^4$$

$$+ \left[-2c_1^2c_2c_3-12c_1^2c_4c_5+4c_1c_2c_3c_4+4c_1^3c_5-2c_2c_3c_4^2-c_2^2c_4c_5-c_1c_2^2c_5\right]x_1^3$$

$$+ \left[4c_1c_2c_3c_5+2c_1c_3^2c_4-c_1^2c_3^2+5c_1c_4c_5^2-4c_1^2c_5^2-c_2^2c_3^2-3c_2c_3c_4c_5+c_1c_2^2c_6\right]x_1^2$$

$$+ \left[c_1c_3^2c_5+c_1c_5^3+4c_1c_4c_5c_6-4c_1^2c_5c_6-c_2c_3^3-c_2c_3c_5^2-2c_2c_3c_4c_6+4c_1c_2c_3c_6 \right.$$

$$\left. +c_2^2c_5c_6\right]x_1$$

$$+\left[c_5^2c_6 - c_2c_3c_5c_6 + c_2^2c_6^2\right] .$$

It is known from the simplicity of the original equations that there will be a unique solution for x_1 provided that $a_1a_4 - a_2a_3 \neq 0$ and hence the coefficients of x_1^4, x_1^3, x_1^2 will be zero. This condition for a unique solution is difficult to obtain algebraically however due to the complexity of the coefficients.

It may be concluded that it is possible to decide in a finite number of steps the solvability of a system of equations obtained from a matrix equation. However it is difficult to derive general properties concerning the coefficient matrices by consideration of the constituent equations. Even with small order matrices the problems of algebraic manipulation soon become apparent.

2.3 SOLUTION OF EQUATIONS BY MULTIVARIABLE RESULTANTS.

As shown in the previous section, a common solution of a pair of equations in two variables $f(x_1x_2) = 0$ and $g(x_1x_2) = 0$ may be found by forming the resultant of f and g with respect to x_2. This gives a polynomial in a single variable $h(x_1)$ and any solution of $h(x_1) = 0$ leads to a solution (x_1x_2) of $f(x_1x_2) = 0$ and $g(x_1x_2) = 0$.

This section extends the idea to n equations in n variables by use of multivariable resultants. The method is described by Hodge and Pedoe [1947]. Again, the general idea is to eliminate one variable at a time until ultimately only one equation is left in a single variable.

Consider first a set of r equations in a single variable

$$f_i(x) = 0 \quad i = 1,2,\ldots,r \quad .$$

The set may be reduced to two equations by introducing additional indeterminates $u_1u_2,\ldots,u_r$ and $v_1v_2,\ldots,v_r$.

Let

$$\phi(x) \equiv u_1f_1 + u_2f_2 + \ldots u_rf_r$$

$$\psi(x) \equiv v_1f_1 + v_2f_2 + \ldots v_rf_r \quad .$$

Forming the resultant of $\phi(x)$ and $\psi(x)$ we obtain R(u v) which is a polynomial in $u_1u_2,\ldots,u_r\ v_1v_2,\ldots,v_r$. Let the coefficient of $u_1^{i_1}\ u_2^{i_2},\ldots,u_r^{i_r}\ v_1^{j_1}\ v_2^{j_2},\ldots,v_r^{j_r}$ be D_s. Then D_s is a polynomial in the coefficients of the original polynomials $f_i(x)$.

These polynomials $D_1D_2,\ldots,D_N$ constitute a resultant system of the set of equations $f_i(x) = 0$.

If the set of equations $f_i(x) = 0$ has a solution say $x = a$ then $f_i(a) = 0$, $i = 1,2,\ldots,r$, and hence $\phi(a) = 0$ and $\psi(a) = 0$ and therefore the resultant R(u v) = 0. This implies that every coefficient D_s vanishes and gives a set of conditions on the coefficients of the polynomials $f_i(x)$ for a solution to exist. Conversely, if each coefficient D_s vanishes then this means that $\phi(x)$ and $\psi(x)$ have a common factor. However, since $\phi(x)$ contains no v_j and $\psi(x)$ contains no u_j then the common factor is independent of the u_iv_j and must therefore be a common factor of the equations $f_i(x) = 0$. Hence the system has a solution.

This technique may be extended to a set of equations in several unknowns. Consider the set

$$g_i(x_1x_2,\ldots,x_n) = 0 \qquad i = 1,2,\ldots,r \quad .$$

Let the polynomial $g_i(x_1,\ldots,x_n)$ be of degree m_i when regarded as a polynomial in x_n with coefficients in $R[x_1,\ldots,x_{n-1}]$ then each polynomial can be written as

$$g_i(x_1x_2,\ldots,x_n) \quad a_i(x_1,\ldots,x_{n-1})x_n^{m_i} + b_i(x_1,\ldots,x_{n-1})x_n^{m_i-1} + \ldots$$

and the techniques previously described may be applied to this set.

The resultant system D_s will be a set of equations in one less variable $(x_1x_2,\ldots,x_{n-1})$. The process may be repeated until all the variables have been eliminated except x_1.

The method is illustrated in the solution of a matrix equation as follows.

Example 2.3.1.

Consider the equation $AX + XB = C$

where $A = \begin{pmatrix} 1 & 2 \\ -1 & 3 \end{pmatrix}$ $B = \begin{pmatrix} 2 & 1 \\ 1 & -1 \end{pmatrix}$ $C = \begin{pmatrix} 3 & 3 \\ -5 & -1 \end{pmatrix}$.

The matrix equation is equivalent to the set of equations

$$f_1 : 3x_1 + x_2 + 2x_3 - 3 = 0$$

$$f_2 : x_1 + 2x_4 - 3 = 0$$

$$f_3 : -x_1 + 5x_3 + x_4 + 5 = 0$$

$$f_4 : -x_2 + x_3 + 2x_4 + 1 = 0 \quad .$$

Introducing the new indeterminates $u_1,u_2,u_3,u_4,\ v_1,v_2,v_3,v_4$.

Let $\Phi \equiv u_1f_1 + u_2f_2 + u_3f_3 + u_4f_4$

$$\psi \equiv v_1f_1 + v_2f_2 + v_3f_3 + v_4f_4$$

Then

$$\Phi(x_1x_2x_3x_4) \equiv (3u_1+u_2-u_3)x_1+(u_1-u_4)x_2+(2u_1+5u_3+u_4)x_3+(2u_2+u_3+2u_4)x_4$$
$$-3u_1-3u_2+5u_3+u_4$$

$$\psi(x_1x_2x_3x_4) \equiv (3v_1+v_2-v_3)x_1+(v_1-v_4)x_2+(2v_1+5v_3+v_4)x_3+(2v_2+v_3+2v_4)x_4$$
$$-3v_1-3v_2+5v_3+v_4 .$$

The resultant R_1 of Φ and ψ with respect to x_4 is formed

$$R_1 = \left[-6x_1-2x_2-4x_3+6\right]u_1v_2+\left[-3x_1-x_2-2x_3+3\right]u_1v_3+\left[-6x_1-2x_2-4x_3+6\right]u_1v_4$$
$$+ \left[6x_1+2x_2+4x_3-6\right]u_2v_1+\left[-3x_1+10x_3+13\right]u_2v_3+\left[-2x_1-2x_2+2x_3+8\right]u_2v_4$$
$$+ \left[3x_1+x_2+2x_3-3\right]u_3v_1+\left[3x_1-10x_3-13\right]u_3v_2+\left[2x_1-x_2-9x_3-9\right]u_3v_4$$
$$+ \left[6x_1+2x_2+4x_3-6\right]u_4v_1+\left[2x_1+2x_2-2x_3-8\right]u_4v_2+\left[-2x_1+x_2+9x_3+9\right]u_4v_3$$

The coefficients of the u_iv_j are all polynomials in the three variables $x_1x_2x_3$.

Setting the coefficients of u_iv_j to zero gives a resultant system of the set of equations $f_i(x_1x_2x_3x_4) = 0$.

The resultant system is

$$g_1(x_1x_2x_3) : 3x_1 + x_2 + 2x_3 - 3 = 0$$

$$g_2(x_1x_2x_3) : 3x_1 - 10x_3 - 13 = 0$$

$$g_3(x_1x_2x_3) : x_1 + x_2 - x_3 - 4 = 0 .$$

This set of equations is the same as the set which is obtained when the variable x_4 is eliminated directly from the original equations by taking the equations in pairs.

The process is now repeated by setting

$$\Phi \equiv u_1 g_1 + u_2 g_2 + u_3 g_3$$

$$\psi \equiv v_1 g_1 + v_2 g_2 + v_3 g_3$$

then

$$\Phi(x_1 x_2 x_3) \equiv (3u_1+3u_2+u_3)x_1 + (u_1+u_3)x_2 + (2u_1-10u_2-u_3)x_3 - 3u_1 - 13u_2 - 4u_3$$

$$\psi(x_1 x_2 x_3) \equiv (3v_1+3v_2+v_3)x_1 + (v_1+v_3) + (2v_1-10v_2-v_3)x_3 - 3v_1 - 13v_2 - 4v_3$$

the resultant R_2 of Φ and ψ with respect to x_3 is formed

$$R_2 = \left[36x_1+10x_2-56\right]u_1v_2 + \left[5x_1+3x_2-11\right]u_1v_3 + \left[-36x_1-10x_2+56\right]u_2v_1 + \left[-7x_1-10x_2+2\right]u_2v_3 + \left[-5x_1-3x_2+11\right]u_3v_1 + \left[7x_1+10x_2-27\right]u_3v_2 \ .$$

Setting the coefficients of the $u_i v_j$ to zero gives the resultant system

$$h_1(x_1 x_2) : 5x_1 + 3x_2 - 11 = 0$$

$$h_2(x_1 x_2) : 7x_1 + 10x_2 - 27 = 0 \ .$$

The resultant R_3 of h_1 and h_2 with respect to x_2 is formed

$$R_3(x_1) = 29x_1 - 29$$

therefore the $h_i(x_1x_2) = 0$ have a common solution if and only if $R_3(x_1) = 0$.

Hence $x_1 = 1$

Substitution in h_1 gives $x_2 = 2$

Substitution in g_1 gives $x_3 = -1$

Substitution in f_2 gives $x_4 = 1$.

Hence the matrix solution is $X = \begin{pmatrix} 1 & 2 \\ -1 & 1 \end{pmatrix}$.

This method could in theory be extended to a matrix equation of any degree involving matrices of any order. Without computer assistance however, problems of computation quickly arise as can be seen in the following case.

Example 2.3.2.

Consider the matrix equation $XAX - B = 0$

where $A = \begin{pmatrix} 2 & 1 \\ -1 & 2 \end{pmatrix}$ $B = \begin{pmatrix} -5 & 0 \\ -1 & 4 \end{pmatrix}$

The matrix equation is equivalent to the four polynomial equations

$$f_1 : 2x_1^2 - x_1x_2 + x_1x_3 + 2x_2x_3 + 5 = 0$$

$$f_2 : 2x_1x_2 - x_2^2 + x_1x_4 + 2x_2x_4 = 0$$

$$f_3 : 2x_1x_3 - x_1x_4 + 2x_3x_4 + x_3^2 + 1 = 0$$

$$f_4 : 2x_2x_3 - x_2x_4 + x_3x_4 + 2x_4^2 + 4 = 0 .$$

Let

$$\Phi \equiv u_1 f_1 + u_2 f_2 + u_3 f_3 + u_3 f_4$$

$$\psi \equiv v_1 f_1 + v_2 f_2 + v_3 f_3 + v_4 f_4$$

then

$$\Phi \equiv 2u_1 x_1^2 - u_2 x_2^2 + u_3 x_3^2 + 2u_4 x_4^2 + \left[-u_1 - 2u_2\right] x_1 x_2 + \left[u_1 + 2u_3\right] x_1 x_3$$

$$+ \left[u_2 - u_3\right] x_1 x_4 + \left[2u_1 + 2u_4\right] x_2 x_3 + \left[2u_2 - u_4\right] x_2 x_4$$

$$+ \left[2u_3 + u_4\right] x_3 x_4 + 5u_1 + u_3 + 4u_4$$

$$\psi \equiv 2v_1 x_1^2 - v_2 x_2^2 + v_3 x_3^2 + 2v_4 x_4^2 + \left[-v_1 - 2v_2\right] x_1 x_2 + \left[v_1 + 2v_3\right] x_1 x_3$$

$$+ \left[v_2 - v_3\right] x_1 x_4 + \left[2v_1 + 2v_4\right] x_2 x_3 + \left[2v_2 - v_4\right] x_2 x_4$$

$$+ \left[2v_3 + v_4\right] x_3 x_4 + 5v_1 + v_3 + 4v_4 .$$

The resultant of Φ and ψ with respect to x_4 is

$$R = 4u_4^2\beta_2^2 - 2u_4\alpha_1\beta_1\beta_2 + 2u_4\alpha_2\beta_1^2 - 8u_4 v_4\alpha_2\beta_2 + 2v_4\alpha_1^2\beta_2$$

$$- 2v_4\alpha_1\alpha_2\beta_1 + 4\alpha_2^2 v_4^2$$

where

$$\alpha_1 = (u_2 - u_3)x_1 + (2u_2 - u_4)x_2 + (2u_3 + u_4)x_3$$

$$\alpha_2 = 2u_1 x_1^2 - u_2 x_2^2 + u_3 x_3^2 + (-u_1 - 2u_2)x_1 x_2 + (u_1 + 2u_3)x_1 x_3$$

$$+ (2u_1 + 2u_4)x_2 x_3 + 5u_1 + u_3 + 4u_4$$

21

$$\beta_1 = (v_2-v_3)x_1 + (2v_2-v_4)x_2 + (2v_3+v_4)x_3$$

$$\beta_2 = 2v_1x_1^2 - v_2x_2^2 + v_3x_3^2 + (-v_1-2v_2)x_1x_2 + (v_1+2v_3)x_1x_3$$

$$+ (2v_1+2v_4)x_2x_3 + 5v_1 + v_3 + 4v_4 \quad .$$

Though this resultant could be evaluated, it would be difficult to achieve without computer assistance since there are 100 terms in u_iv_j. The problems of computation would obviously be greatly increased if the order of the matrices involved was greater than 2×2.

2.4 ELIMINATION METHOD.

A method of elimination which was known to algebraists of the last century has been adapted for the solution of systems of equations by computer. It is described by Williams [1962] and reduces the problem to that of finding the zeros of a polynomial in a single variable. Given a set of equations

$$F_1(x_1x_2,\ldots,x_n) = 0$$

$$F_2(x_1x_2,\ldots,x_n) = 0$$

$$\vdots$$

$$F_n(x_1x_2,\ldots,x_n) = 0$$

the method derives a set of n equations of the form

$$f_1(x_1x_2,\ldots,x_n) = 0$$

$$f_2(x_2x_3,\ldots,x_n) = 0$$

$$f_3(x_3x_4,\ldots,x_n) = 0$$

$$\vdots$$

$$f_n(x_n) = 0$$

and the set of solutions of $F_i(x_1,\ldots,x_n) = 0, \quad i = 1,\ldots,n$ is identical to the set of solutions of $f_i(x_1,\ldots,x_n) = 0$, $i = 1,\ldots,n$. When x_n is obtained from the final equation, the values of $(x_1x_2,\ldots,x_{n-1})$ may be found by back substitution.

The method is illustrated in the following simple example.

Given the two equations

$$A \ldots\ldots\ldots x_1^2 - x_2^2 = 0$$

$$B \ldots\ldots\ldots x_1^2 - x_1 + 2x_2^2 - x_2 - 1 = 0$$

we need to eliminate x_1 and hence obtain an equation in x_2 only.

Let $C = A - B$

$$C \ldots\ldots\ldots\ldots x_1 - 3x_2^2 + x_2 + 1 = 0 .$$

Let $D = B - x_1C$

$$D \ldots\ldots\ldots\ldots \left[3x_2^2-x_2-2\right]x_1 + 2x_2^2 - x_2 - 1 = 0 .$$

Let $E = \left[3x_2^2-x_2-2\right]C - D$

$$E \ldots\ldots\ldots\ldots 9x_2^4 - 6x_2^3 - 6x_2^2 + 2x_2 + 1 = 0 .$$

The reduced system of equations of the required form is therefore

$$C \ldots\ldots\ldots\ldots x_1 - 3x_2^2 + x_2 + 1 = 0$$

$$E \ldots\ldots\ldots\ldots 9x_2^4 - 6x_2^3 - 6x_2^2 + 2x_2 + 1 = 0 .$$

Each solution of E gives a corresponding value of x_1 when substituted in C and the solutions obtained are also solutions of the original equations A & B.

This method may be applied to the constituent equations obtained from a matrix equation as shown in the following examples

Example 2.4.1.

Consider the equation XAX = B where $A = \begin{pmatrix} 1 & -1 \\ 2 & 0 \end{pmatrix}$ $B = \begin{pmatrix} 10 & -2 \\ 9 & 8 \end{pmatrix}$

The matrix equation is equivalent to the four equations

$$f_1 : x_1^2 + 2x_1x_2 - x_1x_3 - 10 = 0$$

$$f_2 : x_1x_2 + 2x_2^2 - x_1x_4 + 2 = 0$$

$$f_3 : x_1x_3 + 2x_1x_4 - x_3^2 - 9 = 0$$

$$f_4 : x_2x_3 + 2x_2x_4 - x_3x_4 - 8 = 0$$

Step 1. Eliminate x_4 between f_2 and f_3 by letting $f_5 = 2x_1f_2 + x_1f_3$

$$\therefore \quad f_5 : 2x_1^2x_2 + 4x_1x_2^2 - 5x_1 + x_1^2x_3 - x_1x_3^2 = 0 \ .$$

Eliminate x_4 between f_3 and f_4 by letting $f_6 = (2x_2-x_3)f_3 - 2x_1f_4$

$$\therefore \quad f_6 : x_3^3 - x_1x_3^2 - 2x_2x_3^2 - 18x_2 + 9x_3 + 16x_1 = 0 \ .$$

After Step 1 we have 3 equations in 3 unknowns

$$f_1 : x_1^2 + 2x_1x_2 - x_1x_3 - 10 = 0$$

$$f_5 : 2x_1^2x_2 + 4x_1x_2^2 - 5x_1 + x_1^2x_3 - x_1x_3^2 = 0$$

$$f_6 : x_3^3 - x_1x_3^2 - 2x_2x_3^2 - 18x_2 + 9x_3 + 16x_1 = 0 \ .$$

Step 2. Eliminate x_2^2 between f_1 and f_5 by letting $f_7 = 2x_2f_1 - f_5$

$$f_7 : -2(x_1x_3+10)x_2 + 5x_1^2 - x_1^2x_3 + x_1x_3^2 = 0 \ .$$

Eliminate x_2 between f_1 and f_7 by letting $f_8 = (x_1x_3+10)f_1 + x_1f_7$

$$\therefore \qquad f_8 : 3x_1^2 - 4x_1x_3 - 20 = 0 \quad .$$

Eliminate x_2 between f_1 and f_6 by letting $f_9 = (x_3^2+9)f_1 + x_1f_6$

$$\therefore \qquad f_9 : 5x_1^2 - 2x_3^2 - 18 = 0 \quad .$$

After Step 2 we have 2 equations in 2 unknowns

$$f_8 : 3x_1^2 - 4x_1x_3 - 20 = 0$$

$$f_9 : 5x_1^2 - 2x_3^2 - 18 = 0 \quad .$$

Step 3. Eliminate x_3 between f_8 and f_9 .

Let $\qquad f_{10} = x_3f_8 - 2x_1f_9$

$$\therefore \qquad f_{10} : (3x_1^2-20)x_3 - 10x_1^3 + 36x_1 = 0 \quad .$$

Let $\qquad f_{11} = (3x^2-20)f_8 + 4x_1f_{10}$

$$\therefore \qquad f_{11} : 31x^4 - 24x_1^2 - 400 = 0 \quad .$$

We now have 4 equations of the required type

$$f_{11}(x_1) = 31x_1^4 - 24x_1^2 - 400 = 0$$

$$f_8(x_1x_3) = 3x_1^2 - 4x_1x_3 - 20 = 0$$

$$f_1(x_1x_2x_3) = x_1^2 + 2x_1x_2 - x_1x_3 - 10 = 0$$

$$f_2(x_1x_2x_4) = x_1x_2 + 2x_2^2 - x_1x_4 + 2 = 0 \quad .$$

From equation $f_{11}(x_1) = 0$ we have $x_1^2 = 4$ or $x_1^2 = -\frac{100}{31}$

$\therefore$ the two real solutions for x_1 are $x_1 = 2$ or $x_1 = -2$.

If $x_1 = 2$ back substitution gives $x_3 = -1$ $x_2 = 1$ $x_4 = 3$

If $x_1 = -2$ back substitution gives $x_3 = -3$ $x_2 = -1$ $x_4 = -3$

$\therefore$ the two real solutions of the matrix equation are

$$X_1 = \begin{pmatrix} 2 & 1 \\ -1 & 3 \end{pmatrix} \qquad X_2 = \begin{pmatrix} -2 & -1 \\ 1 & -3 \end{pmatrix} .$$

Example 2.4.2.

Consider the matrix equation $X^2 = P$ where $P = \begin{pmatrix} -1 & -2 \\ 4 & -1 \end{pmatrix}$

the constituent equations are

$$f_1 : x_1^2 + x_2x_3 + 1 = 0$$

$$f_2 : x_1x_2 + x_2x_4 + 2 = 0$$

$$f_3 : x_1x_3 + x_3x_4 - 4 = 0$$

$$f_4 : x_2x_3 + x_4^2 + 1 = 0 \quad .$$

Eliminating x_4 gives a set of 3 equations in 3 unknowns

$$f_1 : x_1^2 + x_2x_3 + 1 = 0$$

$$f_5 : x_3 + 2x_2 = 0$$

$$f_7 : x_1^2x_3^2 - 8x_1x_3 + x_2x_3^3 + x_3^2 + 16 = 0 \quad .$$

Eliminating x_2 gives 2 equations in 2 unknowns

$$f_8 : 2x_1^2 - x_3^2 + 2 = 0$$

$$f_9 : 2x_1^2x_3^2 - 16x_1x_3 + 2x_3^2 - x_3^4 + 32 = 0 \quad .$$

Eliminating x_1^2 gives

$$f_{10} : x_1x_3 - 2 = 0 \quad .$$

Finally eliminating x_1 we obtain an equation in a single variable and now have four equations of the required type

$$f_{12}(x_3) : x_3^4 - 2x_3^2 - 8 = 0$$

$$f_{10}(x_1x_3) : x_1x_3 - 2 = 0$$

$$f_5(x_2x_3) : x_3 + 2x_2 = 0$$

$$f_2(x_1x_2x_4) : x_1x_2 + x_2x_4 + 2 = 0 \quad .$$

From $f_{12}(x_3) = 0$ we obtain two real solutions $x_3 = 2$ or $x_3 = -2$.

If $x_3 = 2$ back substitution gives $x_1 = 1 \quad x_2 = -1 \quad x_4 = 1$

If $x_3 = -2$ back substitution gives $x_1 = -1 \quad x_2 = 1 \quad x_4 = -1$.

Hence the two matrix solutions are $X = \begin{pmatrix} 1 & -1 \\ 2 & 1 \end{pmatrix} \quad X = \begin{pmatrix} -1 & 1 \\ -2 & -1 \end{pmatrix}$

Williams[1962] described how the complete procedure could be programmed for a digital computer using a polynomial manipulation system. The experience gained from programs using the procedure has been analysed by Moses [1966]. The conclusions

are that the method gives fast and accurate results for small systems of equations. For large systems of equations the final polynomial in a single variable can be of such a high degree that it is virtually impossible to solve for all the roots. Since a matrix equation involving m×m matrices is equivalent to m^2 polynomial equations this method would only give accurate results for matrices of low order.

2.5 DIRECT SOLUTION.

In the case of matrix equations involving 2×2 matrices it is possible in some cases to obtain a solution by solving the constituent equations directly making use of the fact that they are not completely general polynomials but have a special form. It is possible then that any results obtained in the 2×2 case could be extended to matrices of larger order. Consider the equation $X^2 = A$ where $A = \begin{pmatrix} a_1 & a_2 \\ a_3 & a_4 \end{pmatrix}$

the equivalent system of equations is

(1) $x_1^2 + x_2x_3 = a_1$

(2) $x_2(x_1+x_4) = a_2$

(3) $x_3(x_1+x_4) = a_3$

(4) $x_4^2 + x_2x_3 = a_4$.

From equations (2) and (3) $x_3 = \frac{a_3x_2}{a_2}$, when $a_2 \neq 0$.

Substituting in (1) and (4) gives $x_1 = \sqrt{a_1 - \frac{a_3x_2^2}{a_2}}$. $x_4 = \sqrt{a_4 - \frac{a_3x_2^2}{a_2}}$

and since $x_1 + x_4 = \frac{a_2}{x_2}$

then $$\sqrt{a_1 - \frac{a_3 x_2^2}{a_2}} + \sqrt{a_4 - \frac{a_3 x_2^2}{a_2}} = \frac{a_2}{x_2} \quad .$$

Simplifying gives a quadratic in x_2^2

$$\left\{4\left[a_1 a_4 - a_2 a_3\right] - \left[a_1 + a_4\right]^2\right\} x_2^4 + 2a_2^2(a_1+a_4)x_2^2 - a_2^4 = 0$$

$$\therefore \quad x_2^2 = \frac{a_2^2}{\text{Tr } A + 2\sqrt{|A|}} \quad \text{or} \quad x_2^2 = \frac{a_2^2}{\text{Tr } A - 2\sqrt{|A|}}$$

where Tr A = Trace A

$|A|$ = det A

and neither $\text{Tr } A \pm 2\sqrt{|A|}$ is zero.

Having obtained x_2 then x_3 may be obtained from

$$x_3 = \frac{a_3 x_2}{a_2}$$

to obtain the corresponding values of x_1 and x_4.

Subtracting Equation (4) from Equation (1) gives

$$x_1^2 - x_4^2 = a_1 - a_4$$

and hence $$(x_1 - x_4)(x_1 + x_4) = a_1 - a_4 \quad .$$

But $$x_1 + x_4 = \frac{a_2}{x_2} \qquad \therefore \quad x_1 - x_4 = \frac{x_2}{a_2}(a_1 - a_4) \quad .$$

Hence $$x_1 = \frac{1}{2}\left[\frac{a_2}{x_2} + x_2 \frac{(a_1 - a_4)}{a_2}\right]$$

and $$x_4 = \frac{1}{2}\left[\frac{a_2}{x_2} - x_2 \frac{(a_1 - a_4)}{a_2}\right].$$

Since there may be 4 possible real values for x_2 and x_3 x_1 x_4 may be obtained in terms of x_2, then there are four possible matrix solutions.

Example 2.5.1.

Consider the equation $X^2 = A$ where $A = \begin{pmatrix} -1 & 8 \\ -4 & 7 \end{pmatrix}$

then Trace A = 6 and det A = 25

$$\therefore \quad x_2^2 = \frac{64}{6+10} \quad \text{or} \quad x_2^2 = \frac{64}{6-10}$$

$$\therefore \quad x_2 = 2, -2, 4i, -4i \quad .$$

If $x_2 = 2$ then $x_3 = \frac{-4\times 2}{8} = -1$

$$x_1 = \frac{1}{2}\left[\frac{8}{2} + \frac{2(-8)}{8}\right] = 1$$

$$x_4 = \frac{1}{2}\left[\frac{8}{2} - \frac{2(-8)}{8}\right] = 3 \quad .$$

Hence the matrix solution corresponding to this value of x_2 is

$$X = \begin{pmatrix} 1 & 2 \\ -1 & 3 \end{pmatrix} .$$

If $x_2 = -2$ then $x_3 = 1 \quad x_1 = -1 \quad x_4 = -3$

$\therefore$ the other real matrix solution is $X = \begin{pmatrix} -1 & -2 \\ 1 & -3 \end{pmatrix}$

The constituent equations may be used to consider special cases, for example, the equation $X^2 = A$ where A is singular. Putting det A = 0 in the formula obtained for x_2 gives

$$x_2 = \frac{a_2}{\pm\sqrt{a_1+a_4}} \quad \text{and hence} \quad x_3 = \frac{a_3}{\pm\sqrt{a_1+a_4}} \quad .$$

Similarly placing the value for x_2 in the formulae for x_1 and x_4 gives

$$x_1 = \frac{a_1}{\pm\sqrt{a_1+a_4}} \quad \text{and} \quad x_4 = \frac{a_4}{\pm\sqrt{a_1+a_4}} \quad .$$

Hence if A is singular the equation $X^2 = A$, where A is a 2×2 matrix, has only two solutions which are

$$X = \frac{1}{\sqrt{a_1+a_4}} \begin{pmatrix} a_1 & a_2 \\ a_3 & a_4 \end{pmatrix} \quad \text{and} \quad X = \frac{1}{\sqrt{a_1+a_4}} \begin{pmatrix} -a_1 & -a_2 \\ -a_3 & -a_4 \end{pmatrix}.$$

Another special case to consider is the equation $X^2 = A$ where A is a derogatory matrix, i.e. its minimum polynomial is of lower degree than its characteristic polynomial.

A 2×2 derogatory matrix is of the form $\begin{pmatrix} \alpha & 0 \\ 0 & \alpha \end{pmatrix}$

Hence $a_2 = a_3 = 0$, Trace A = 2α and Det A = α^2.

Putting these values in the formula obtained for x_2 gives

$$x_2^2 = \frac{0}{2\alpha-2\alpha} \quad .$$

Hence this formula does not lead to a value for x_2.

Considering the constituent equations when A is derogatory we have

(1) $\quad x_1^2 + x_2x_3 = \alpha$

(2) $\quad x_2(x_1+x_4) = 0$

(3) $\quad x_3(x_1+x_4) = 0$

(4) $\quad x_4^2 + x_2x_3 = \alpha \quad .$

There are two cases to consider

(1) If $x_1 + x_4 \neq 0$ then $x_2 = 0$, $x_3 = 0$ & $x_1 = \sqrt{\alpha}$, $x_4 = \sqrt{\alpha}$

Hence there are solutions $X = \begin{pmatrix} \sqrt{\alpha} & 0 \\ 0 & \sqrt{\alpha} \end{pmatrix}$.

(2) If $x_1 + x_4 = 0$ then setting $x_1 = a$ and $x_4 = -a$

from equation (1) $x_2x_3 = \alpha - a^2$

Hence if $x_3 = b$ $x_2 = \frac{\alpha - a^2}{b}$.

We therefore have the result that if A is derogatory the equation $X^2 = A$ has an infinite number of solutions of the form

$$X = \begin{pmatrix} a & \frac{\alpha - a^2}{b} \\ b & -a \end{pmatrix}$$

Example 2.5.2.

Consider the equation $X^2 = A$ where $A = \begin{pmatrix} 6 & 2 \\ 3 & 1 \end{pmatrix}$

Since A is singular there are only two solutions which may be obtained by substituting in the formula

$$X = \frac{1}{\sqrt{a_1 + a_4}} \begin{pmatrix} a_1 & a_2 \\ a_3 & a_4 \end{pmatrix}$$

Hence the two solutions are $\begin{pmatrix} \frac{6}{\sqrt{7}} & \frac{2}{\sqrt{7}} \\ \frac{3}{\sqrt{7}} & \frac{1}{\sqrt{7}} \end{pmatrix}$ and $\begin{pmatrix} \frac{-6}{\sqrt{7}} & \frac{-2}{\sqrt{7}} \\ \frac{-3}{\sqrt{7}} & \frac{-1}{\sqrt{7}} \end{pmatrix}$.

Example 2.5.3.

Consider the equation $X^2 = A$ where $A = \begin{pmatrix} 4 & 0 \\ 0 & 4 \end{pmatrix}$

Since A is derogatory there are 4 diagonal solutions

$$\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} \quad \begin{pmatrix} -2 & 0 \\ 0 & 2 \end{pmatrix} \quad \begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix} \quad \begin{pmatrix} -2 & 0 \\ 0 & -2 \end{pmatrix}$$

and also an infinite number of solutions of the form

$$X = \begin{pmatrix} a & \frac{4-a^2}{b} \\ b & -a \end{pmatrix} \quad \text{where a and b are arbitrary numbers.}$$

Consider the equation $XAX = B$ where $A = \begin{pmatrix} a_1 & 0 \\ 0 & a_2 \end{pmatrix}$ $B = \begin{pmatrix} b_1 & b_2 \\ b_3 & b_4 \end{pmatrix}$

The constituent equations are

(1) $a_1x_1^2 + a_2x_2x_3 - b_1 = 0$

(2) $a_1x_1x_2 + a_2x_2x_4 - b_2 = 0$

(3) $a_1x_1x_3 + a_2x_3x_4 - b_3 = 0$

(4) $a_1x_2x_3 + a_2x_4^2 - b_4 = 0$.

From equations (2) and (3) $x_3 = \frac{b_3x_2}{b_2}$, $b_2 \neq 0$.

Substituting in (1) and (4) gives

$$x_1 = \sqrt{\frac{b_1b_2 - a_2b_3x_2^2}{a_1b_2}} \qquad x_4 = \sqrt{\frac{b_2b_3 - a_1b_3x_2^2}{a_2b_2}} \quad , \; a_1, a_2 \neq 0.$$

and since $a_1x_1 + a_2x_4 = \frac{b_2}{x_2}$

a polynomial in x_2 may be obtained

$$b_2^2 \left\{(4a_1a_2b_2b_3+a_1^2b_1^2+a_2^2b_4^2-2a_1a_2b_1b_4)x_2^4 + (-2a_1b_1b_2^2-2a_2b_2^2b_4)x_2^2 + b_2^4\right\} = 0$$

and setting P = Trace (A B) and $|A|$ = det A $|B|$ = det B

then $$b_2^2 \left\{(P^2-4|A||B|)x_2^4 - 2b_2^2Px_2^2 + b_2^4\right\} = 0$$

and if $b_2 \neq 0$ then $x_2^2 = \dfrac{b_2^2}{P-2\sqrt{|AB|}}$ or $x_2^2 = \dfrac{b_2^2}{P+2\sqrt{|AB|}}$ and neither $P \pm \sqrt{|AB|}$ is zero.

Hence there may be 4 possible real values for x_2 and therefore 4 possible matrix solutions.

Example 2.5.4.

Consider the equation XAX = B where $A = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$ $B = \begin{pmatrix} -3 & 2 \\ -1 & -2 \end{pmatrix}$

then P = Trace (AB) = −7 and $|A| = 2$ $|B| = 8$

$$\therefore \quad x_2^2 = \frac{4}{-7-2\sqrt{16}} \quad \text{or} \quad x_2^2 = \frac{4}{-7+2\sqrt{16}}$$

$$\therefore \quad x_2 = 2, \ -2, \ \frac{2i}{\sqrt{15}}, \ \frac{-2i}{\sqrt{15}}$$

$\therefore$ By back substitution to obtain x_1 x_3 x_4 the two real solutions of the equation are

$$X = \begin{pmatrix} 1 & 2 \\ -1 & 0 \end{pmatrix} \qquad X = \begin{pmatrix} -1 & -2 \\ 1 & 0 \end{pmatrix} .$$

Consider the equation $X^2 + X - A = 0$ where $X = \begin{pmatrix} x_1 & x_2 \\ x_3 & x_4 \end{pmatrix}$ $A = \begin{pmatrix} a_1 & a_2 \\ a_3 & a_4 \end{pmatrix}$

The constituent equations are

(1) $x_1^2 + x_2x_3 + x_1 = a_1$

(2) $x_2(x_1+x_4+1) = a_2$

(3) $x_3(x_1+x_4+1) = a_3$

(4) $x_4^2 + x_2x_3 + x_4 = a_4$

Eliminating x_1 x_3 x_4 a polynomial in x_2 is obtained

$$16a_2^4 \left\{(P^2-4|A|)x_2^4 + a_2^2(-1-2P)x_2^2 + a_2^4\right\} = 0$$

where P = Trace A and $|A|$ = det A

and if $a_2 \neq 0$ then $x_2^2 = \frac{a_2^2}{2}\left\{\frac{(1+2P) \pm \sqrt{1+4P+16|A|}}{P^2 - 4|A|}\right\}$

and $x_3 = \frac{a_3}{a_2} x_2$

$$x_1 = \frac{1}{2}\left\{\frac{a_2}{x_2} - 1 + \frac{(a_1-a_4)}{a_2} x_2\right\}$$

$$x_4 = \frac{1}{2}\left\{\frac{a_2}{x_2} - 1 - \frac{(a_1-a_4)}{a_2} x_2\right\}$$

Example 2.5.5.

Consider the equation $X^2 + X - A = 0$ where $A = \begin{pmatrix} 0 & -5 \\ 10 & 10 \end{pmatrix}$

then $P = \text{Trace } A = 10$ and $\det A = 50$.

Substituting these values in the formula for x_2 gives

$$x_2^2 = \frac{25}{2}\left\{\frac{21 \pm \sqrt{841}}{100 - 200}\right\} \quad \therefore \quad x_2^2 = -\frac{25}{4} \text{ or } x_2^2 = 1$$

$\therefore$ the two real solutions for x_2 are 1 and -1 .

If $x_2 = 1$ then substitution in the formulae for x_1 x_3 x_4 gives

$$x_1 = -2 \qquad x_3 = -2 \qquad x_4 = -4$$

$\therefore$ a solution of the matrix equation is $X = \begin{pmatrix} -2 & 1 \\ -2 & -4 \end{pmatrix}$.

If $x_2 = -1$ then substitution in the formulae gives

$$x_1 = 1 \qquad x_3 = 2 \qquad x_4 = 3$$

$\therefore$ a second solution of the matrix equation is $X = \begin{pmatrix} 1 & -1 \\ 2 & 3 \end{pmatrix}$

It may be concluded that in consideration of the equations $X^2 = A$, $XAX = B$, $X^2 + X - A = 0$ where X, A, B are 2×2 matrices, the element x_2 can always be obtained directly as a function of the Trace and determinant of the coefficient matrices and by back substitution the other elements of X may also be obtained.

2.6 CONCLUSION.

Since any matrix equation can be expressed as a set of polynomial equations, solutions may be sought by consideration of

the constituent equations. Since the existence of a solution to the constituent equations depends on the values of the coefficients it is possible that consideration of the constituent equations could give conditions on the coefficient matrices necessary for a solution.

This is clearly obtained in the simple case of the equation $AX = B$ where $A = \begin{pmatrix} a_1 & a_2 \\ a_3 & a_4 \end{pmatrix}$ $X = \begin{pmatrix} x_1 & x_2 \\ x_3 & x_4 \end{pmatrix}$ $B = \begin{pmatrix} b_1 & b_2 \\ b_3 & b_4 \end{pmatrix}$

This is equivalent to the set of equations

(1) $a_1x_1 + a_2x_3 = b_1$

(2) $a_1x_2 + a_2x_4 = b_2$

(3) $a_3x_1 + a_4x_3 = b_3$

(4) $a_3x_2 + a_4x_4 = b_4$

The solution of this system is

$$x_1 = \frac{a_4b_1 - a_2b_3}{a_1a_4 - a_2a_3} \qquad x_2 = \frac{a_4b_2 - a_2b_4}{a_1a_4 - a_2a_3} \qquad x_3 = \frac{a_1b_3 - a_3b_1}{a_1a_4 - a_2a_3}$$

$$x_4 = \frac{a_1b_4 - a_3b_2}{a_1a_4 - a_2a_3} .$$

The constituent equations clearly have a unique solution provided that $a_1a_4 - a_2a_3 \neq 0$ and this gives the condition on the coefficient matrix that $\det A \neq 0$.

In matrix equations of higher degree or involving matrices of higher order, while it is possible to obtain a particular solution with computer methods by consideration of the constituent

equations, it is difficult to obtain a general solution and hence deduce any conditions on the matrix coefficients.

The results and difficulties encountered in applying the methods discussed in this chapter appear to support the opinion of Ingraham [1941] who stated that expressing the matrix equation as a set of polynomial equations was the 'worst possible algorithm'. He said that a method which shows that ultimately a matrix problem can be solved in a finite number of steps but shows little else is of limited value. His view was that methods which essentially use the matrical properties of the elements of the equation were the only ones worth considering.

All the methods considered in this chapter suffer from the same disadvantages of practical computation. The degree of the final equation in a single variable increases very rapidly as the order of the matrices involved increases. In the case of the quadratic matrix equation it will be shown in Chapter 4 that if the matrices are m×m then there can be ${}^{2m}C_m$ matrix solutions. Hence for a quadratic matrix equation involving 4×4 matrices the number of possible solutions would be ${}^{8}C_4 = 70$. The final equation in a single variable would therefore be a polynomial equation of degree 70.

It must be concluded that computational difficulties exist in attempting to deduce matrical properties for solvability of matrix equations by consideration of the constituent equations.

CHAPTER 3

A Review of Established Methods of Solution of Various Matrix Equations

3.1 HISTORICAL BACKGROUND.

The study of matrix equations began as long ago as 1858 when Cayley first discussed the equation $X^2 = P$ for matrices of order 2×2 and 3×3. He showed that a square matrix satisfies an algebraic equation of degree equal to its own order. This equation $\det [P - \lambda I] = 0$ is now known as the characteristic equation of P.

Cayley used the characteristic equation in the 2×2 case to find solutions of the equation $X^2 = P$. He showed that by substituting the matrix P for X^2 in the characteristic equation of X, a linear equation in X could be obtained from which X could be obtained in terms of its own characteristic coefficients a_1, a_2 and the matrix P

$$X = -\frac{1}{a_1}(P + a_2 I) \quad .$$

By using the fact that $a_1 = -$ Trace X and $a_2 = \det X$, two equations in the unknowns a_1 and a_2 are obtained and solutions of these equations lead to matrix solutions X of the equation $X^2 = P$.

Sylvester [1884,(A)] published many new results concerning matrix equations. He discovered that in the 2×2 case the equation $X^2 = P$ has 4 solutions if P has distinct eigenvalues and 2 solutions if P has equal eigenvalues.

In considering the equation XP = PX Sylvester [1884,(B)] applied the name derogatory matrix to the class of matrices which satisfy an

equation whose degree is less than the order of the matrix. He obtained conditions for the solution of the equation PX = XQ in the 2×2 case by considering the 4 constituent equations in the elements of P, Q and X. He showed that a non zero solution existed only if

$$(\alpha_2-\beta_2)^2 + (\beta_1-\alpha_1)(\alpha_2\beta_1-\alpha_1\beta_2) = 0$$

where α_1 α_2 and β_1 β_2 are the characteristic coefficients of P and Q.

This expression is the resultant of the characteristic polynomials of P and Q. Hence, the condition that this resultant should be equal to zero is precisely the condition that the polynomials should have a common factor and hence that P and Q should have a common eigenvalue.

Sylvester [1884, (c)] went on to extend this result to matrices of any order and also obtained results concerning the general linear equation $A_1XB_1 + A_2XB_2 + \ldots + A_nXB_n = C$. He expressed it as

$$\left[A_1 \otimes B_1^T + A_2 \otimes B_2^T + \ldots + A_n \otimes B_n^T\right] \underline{x} = \underline{c}$$

where $\otimes$ denotes the Direct or Kronecker product and $\underline{x}$ and $\underline{c}$ are column vectors formed from the rows of X and C respectively taken in order. The matrix

$$\left[A_1 \otimes B_1^T + A_2 \otimes B_2^T + \ldots + A_n \otimes B_n^T\right]$$

was called by him the nivellateur although he did not recognise it as the sum of direct products.

In the same year [1884] Sylvester (D) began to study the

unilateral matrix quadratic equation. Considering the matrix equation

$$X^2 - 2PX + Q = 0$$

where X, P, Q are 2×2 matrices, and defining the characteristic equation of X as

$$X^2 - 2a_1X + a_2I = 0 \ ,$$

Sylvester showed that X could be expressed in terms of the unknown characteristic coefficients as

$$X = \frac{1}{2} [P-a_1I]^{-1} [Q-a_2I] \quad .$$

Using the fact that $2a_1$ = Trace X and a_2 = det X then a_1 and a_2 can be obtained in terms of the known elements of P and Q and hence the matrix solution X obtained.

Sylvester also showed in the 2×2 case that every characteristic root of a solution X of

$$A_oX^2 + A_1X + A_2 = 0$$

is a root of

$$\det|A_o\lambda^2 + A_1\lambda + A_2| = 0 \quad .$$

He suggested that this could be extended to the unilateral equation of degree n. This was later proved by Buchheim [1884] who showed that in general, if X satisfies

$$A_oX^n + A_1X^{n-1} + \ldots + A_n = 0$$

then every characteristic root of X satisfies

$$\det|A_o\lambda^n + A_1\lambda^{n-1} + \ldots . A_n| = 0.$$

Sylvester (E) suggested that if the A_i and X are m×m matrices then the characteristic polynomial of a solution X will be a

factor of degree m of the determinant

$$\det\left[A_o\lambda^n + A_1\lambda^{n-1} + \ldots\ldots A_n\right] \quad .$$

Since the determinant is of degree mn then the maximum number of solutions will be the number of combinations of mn elements chosen m at a time.

He suggested that solutions of the matrix equation may be obtained by choosing a factor $\phi(\lambda)$ of degree m from

$$\det\left|A_o\lambda^n + A_1\lambda^{n-1} + \ldots\ldots A_n\right|$$

as the characteristic polynomial of a solution X. Then by combining $\phi(X) = 0$ with

$$A_oX^n + A_1X^{n-1} + \ldots\ldots A_n = 0,$$

higher powers of X are eliminated until a linear equation in X is obtained from which the solution may be found.

Sylvester's results and publications in 1884 provided the foundation for further study which was subsequently undertaken by many mathematicians.

Frobenius [1896] studied the equation $X^2 = P$ and Baker [1925] and Dickson [1926] later extended his work to find all the solutions of $X^m = P$ which are expressible as polynomials in P.

Kreis [1906], Roth [1928] and Franklin [1932] studied the equation $p(X) = A$ where $p(\lambda)$ is a polynomial with complex coefficients. Kreis and Roth obtained solutions which are polynomials in A and Franklin gave a method for finding solutions which are not expressible as polynomials in A.

Roth [1930] considered the solution of the general unilateral

matrix equation

$$A_o X^n + A_1 X^{n-1} + \ldots . A_n = 0$$

where the A_i are p×q matrices and the X is a q×q matrix. This method is described later in this chapter along with several other methods which have been established since then.

Bell [1950] has shown that the unilateral matrix equation has an infinite family of solutions if and only if there exist two distinct solutions X_1 and X_2 which are similar.

A great deal of research into matrix methods and equations has taken place in the last fifty years and the field of knowledge has expanded rapidly. With the advent of computers, numerical methods have advanced tremendously.

The following sections include a small selection of methods for solving various matrix equations.

3.2 A SURVEY OF SOME METHODS OF SOLUTION OF THE UNILATERAL MATRIX EQUATION.

This section is devoted to a study of four methods of solution of the unilateral matrix equation

$$A_o X^n + A_1 X^{n-1} + A_2 X^{n-2} + \ldots . + A_n = 0 \quad . \qquad (3.2.1)$$

The coefficients A_i $i = 0,1,2,\ldots,n$ are matrices with constant elements and X is a square matrix of unknown elements.

Defining the lambda matrix $A(\lambda)$ as

$$A(\lambda) = A_o \lambda^n + A_1 \lambda^{n-1} + \ldots . + A_n$$

then it can be shown [Lancaster, 1966] that if X is a solution of (3.2.1) then $[\lambda I - X]$ is a right (left) factor of $A(\lambda)$.

Proof.

$$A(\lambda) = A_o\lambda^n + A_1\lambda^{n-1} + ... + A_n$$

$$= A_o\lambda^{n-1}(\lambda I-X) + (A_oX+A_1)\lambda^{n-1} + A_2\lambda^{n-2} + ... + A_n$$

$$= \left[A_o\lambda^{n-1} + (A_oX+A_1)\lambda^{n-2}\right](\lambda I-X)$$

$$+ (A_oX^2+A_1X+A_2)\lambda^{n-2} + A_3\lambda^{n-3} + ... + A_n$$

$$= \left[A_o\lambda^{n-1} + (A_oX+A_1)\lambda^{n-2} + (A_oX^2+A_1X+A_2)\lambda^{n-3} + ...\right.$$

$$\left.(A_oX^{n-1}+A_1X^{n-2} + ... A_{n-1})\right](\lambda I-X)$$

$$+ A_oX^n + A_1X^{n-1} + ... A_{n-1}X + A_n$$

Hence the lamda matrix $A(\lambda)$ may be written as

$$A(\lambda) = Q_1(\lambda)[\lambda I-X] + A_oX^n + A_1X^{n-1} + ... A_{n-1}X + A_n \quad .$$

Similarly it may be shown that $A(\lambda)$ may be written as

$$A(\lambda) = [\lambda I-X]Q_2(\lambda) + A_oX^n + A_1X^{n-1} + ... A_{n-1}X + A_n$$

where $Q_1(\lambda)$, $Q_2(\lambda)$ are polynomials in λ with matrix coefficients, and hence if X is a solution of (3.2.1) then

$$A(\lambda) = Q_1(\lambda)[\lambda I-X] \quad \text{or} \quad A(\lambda) = [\lambda I-X]Q_2(\lambda) \quad .$$

From this it can be seen that

$$\det.A(\lambda) = \det Q(\lambda).\ \det[\lambda I-X]$$

and hence $\det[\lambda I-X]$ is a factor of $\det A(\lambda)$ or the characteristic polynomial of a solution X is a factor of $\det.A(\lambda)$.

This result is frequently used in the methods of solution which follow.

Method 1.

This method is described by Gantmacher [1959] and uses the result previously proved, that if X is a solution of (3.2.1) then the characteristic polynomial of X is a factor of det $A(\lambda)$.

Since X satisfies its own characteristic equation then X also satisfies the scalar equation

$$g(\lambda) = 0 \quad \text{where} \quad g(\lambda) = \det A(\lambda).$$

If $g(\lambda) = 0$ has solutions $\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_r$ then $g(X) = 0$ has an infinite number of solutions of the form TDT^{-1} where D is a diagonal matrix with λ_i on the diagonal.

Since solutions of $g(X) = 0$ are of the form TDT^{-1} then solutions of the same form are sought for equation (3.2.1).

Substitution in the equation gives

$$A_o\left[TDT^{-1}\right]^n + A_1\left[TDT^{-1}\right]^{n-1} + \ldots. + A_n = 0$$

$$\therefore \quad A_o TD^n T^{-1} + A_1 TD^{n-1} T^{-1} + \ldots. + A_n = 0$$

and multiplying on the right by T

$$A_o TD^n + A_1 TD^{n-1} + \ldots. A_n T = 0 \quad .$$

Since the A_i and D are known this gives a linear equation in T. Solving for T gives the particular transforming matrix which is necessary to obtain a solution of (3.2.1).

Example 3.2.1.

Consider the equation $A_o X^2 + A_1 X + A_2 = 0$

where

$$A_o = \begin{pmatrix} 3 & 1 \\ 4 & 2 \end{pmatrix} \qquad A_1 = \begin{pmatrix} 2 & 3 \\ 1 & 2 \end{pmatrix} \qquad A_2 = \begin{pmatrix} -16 & -15 \\ -18 & -17 \end{pmatrix}$$

Then $g(\lambda) = \det \left[A_o\lambda^2 + A_1\lambda + A_2\right]$

$= (\lambda-2)(\lambda-1)(\lambda+1)(2\lambda+1)$

$\therefore$ $g(\lambda) = 0$ has solutions $\lambda_1 = 2 \quad \lambda_2 = 1 \quad \lambda_3 = -1 \quad \lambda_4 = -\frac{1}{2}$.

Choosing $D = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$ then $D^2 = \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix}$

and setting $T = \begin{pmatrix} t_1 & t_2 \\ t_3 & t_4 \end{pmatrix}$

then $A_oTD^2 + A_1TD + A_2T = 0$

becomes $\begin{pmatrix} -11t_1-11t_3 & -5t_4 \\ -13t_1-13t_3 & -5t_4 \end{pmatrix} = 0$

$\therefore$ T is any matrix with $t_1 = -t_3$ and $t_4 = 0$

$\therefore$ $T = \begin{pmatrix} p & q \\ -p & 0 \end{pmatrix}$ and $T^{-1} = \begin{pmatrix} 0 & -\frac{1}{p} \\ \frac{1}{q} & \frac{1}{q} \end{pmatrix}$

$\therefore$ $X = TDT^{-1} = \begin{pmatrix} p & q \\ -p & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} 0 & -\frac{1}{p} \\ \frac{1}{q} & \frac{1}{q} \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 0 & 1 \end{pmatrix}$.

Other solutions may be obtained from different choices of elements for D

$D = \begin{pmatrix} -1 & 0 \\ 0 & 2 \end{pmatrix}$ gives $X = \begin{pmatrix} 2 & 3.4 \\ 0 & -1 \end{pmatrix}$

$$D = \begin{pmatrix} -0.5 & 0 \\ 0 & 2 \end{pmatrix} \quad \text{gives} \quad X = \begin{pmatrix} 2 & 2.5 \\ 0 & -0.5 \end{pmatrix}$$

$$D = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \quad \text{gives} \quad X = \begin{pmatrix} -16 & -17 \\ 15 & 16 \end{pmatrix}$$

$$D = \begin{pmatrix} -1 & 0 \\ 0 & -0.5 \end{pmatrix} \quad \text{gives} \quad X = \begin{pmatrix} -4.75 & -4.25 \\ 3.75 & 3.25 \end{pmatrix} .$$

However the choice of $D = \begin{pmatrix} 1 & 0 \\ 0 & -\frac{1}{2} \end{pmatrix}$ does not lead to a solution since this gives $T = \begin{pmatrix} a & b \\ -a & -b \end{pmatrix}$ which is singular and hence a solution of the form TDT^{-1} does not exist for this choice of D.

This shows that every factor of det $A(\lambda)$ is not necessarily the characteristic polynomial of a solution X of the equation (3.2.1).

Method 2.

This method also obtains the solution by finding a diagonal matrix D and a suitable transforming matrix. It is described in Dennis, Traub and Weber [1976] and is virtually the same as Method 1 but the terminology is different and is in fact the same as is used by Lancaster.

The terms latent roots and latent vectors are used. They are defined as follows.

A solution of λ_1 of $g(\lambda) = 0$ where $g(\lambda) = \det |A(\lambda)|$ is called a latent root of the lambda matrix $A(\lambda)$ and a vector $\underline{b}$ is called a latent vector if, for a particular latent root λ_1

then $A(\lambda_1).\ \underline{b} = 0$.

The method uses the fact that if X is an m×m matrix and if $\underline{b}_1\ \underline{b}_2,\ldots,\underline{b}_m$ are linearly independent (m×1) latent vectors corresponding to the latent roots $\lambda_1,\lambda_2,\ldots,\lambda_m$ then

$X = TDT^{-1}$ is a solution of (3.2.1)

where $T = (\underline{b}_1\ \underline{b}_2,\ldots,\underline{b}_m)$

and $$D = \begin{pmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \lambda_m \end{pmatrix} .$$

Example 3.2.2.

Consider the equation $A_oX^2 + A_1X + A_2 = 0$

where

$$A_o = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad A_1 = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 2 \\ 0 & 1 & 1 \end{pmatrix} \quad A_2 = \begin{pmatrix} -4 & -2 & -2 \\ -2 & -1 & -1 \\ 0 & 2 & -2 \end{pmatrix}$$

then $g(\lambda) = \det |A(\lambda)| = \lambda^2(\lambda-2)^2(\lambda+2)(\lambda+3)$

$\therefore$ The latent roots are $\lambda_1 = 0 \quad \lambda_2 = 2 \quad \lambda_3 = -2 \quad \lambda_4 = -3$.

Taking $\lambda_1 = 0$ then $A(\lambda_1) = \begin{pmatrix} -4 & -2 & -2 \\ -2 & -1 & -1 \\ 0 & 2 & -2 \end{pmatrix}$.

The latent vector corresponding to $\lambda_1 = 0$ is $\begin{pmatrix} -a \\ a \\ a \end{pmatrix}$.

Similarly the latent vectors corresponding to

$$\lambda_2 = 2 \quad \lambda_3 = -2 \quad \lambda_4 = -3 \text{ are } \begin{pmatrix} b \\ -c \\ c \end{pmatrix} \quad \begin{pmatrix} -2d \\ -d \\ d \end{pmatrix} \quad \begin{pmatrix} 5e \\ 4e \\ e \end{pmatrix} .$$

Choosing $D_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & -2 \end{pmatrix}$ then $T = \begin{pmatrix} -a & b & -2d \\ a & -c & -d \\ a & c & d \end{pmatrix}$

and $X = TD_1T^{-1} = \begin{pmatrix} -\frac{2}{3} & -\frac{5}{3} & 1 \\ -\frac{4}{3} & -\frac{1}{3} & -1 \\ \frac{4}{3} & \frac{1}{3} & 1 \end{pmatrix}$.

Choosing $D_2 = \begin{pmatrix} 2 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -3 \end{pmatrix}$ then $T = \begin{pmatrix} b & -2d & 5e \\ -c & -d & 4e \\ c & d & e \end{pmatrix}$

and $X = TD_2T^{-1} = \begin{pmatrix} -\frac{2}{3} & -\frac{43}{15} & -\frac{1}{5} \\ -\frac{4}{3} & -\frac{14}{15} & -\frac{8}{5} \\ \frac{4}{3} & -\frac{31}{15} & -\frac{7}{5} \end{pmatrix}$.

Choosing $D_3 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -3 \end{pmatrix}$ then $T = \begin{pmatrix} -a & -2d & 5e \\ a & -d & 4e \\ a & d & e \end{pmatrix}$

and $$X = TD_3T^{-1} = \begin{pmatrix} -2 & -1 & -1 \\ -2 & 0 & -2 \\ -\frac{4}{3} & \frac{5}{3} & -3 \end{pmatrix}$$

Choosing $$D_4 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & -3 \end{pmatrix}$$ then $$T = \begin{pmatrix} -a & b & 5e \\ a & -c & 4e \\ a & c & e \end{pmatrix}$$

and $$X = TD_4T^{-1} = \begin{pmatrix} -\frac{4}{3} & -\frac{7}{3} & 1 \\ -\frac{5}{3} & -\frac{2}{3} & -1 \\ 0 & -1 & 1 \end{pmatrix} .$$

Method 3.

An algorithm for the solution of (3.2.1) is given by Ingraham [1941]. The following results and definitions are used.

(1) A matrix with elements which are polynomials in λ is said to be unimodular if its determinant is a constant k., $k \neq 0$.

(2) A matrix $A(\lambda)$ is in upper (lower) triangular form if all the elements below (above) the main diagonal are zero. It is in canonical triangular form if all the elements above (below) the main diagonal are of lower degree than the elements in the same column on the main diagonal and if when a zero occurs on the main diagonal the whole row in which it occurs is zero.

A description of the algorithm follows.

As in the previous methods we define the lambda matrix

$$A(\lambda) = A_o\lambda^n + A_1\lambda^{n-1} + \ldots + A_n \quad .$$

As shown previously, if X is a solution of (3.2.1) then $[\lambda I-X]$ is a right factor of $A(\lambda)$.

Then $A(\lambda) = P(\lambda)[\lambda I-X]$ for some lambda matrix $P(\lambda)$ and if U is a unimodular lambda matrix, then

$$UA(\lambda) = UP(\lambda)[\lambda I-X] \quad .$$

Hence if X is a solution of (3.2.1) then $[\lambda I-X]$ is a right factor of $UA(\lambda)$.

Moreover for any lambda matrix A, a particular unimodular matrix U may be chosen so that $UA(\lambda)$ is in canonical triangular form. Writing A instead of $A(\lambda)$ for ease of notation and assuming that any new matrices introduced in the following algorithm are lambda matrices, then since $[\lambda I-X]$ is a right factor of UA we may write $UA = H[\lambda I-X]$ where H is a lambda matrix.

$\therefore$ $UA = HV^{-1}V[\lambda I-X]$ where V is the unimodular matrix such that $V[\lambda I-X]$ is in canonical triangular form.

$\therefore$ $UA = HV^{-1}[V(\lambda I-X)]$

and hence $V[\lambda I-X]$ is a right factor of UA.

The problem is therefore reduced to finding the triangular factors of UA which are the canonical triangular forms of matrices of the type $[\lambda I-X]$ where X is independent of λ.

Let $T = V[\lambda I-X]$

Then $UA = (HV^{-1})T$ and since UA and T are in canonical

triangular form then (HV^{-1}) is also in triangular form and hence the diagonal elements of T are factors of the corresponding elements of UA.

The problem is now to determine an X such that T satisfies the required conditions.

If $T = V[\lambda I - X]$

then $V^{-1} = (\lambda I - X)T^{-1}$

and since V is defined to be unimodular then V^{-1} will have elements which are polynomials in λ rather than rational functions of λ, since the determinant of V is a constant function.

The problem is therefore reduced to finding an X such that $[\lambda I - X]T^{-1}$ is a matrix with elements which are polynomials in λ.

The algorithm is illustrated in the following example.

Example 3.2.3.

Consider the equation $A_o X^2 + A_1 X + A_2 = 0$

where $A_o = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ $A_1 = \begin{pmatrix} -2 & 1 \\ 1 & 0 \end{pmatrix}$ $A_2 = \begin{pmatrix} 1 & -1 \\ 1 & -1 \end{pmatrix}$

$\therefore$ $A(\lambda) = \begin{pmatrix} \lambda^2-2\lambda+1 & \lambda-1 \\ \lambda+1 & \lambda^2-1 \end{pmatrix}$

and $UA = \begin{pmatrix} 1 & -\frac{\lambda^3}{4} + \frac{3}{4}\lambda^2 + \frac{\lambda}{2} - 1 \\ 0 & \lambda(\lambda-1)(\lambda+1)(\lambda-2) \end{pmatrix}$

where $U = \begin{pmatrix} \frac{1}{4} & \frac{(-\lambda+3)}{4} \\ -\lambda-1 & \lambda^2-2\lambda+1 \end{pmatrix}$

Now writing $UA = (HV^{-1})T$ where $T = V[\lambda I - X]$ and T is in canonical triangular form, then the diagonal elements of T are factors of the corresponding elements of UA which are 1 and $\lambda(\lambda-1)(\lambda+1)(\lambda-2)$.

Choosing 1 and $\lambda(\lambda+1)$ as the diagonal elements of T then the diagonal elements of (HV^{-1}) are 1 and $(\lambda-1)(\lambda-2)$

$$\therefore (HV^{-1})T = \begin{pmatrix} 1 & h_1\lambda+h_2 \\ 0 & (\lambda-1)(\lambda-2) \end{pmatrix} \begin{pmatrix} 1 & t_1\lambda+t_2 \\ 0 & \lambda(\lambda+1) \end{pmatrix}$$

$$= \begin{pmatrix} 1 & h_1\lambda^3+(h_1+h_2)\lambda^2+(h_2+t_1)\lambda+t_2 \\ 0 & \lambda(\lambda-1)(\lambda+1)(\lambda-2) \end{pmatrix}$$

$$\text{But } (HV^{-1})T = UA = \begin{pmatrix} 1 & -\frac{\lambda^3}{4}+\frac{3}{4}\lambda^2+\frac{\lambda}{2}-1 \\ 0 & \lambda(\lambda-1)(\lambda+1)(\lambda-2) \end{pmatrix} .$$

$$\text{Hence } \quad h_1\lambda^3 + (h_1+h_2)\lambda^2 + (h_2+t_1)\lambda + t_2 \equiv -\frac{\lambda^3}{4}+\frac{3}{4}\lambda^2+\frac{\lambda}{2}-1$$

and comparing coefficients of λ gives

$$h_1 = -\frac{1}{4} \qquad h_2 = 1 \qquad t_1 = -\frac{1}{2} \qquad t_2 = -1 \ .$$

We have therefore obtained the matrix T

$$T = \begin{pmatrix} 1 & -\frac{\lambda}{2}-1 \\ 0 & \lambda(\lambda+1) \end{pmatrix} .$$

Now $T = V[\lambda I - X]$

$$\therefore \quad V^{-1} = [\lambda I - X]T^{-1}$$

and since V is defined to be unimodular, the matrix $[\lambda I-X]T^{-1}$

must have elements which are polynomials in λ rather than rational functions of λ.

Let $X = \begin{pmatrix} x_1 & x_2 \\ x_3 & x_4 \end{pmatrix}$

then $[\lambda I - X]T^{-1} = \begin{pmatrix} \lambda - x_1 & -x_2 \\ -x_3 & \lambda - x_4 \end{pmatrix} \begin{pmatrix} 1 & \frac{\frac{\lambda}{2} + 1}{\lambda(\lambda+1)} \\ 0 & \frac{1}{\lambda(\lambda+1)} \end{pmatrix}$

$$= \begin{pmatrix} \lambda - x_1 & \dfrac{\frac{\lambda^2}{2} + \left(1 - \frac{x_1}{2}\right)\lambda - x_1 - x_2}{\lambda(\lambda+1)} \\ -x_3 & \dfrac{\lambda\left(1 - \frac{x_3}{2}\right) - x_3 - x_4}{\lambda(\lambda+1)} \end{pmatrix} .$$

Hence $\lambda(\lambda+1)$ must be a factor of $\frac{\lambda^2}{2} + \left(1 - \frac{x_1}{2}\right)\lambda - x_1 - x_2$.

This imposes the restriction that $x_1 = 1$ and $x_2 = -1$ and $\lambda(\lambda+1)$ must be a factor of $\lambda\left(1 - \frac{x_3}{2}\right) - x_3 - x_4$.

This imposes the restriction that $x_3 = 2$ and $x_4 = -2$.

Hence we have obtained the values of $x_1\ x_2\ x_3\ x_4$ which ensure that the elements of $[\lambda I - X]T^{-1}$ are polynomials in λ.

For these values then X is a solution of the equation

$$A_o X^2 + A_1 X + A_2 = 0$$

$$\therefore \quad X = \begin{pmatrix} 1 & -1 \\ 2 & 2 \end{pmatrix} .$$

Other solutions may be obtained by choosing other combinations of factors for the diagonal elements of T.

E.g., choosing the diagonal elements to be 1 and $(\lambda-1)(\lambda+1)$ leads to

$$T = \begin{pmatrix} 1 & \frac{\lambda}{4} - \frac{1}{4} \\ 0 & (\lambda-1)(\lambda+1) \end{pmatrix}$$

and $$X = \begin{pmatrix} -1 & 0 \\ -4 & 1 \end{pmatrix} .$$

Ingraham gives an algorithm for obtaining a solution X from the matrix T.

The steps are as follows:

(1) Augment each element of T by the proper powers of λ with zero coefficients so that terms of the same degree as in the corresponding diagonal elements appear in each column.

(2) Break up each column into separate columns each one of which involves monomials of the same degree in λ.

e.g. $$\begin{pmatrix} a\lambda+b \\ c\lambda+d \end{pmatrix} \rightarrow \begin{pmatrix} a\lambda & b \\ c\lambda & d \end{pmatrix} .$$

(3) Delete the columns which do not involve λ.

(4) Set $\lambda = 1$ obtaining the matrix D.

A necessary and sufficient condition that there exists a matrix X such that $T = V[\lambda I - X]$ is that D is non-singular.

Let $$T = T_o + T_1\lambda + T_2\lambda^2 + \ldots . T_n\lambda^n$$

A solution X may be obtained from the matrix D.

First find the matrix B as follows:

If $t_{jj} \neq 1$ then $(b_{j1}\ b_{j2} \ldots . b_{jm}) = (0\ldots.0, -1, 0\ldots.0)D^{-1}$

where the -1 occurs in the position corresponding to the column of D containing $t_{1_{jj}}$ (where $t_{kj} = t_{o_{kj}} + t_{1_{kj}}\lambda + t_{2_{kj}}\lambda^2 \ldots\ldots$).

If $t_{jj} = 1$ then $(b_{j1}\ b_{j2}\ \ldots\ b_{jm}) = \alpha_j D^{-1}$ where α_j is a $1\times m$ vector obtained by taking the j^{th} row of the matrix obtained in Step 2 by deleting the columns involving the leading term of t_{kk} for every k and setting $\lambda = 1$.

Then the solution is $X = BT_o$.

The algorithm may be illustrated by finding further solutions for example 3.2.3.

Taking $T = \begin{pmatrix} 1 & \lambda-1 \\ 0 & \lambda(\lambda-2) \end{pmatrix}$

Step (2) $\begin{pmatrix} 1 & 0 & \lambda & -1 \\ 0 & \lambda^2 & -2\lambda & 0 \end{pmatrix}$

Step (3) $\begin{pmatrix} 0 & \lambda \\ \lambda^2 & -2\lambda \end{pmatrix}$

Step (4) $D = \begin{pmatrix} 0 & 1 \\ 1 & -2 \end{pmatrix}$ and hence $D^{-1} = \begin{pmatrix} 2 & 1 \\ 1 & 0 \end{pmatrix}$

$$(b_{11}\ b_{12}) = (1 \quad -1)\begin{pmatrix} 2 & 1 \\ 1 & 0 \end{pmatrix} = (1 \quad 1)$$

$$(b_{21}\ b_{22}) = (0 \quad -1)\begin{pmatrix} 2 & 1 \\ 1 & 0 \end{pmatrix} = (-1 \quad 0)$$

$\therefore$ $B = \begin{pmatrix} 1 & 1 \\ -1 & 0 \end{pmatrix}$ and a solution X is obtained from $X = BT_o$

$$\therefore \quad X = \begin{pmatrix} 1 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} .$$

Taking $T = \begin{pmatrix} 1 & \lambda-1 \\ 0 & \lambda(\lambda-1) \end{pmatrix}$ gives $D = \begin{pmatrix} 0 & 1 \\ 1 & -1 \end{pmatrix}$ $\therefore D^{-1} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$

$$(b_{11}\ b_{12}) = (1 \quad -1) \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} = (0 \quad 1)$$

$$(b_{21}\ b_{22}) = (0 \quad -1) \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} = (-1 \quad 0)$$

$$\therefore \quad X = BT_o = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ -1 & 1 \end{pmatrix}$$

Taking $T = \begin{pmatrix} 1 & \lambda-1 \\ 0 & (\lambda-1)(\lambda-2) \end{pmatrix}$ gives $D = \begin{pmatrix} 0 & 1 \\ 1 & -3 \end{pmatrix}$ and $D^{-1} = \begin{pmatrix} 3 & 1 \\ 1 & 0 \end{pmatrix}$

Also $B = \begin{pmatrix} 2 & 1 \\ -1 & 0 \end{pmatrix}$

$$\therefore \quad X = BT_o = \begin{pmatrix} 2 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 0 & 2 \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ -1 & 1 \end{pmatrix}$$

Taking $T = \begin{pmatrix} 1 & \frac{1}{2}\lambda \\ 0 & (\lambda+1)(\lambda-2) \end{pmatrix}$ gives $D = \begin{pmatrix} 0 & \frac{1}{2} \\ 1 & -1 \end{pmatrix}$ and $D^{-1} = \begin{pmatrix} 2 & 1 \\ 2 & 0 \end{pmatrix}$

Also $B = \begin{pmatrix} 1 & \frac{1}{2} \\ -2 & 0 \end{pmatrix}$

$$\therefore \quad X = BT_o = \begin{pmatrix} 1 & \frac{1}{2} \\ -2 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & -2 \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ -2 & 0 \end{pmatrix}$$

Since all combinations of factors of UA have now been used this shows that there are 6 solutions for the equation

$$A_o X^2 + A_1 X + A_2 = 0 \quad .$$

Method 4.

This method applies to the case when the coefficient matrices A_i of the unilateral matrix equation (3.2.1) are not necessarily square. It is described by Roth [1930].

Let the coefficient matrices A_i be p×q matrices and let X be a q×q matrix.

Again the lambda matrix $A(\lambda)$ is used where

$$A(\lambda) = A_o \lambda^n + A_1 \lambda^{n-1} + ... + A_n \quad .$$

The following properties hold:

(1) For every characteristic value λ_i of a solution X of (3.2.1) the rank of $A(\lambda_i)$ is less than q.

(2) If p > q the determinant of a q×q matrix formed from $A(\lambda)$ by deleting any (p-q) of its rows is divisible by the characteristic polynomial of a solution X or is identically zero.

(3) If p < q a solution X of (3.2.1) may have arbitrary characteristic values.

The following example illustrates how property (2) may be used to obtain a solution of (3.2.1) when p > q.

Example 3.2.4.

Consider the equation $A_o X^2 + A_1 X + A_2 = 0$

where $A_o = \begin{pmatrix} -2 & 1 \\ 0 & 2 \\ 1 & 0 \end{pmatrix}$ $A_1 = \begin{pmatrix} 1 & 0 \\ 2 & -1 \\ 0 & 3 \end{pmatrix}$ $A_2 = \begin{pmatrix} 5 & -3 \\ -16 & 6 \\ -13 & 3 \end{pmatrix}$

then $A(\lambda) = \begin{pmatrix} -2\lambda^2+\lambda+5 & \lambda^2-3 \\ 2\lambda-16 & 2\lambda^2-\lambda+6 \\ \lambda^2-13 & 3\lambda+3 \end{pmatrix}$

Deleting Row 1 of $A(\lambda)$, the determinant of the 2×2 matrix so formed is

$$\det \begin{pmatrix} 2\lambda-16 & 2\lambda^2-\lambda+6 \\ \lambda^2-13 & 3\lambda+3 \end{pmatrix}$$

$$= (\lambda^2-3\lambda+2)(-2\lambda^2-5\lambda+15)$$

Deleting Row 2 of $A(\lambda)$, the remaining determinant is

$$\det \begin{pmatrix} -2\lambda^2+\lambda+5 & \lambda^2-3 \\ \lambda^2-13 & 3\lambda+3 \end{pmatrix}$$

$$= (\lambda^2-3\lambda+2)(-\lambda^2-9\lambda-12)$$

Deleting Row 3 of $A(\lambda)$, the remaining determinant is

$$\det \begin{pmatrix} -2\lambda^2+\lambda+5 & \lambda^2-3 \\ 2\lambda-16 & 2\lambda^2-\lambda+6 \end{pmatrix}$$

$$= (\lambda^2-3\lambda+2)(-4\lambda^2-10\lambda-9) \quad .$$

By property (2) the three determinants obtained must be divisible by the characteristic polynomial of a solution X. It is clear that all three determinants are divisible by the factor $(\lambda^2-3\lambda+2)$ and hence this may be taken as the characteristic polynomial of a solution X.

Since the roots of $\lambda^2-3\lambda+2 = 0$ are $\lambda = 1$ and $\lambda = 2$, then the Jordan Normal Form of X is $\bar{X} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$ and solutions may be sought of the type $X = T\bar{X}T^{-1}$.

Following Method I we must find a matrix T such that

$$A_o T\bar{X}^2 + A_1 T\bar{X} + A_2 T = 0.$$

Let $T = \begin{pmatrix} t_1 & t_2 \\ t_3 & t_4 \end{pmatrix}$ then since $\bar{X} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$ and $\bar{X}^2 = \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix}$

substitution in the equation gives $t_3 = 2t_1$ and $t_2 = t_4$.

Hence $T = \begin{pmatrix} a & b \\ 2a & b \end{pmatrix}$ and $T^{-1} = \begin{pmatrix} -\frac{1}{a} & \frac{1}{a} \\ \frac{2}{b} & -\frac{1}{b} \end{pmatrix}$

$$\therefore \quad X = T\bar{X}T^{-1}$$

$$= \begin{pmatrix} a & b \\ 2a & b \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} -\frac{1}{a} & \frac{1}{a} \\ \frac{2}{b} & -\frac{1}{b} \end{pmatrix}$$

$$\therefore \quad X = \begin{pmatrix} 3 & -1 \\ 2 & 0 \end{pmatrix}.$$

We now consider the case where the coefficient matrices A_i

of equation (3.2.1) are p×q matrices and p < q.

As stated in property (1), the rank of $A(\lambda_i)$ is less than q, for every characteristic value λ_i of a solution X. When p < q this condition is always satisfied by an arbitrary λ_i.

Let $X = T\bar{X}T^{-1}$ be a solution of (3.2.1) where $\bar{X}$ is the Jordan Normal Form of X and X has arbitrary characteristic values ρ_i.

Roth describes a method for obtaining the transforming matrix T necessary to obtain solutions X.

If $(\rho_i-\lambda)^{m_i}$ are the elementary divisors of the matrix X and if $\underline{t}_j(\lambda)$ are matrices satisfying the identity $A(\lambda)\ \underline{t}_j(\lambda) \equiv 0$, then the m_i columns of the transforming matrix T corresponding to the particular characteristic value ρ_i are given by

$$t_{m_i}(\rho_i) \equiv \left\{ \underline{t}(\rho_i)\ \underline{t}'(\rho_i) \ldots\ \underline{t}^{(m_i-1)}\ \frac{(\rho_i)}{(m_i-1)!} \right\}$$

The method is illustrated in the following example.

Example 3.2.5.

Consider the equation $A_oX^2 + A_1X + A_2 = 0$

where $A_o = (2 \quad 1)$ $A_1 = (-1 \quad 1)$ $A_2 = (1 \quad -2)$

then $A(\lambda) = (2\lambda^2-\lambda+1 \quad \lambda^2+\lambda-2)$.

We need to find the matrix $\underline{t}(\lambda)$ such that $A(\lambda).\underline{t}(\lambda) \equiv 0$.

Let $\underline{t}(\lambda) \equiv \underline{t}_o + \underline{t}_1\lambda + \underline{t}_2\lambda^2$

then $A(\lambda)\ \underline{t}(\lambda) \equiv \left[A_o\lambda^2+A_1\lambda+ A_2\right]\left[\underline{t}_o+\underline{t}_1\lambda+\underline{t}_2\lambda^2\right]$

$$\equiv A_o\underline{t}_2\lambda^4 + \left[A_o\underline{t}_1+A_1\underline{t}_2\right]\lambda^3 + \left[A_o\underline{t}_o+A_1\underline{t}_1+A_2\underline{t}_2\right]\lambda^2$$

$$+ \left[A_1\underline{t}_o + A_2\underline{t}_1\right]\lambda + A_2\underline{t}_o \quad .$$

If $A(\lambda)\,\underline{t}(\lambda) \equiv 0$ then the coefficients of λ^r are all zero

$$\therefore \quad A_o \underline{t}_2 = 0$$

$$A_o \underline{t}_1 + A_1 \underline{t}_2 = 0$$

$$A_o \underline{t}_o + A_1 \underline{t}_1 + A_2 \underline{t}_2 = 0$$

$$A_1 \underline{t}_o + A_2 \underline{t}_1 = 0$$

$$A_2 \underline{t}_o = 0 \quad .$$

Solving these equation gives

$$\underline{t}_o = a \begin{pmatrix} -2 \\ -1 \end{pmatrix} \qquad \underline{t}_1 = a \begin{pmatrix} 1 \\ 1 \end{pmatrix} \qquad \underline{t}_2 = a \begin{pmatrix} 1 \\ -2 \end{pmatrix}$$

$$\therefore \quad \underline{t}(\lambda) \equiv a \left[\begin{pmatrix} -2 \\ -1 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \end{pmatrix} \lambda + \begin{pmatrix} 1 \\ -2 \end{pmatrix} \lambda^2 \right]$$

$$\equiv a \begin{bmatrix} -2+\lambda+\lambda^2 \\ -1+\lambda-2\lambda^2 \end{bmatrix}$$

$$\text{and} \quad \underline{t}'(\lambda) \equiv a \begin{bmatrix} 1+2\lambda \\ 1-4\lambda \end{bmatrix}$$

If $(\rho_1-\lambda)^2$ is the characteristic polynomial of X then the Jordan Normal Form of X is $\begin{pmatrix} \rho_1 & 1 \\ 0 & \rho_1 \end{pmatrix}$ and the columns of the transforming matrix T are given by

$$(\underline{t}(\rho_1), \quad \underline{t}'(\rho_1))$$

If however $(\rho_1-\lambda)(\rho_2-\lambda)$ is the characteristic polynomial of X then the Jordan Normal Form of X is $\begin{pmatrix} \rho_1 & 0 \\ 0 & \rho_2 \end{pmatrix}$ and the columns of the transforming matrix T are given by $(\underline{t}(\rho_1)\ \underline{t}(\rho_2))$.

Hence solutions are given by

$$X_1 = T_1 \begin{pmatrix} \rho_1 & 1 \\ 0 & \rho_1 \end{pmatrix} T_1^{-1} \quad \text{where} \quad T_1 = a \begin{pmatrix} \rho_1^2+\rho_1-2 & 2\rho_1+1 \\ -2\rho_1^2+\rho_1-1 & -4\rho_1+1 \end{pmatrix}$$

$$X_2 = T_2 \begin{pmatrix} \rho_1 & 0 \\ 0 & \rho_2 \end{pmatrix} T_2^{-1} \quad \text{where} \quad T_2 = a \begin{pmatrix} \rho_1^2+\rho_1-2 & \rho_2^2+\rho_2-2 \\ -2\rho_1^2+\rho_1-1 & -2\rho_2^2+\rho_2-1 \end{pmatrix}$$

E.g. if $\rho_1 = 1$ and the characteristic polynomial of a solution X is $(1-\lambda)^2$ then $T_1 = a \begin{pmatrix} 0 & 3 \\ -2 & -3 \end{pmatrix}$

and a solution $X_1 = a \begin{pmatrix} 0 & 3 \\ -2 & -3 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} -\frac{1}{2a} & -\frac{1}{2a} \\ \frac{1}{3a} & 0 \end{pmatrix}$

$$X_1 = \begin{pmatrix} 1 & 0 \\ -\frac{2}{3} & 1 \end{pmatrix} .$$

If $\rho_1 = 1$ and $\rho_2 = 2$ and the characteristic polynomial of a solution $X = (1-\lambda)(2-\lambda)$ then $T_2 = a \begin{pmatrix} 0 & 4 \\ -2 & -7 \end{pmatrix}$

and a solution $X_2 = a \begin{pmatrix} 0 & 4 \\ -2 & -7 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} -\frac{7}{8a} & -\frac{1}{2a} \\ \frac{1}{4a} & 0 \end{pmatrix}$

$$X_2 = \begin{pmatrix} 2 & 0 \\ -1\frac{3}{4} & 1 \end{pmatrix} .$$

This illustrates the fact that the equation $A_o X^2 + A_1 X + A_2 = 0$ has an infinite number of solutions when the coefficient matrices are p×q matrices such that $p < q$.

3.3 THE QUADRATIC MATRIX EQUATION.

The methods described in 3.2 for the solution of the unilateral matrix equation of degree n may obviously be applied to the unilateral quadratic matrix equation. They may also be used to solve the quadratic equation

$$XEX + DX + XF + G = 0 \qquad (3.3.1)$$

since this can be made unilateral by the substitution

$$Z = XE + D \quad .$$

The equation then becomes

$$Z^2 + Z(E^{-1}FE - D) + GE - DE^{-1}FE = 0 \quad .$$

This substitution can only be made however when E is non singular.

There are many methods which are specifically designed for the solution of the quadratic matrix equation and in particular the algebraic matrix Riccati equation where P is the symmetric matrix which is the solution of

$$PBR^{-1}B^TP - A^TP - PA - Q = 0 \quad . \qquad (3.3.2)$$

This equation has been studied widely because of its importance in control theory.

In this section five methods are described for the solution of the Quadratic Matrix equation.

Method 1.

This method for the unilateral quadratic matrix equation relies on the formation of a $2m \times 2m$ matrix and is described by Roth [1950].

Consider the equation $X^2 + A_1X + A_2 = 0$ where X, A_1 A_2 are $m \times m$ matrices.

The $2m \times 2m$ matrix R is formed where $R = \begin{pmatrix} 0 & I \\ -A_2 & -A_1 \end{pmatrix}$.

Let $T = \begin{pmatrix} X & I \\ I & 0 \end{pmatrix}$ and $T^{-1} = \begin{pmatrix} 0 & I \\ I & -X \end{pmatrix}$

then $TRT^{-1} = \begin{pmatrix} X+A_1 & -(X^2+A_1X+A_2) \\ I & -X \end{pmatrix}$

and if X is a solution of $X^2 + A_1X + A_2 = 0$

then $TRT^{-1} = \begin{pmatrix} X+A_1 & 0 \\ I & -X \end{pmatrix}$.

Hence the matrices R and $\begin{pmatrix} X+A_1 & 0 \\ I & -X \end{pmatrix}$ are similar

and $\det [R-\lambda I] = \det [(X+A_1)-\lambda I] . \det [-X-\lambda I]$

$\therefore$ $\det [R-\lambda I]$ is reducible to polynomials $f(\lambda)$ $g(\lambda)$ which are the characteristic polynomials of $(X+A_1)$ and $(-X)$ respectively.

Hence $f(X+A_1) = 0$ and $g(-X) = 0$.

Let $$f(R) = \begin{pmatrix} U & M \\ V & N \end{pmatrix}$$

then $$f(TRT^{-1}) = T\, f(R)\, T^{-1}$$

$\therefore$ $$T\, f(R)\, T^{-1} = f \begin{pmatrix} X+A_1 & 0 \\ I & -X \end{pmatrix}$$

But $$T\, f(R)\, T^{-1} = \begin{pmatrix} X & I \\ I & 0 \end{pmatrix} \begin{pmatrix} U & M \\ V & N \end{pmatrix} \begin{pmatrix} 0 & I \\ I & -X \end{pmatrix}$$

$$= \begin{pmatrix} XM+N & XU+V - (XM+N)X \\ M & U - MX \end{pmatrix}$$

and $$f \begin{pmatrix} X+A_1 & 0 \\ I & -X \end{pmatrix} = \begin{pmatrix} f(X+A_1) & 0 \\ * & f(-X) \end{pmatrix} .$$

But since $f(\lambda)$ is the characteristic polynomial of $(X+A_1)$ then

$$f(X+A_1) = 0 .$$

Hence $$\begin{pmatrix} XM+N & XU+V-(XM+N)X \\ M & U-MX \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ * & f(-X) \end{pmatrix}$$

$\therefore$ $XM+N = 0$ and $XU+V - (XM+N)X = 0$

are simultaneously satisfied whether M is singular or not and X may be obtained from

either $X = -NM^{-1}$ or $X = -VU^{-1}$.

The method therefore consists of forming a $2m \times 2m$ matrix R from the coefficient matrices of the equation. The characteristic

polynomial of R is then derived and a factor $f(\lambda)$ of degree m is chosen. The matrix $f(R)$ is then evaluated and a solution X may be obtained from this matrix. The following example is an illustration of this method.

Example 3.3.1.

Consider the equation $X^2 + A_1X + A_2 = 0$

where $A_1 = \begin{pmatrix} 1 & -2 \\ 0 & 1 \end{pmatrix}$ $A_2 = \begin{pmatrix} 0 & -2 \\ 4 & -10 \end{pmatrix}$.

The 4×4 matrix obtained is $R = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 2 & 1 & -2 \\ -4 & 10 & 0 & 1 \end{pmatrix}$.

and $\det [R-\lambda I] = \lambda^4 - 2\lambda^3 - 9\lambda^2 + 2\lambda + 8$

$= (\lambda^2-5\lambda+4)(\lambda^2+3\lambda+2)$.

Choosing the factor $(\lambda^2+3\lambda+2)$ as $f(\lambda)$

then $f(R) = \begin{pmatrix} 2 & 2 & 4 & -2 \\ -4 & 12 & 0 & 4 \\ 8 & -12 & 6 & -8 \\ -16 & 40 & -4 & 16 \end{pmatrix}$

$\therefore$ $U = \begin{pmatrix} 2 & 2 \\ -4 & 12 \end{pmatrix}$ $M = \begin{pmatrix} 4 & -2 \\ 0 & 4 \end{pmatrix}$ $V = \begin{pmatrix} 8 & -12 \\ -16 & 40 \end{pmatrix}$

$N = \begin{pmatrix} 6 & -8 \\ -4 & 16 \end{pmatrix}$

and a solution $X = -NM^{-1}$

$$\therefore \quad X = \begin{pmatrix} -6 & 8 \\ 4 & -16 \end{pmatrix} \begin{pmatrix} \frac{1}{4} & \frac{1}{8} \\ 0 & \frac{1}{4} \end{pmatrix}$$

$$\therefore \quad X = \begin{pmatrix} -1.5 & 1.25 \\ 1 & -3.5 \end{pmatrix}$$

A different solution may be obtained by choosing the factor $(\lambda^2-5\lambda+4)$ as $f(\lambda)$

then $$f(R) = \begin{pmatrix} 4 & 2 & -4 & -2 \\ -4 & 14 & 0 & -4 \\ 8 & -28 & 0 & 8 \\ 16 & -40 & -4 & 10 \end{pmatrix}$$

and $$M = \begin{pmatrix} 4 & -2 \\ 0 & 4 \end{pmatrix} \qquad N = \begin{pmatrix} 0 & 8 \\ -4 & 10 \end{pmatrix}$$

and a solution $X = -NM^{-1}$

$$\therefore \quad X = \begin{pmatrix} 0 & -8 \\ 4 & -10 \end{pmatrix} \begin{pmatrix} -\frac{1}{4} & \frac{1}{8} \\ 0 & -\frac{1}{4} \end{pmatrix}$$

$$\therefore \quad X = \begin{pmatrix} 0 & 2 \\ -1 & 3 \end{pmatrix}$$

Method 2.

An adaptation of Method 1 for the solution of the general

matrix Riccati equation (3.3.1) is described by Freested, Webber and Bass [1968].

Consider the equation $XEX + DX + XF + G = 0$ where all the matrices are m×m matrices.

The 2m×2m matrix H is formed

where $$H = \begin{pmatrix} -F & E \\ -G & D \end{pmatrix}$$

Let $$T = \begin{pmatrix} I & 0 \\ X & I \end{pmatrix} \quad \text{and} \quad T^{-1} = \begin{pmatrix} I & 0 \\ -X & I \end{pmatrix}$$

so that $$THT^{-1} = \begin{pmatrix} -F-EX & E \\ -XF-G-XEX-DX & XE+D \end{pmatrix} .$$

Then if X is a solution of (3.3.1)

$$THT^{-1} = \begin{pmatrix} -F-EX & E \\ 0 & XE+D \end{pmatrix}$$

Let $$THT^{-1} = \hat{H} = \begin{pmatrix} \hat{F} & E \\ 0 & \hat{D} \end{pmatrix} \quad \text{where} \quad \begin{aligned} \hat{F} &= -F - EX \\ \hat{D} &= XE + D \end{aligned}$$

then the matrices H and $\hat{H}$ are similar and

$$\det [\lambda I - H] = \det [\lambda I - \hat{H}] = \det [\lambda I - \hat{F}] . \det [\lambda I - \hat{D}] .$$

Let $f(\lambda) = \det [\lambda I - \hat{F}]$ and let $f(H) = \begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix}$

then since $THT^{-1} = \hat{H}$

$$Tf(H)T^{-1} = f(\hat{H}) = \begin{pmatrix} f(\hat{F}) & * \\ 0 & f(\hat{D}) \end{pmatrix} .$$

$$\text{But } T f(H)T^{-1} = \begin{pmatrix} I & 0 \\ X & I \end{pmatrix} \begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix} \begin{pmatrix} I & 0 \\ -X & I \end{pmatrix}$$

$$= \begin{pmatrix} H_{11}-H_{12}X & H_{12} \\ X(H_{11}-H_{12}X)+H_{21}-H_{22}X & XH_{12}+H_{22} \end{pmatrix}$$

$$\therefore \begin{pmatrix} H_{11}-H_{12}X & H_{12} \\ X(H_{11}-H_{12}X)+H_{21}-H_{22}X & XH_{12}+H_{22} \end{pmatrix} = \begin{pmatrix} f(\hat{F}) & * \\ 0 & f(\hat{D}) \end{pmatrix}$$

But $f(\hat{F}) = 0$ since $f(\lambda)$ is the characteristic polynomial of $\hat{F}$.

Hence $H_{11} - H_{12}X = 0$ and $X(H_{11}-H_{12}X) + H_{21} - H_{22}X = 0$

$\therefore$ X can be obtained from either $X = H_{12}^{-1} H_{11}$ or $X = H_{22}^{-1} H_{21}$,

provided one of H_{12} and H_{22} is nonsingular.

Example 3.3.2.

Consider the equation $XEX + DX + XF + G = 0$

where $E = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$ $D = \begin{pmatrix} -1 & 1 \\ 1 & 0 \end{pmatrix}$ $F = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$

$G = \begin{pmatrix} -23 & 30 \\ -20 & 26 \end{pmatrix}$.

Forming the 4×4 matrix $H = \begin{pmatrix} -1 & 1 & 1 & 0 \\ 1 & -1 & 0 & -1 \\ 23 & -30 & -1 & 1 \\ 20 & -26 & 1 & 0 \end{pmatrix}$

and $\det [\lambda I - H] = \lambda^4 + 3\lambda^3 - 48\lambda^2 - 77\lambda - 9$

$$= [\lambda^2 + 8\lambda + 1][\lambda^2 - 5\lambda - 9] \quad .$$

Choosing the factor $[\lambda^2 + 8\lambda + 1]$ as $f(\lambda)$

then $$f(H) = \begin{pmatrix} 18 & -24 & 6 & 0 \\ -14 & 21 & 0 & -7 \\ 128 & -183 & 18 & 37 \\ 137 & -192 & 27 & 28 \end{pmatrix}$$

$$\therefore \quad H_{11} = \begin{pmatrix} 18 & -24 \\ -14 & 21 \end{pmatrix} \qquad H_{12} = \begin{pmatrix} 6 & 0 \\ 0 & -7 \end{pmatrix} \qquad H_{21} = \begin{pmatrix} 128 & -183 \\ 137 & -192 \end{pmatrix}$$

$$H_{22} = \begin{pmatrix} 18 & 37 \\ 27 & 28 \end{pmatrix}$$

and a solution X may be obtained from $X = H_{12}^{-1} H_{11}$

$$\therefore \quad X = \begin{pmatrix} 6 & 0 \\ 0 & -7 \end{pmatrix}^{-1} \begin{pmatrix} 18 & -24 \\ -14 & 21 \end{pmatrix}$$

$$\therefore \quad X = \begin{pmatrix} \frac{1}{6} & 0 \\ 0 & -\frac{1}{7} \end{pmatrix} \begin{pmatrix} 18 & -24 \\ -14 & 21 \end{pmatrix}$$

$$\therefore \quad X = \begin{pmatrix} 3 & -4 \\ 2 & -3 \end{pmatrix}$$

Another solution may be obtained by choosing the factor $[\lambda^2 - 5\lambda - 9]$ for $f(\lambda)$.

Then $$f(H) = \begin{pmatrix} 21 & -37 & -7 & 0 \\ -27 & 24 & 0 & 6 \\ -171 & 207 & 21 & 24 \\ -123 & 146 & 14 & 18 \end{pmatrix}$$

$$\therefore \quad H_{11} = \begin{pmatrix} 21 & -37 \\ -27 & 24 \end{pmatrix} \qquad H_{12} = \begin{pmatrix} -7 & 0 \\ 0 & 6 \end{pmatrix}$$

and a solution $X = H_{12}^{-1} H_{11} = \begin{pmatrix} -3 & \frac{37}{7} \\ -\frac{9}{2} & 4 \end{pmatrix}$.

Method 3.

This method may be applied to find the solution of the Matrix Riccati equation. The method involves the computation of the eigenvectors of a 2m×2m matrix and the solution is then described in terms of these eigenvectors.

Consider equation (3.3.1) XEX + DX + XF + G = 0.

Potter [1966] shows that every solution X of this equation has the form $X = (\underline{b}_1\ \underline{b}_2,\ldots,\underline{b}_m)(\underline{c}_1\ \underline{c}_2,\ldots,\underline{c}_m)^{-1}$ where the column vectors $\underline{b}_i$ and $\underline{c}_i$ are the upper and lower halves of an eigenvector of the 2m×2m matrix M where

$$M = \begin{pmatrix} D & G \\ -E & -F \end{pmatrix} .$$

This can be verified quite simply as follows:

Let T be any matrix which transforms M into its Jordan Normal Form. Then the columns of T are the eigenvectors of M.

Let $T = \begin{pmatrix} T_1 & T_2 \\ T_3 & T_4 \end{pmatrix}$ and let the Jordan Normal Form of M be $\begin{pmatrix} J_1 & J_2 \\ 0 & J_3 \end{pmatrix}$ where the matrices J_1 and J_3 are upper triangular

then $T^{-1}MT = J$

or $MT = TJ$

$$\therefore \quad \begin{pmatrix} D & G \\ -E & -F \end{pmatrix} \begin{pmatrix} T_1 & T_2 \\ T_3 & T_4 \end{pmatrix} = \begin{pmatrix} T_1 & T_2 \\ T_3 & T_4 \end{pmatrix} \begin{pmatrix} J_1 & J_2 \\ 0 & J_3 \end{pmatrix}$$

$$\therefore \quad \begin{pmatrix} DT_1+GT_3 & DT_2+GT_4 \\ -ET_1-FT_3 & -ET_2-FT_4 \end{pmatrix} = \begin{pmatrix} T_1J_1 & T_1J_2+T_2J_3 \\ T_3J_1 & T_3J_2+T_4J_3 \end{pmatrix}.$$

From this we can obtain the two equations

$$DT_1 + GT_3 = T_1J_1 \qquad \ldots\ldots \quad (a)$$

$$-ET_1 - FT_3 = T_3J_1 \qquad \ldots\ldots \quad (b)$$

Multiplying equation (a) on the right by T_3^{-1} gives

$$DT_1T_3^{-1} + G = T_1J_1T_3^{-1} \qquad \ldots\ldots \quad (c)$$

Multiplying equation (b) on the left by $T_1T_3^{-1}$ and on the right by T_3^{-1} gives

$$-T_1T_3^{-1}ET_1T_3^{-1} - T_1T_3^{-1}F = T_1J_1T_3^{-1} \qquad \ldots\ldots \quad (d)$$

Subtracting equation (d) from equation (c) we obtain

$$(T_1T_3^{-1})E(T_1T_3^{-1}) + D(T_1T_3^{-1}) + (T_1T_3^{-1})F + G = 0$$

and hence $X = T_1 T_3^{-1}$ is a solution of the equation (3.3.1) and since the columns of T_1, T_3 are the upper and lower halves of eigenvectors of M the result is verified.

Potter shows that in consideration of the equation (3.3.2)

$$PBR^{-1}B^{T}P - A^{T}P - PA - Q = 0$$

then $$M = \begin{pmatrix} -A^{T} & -Q \\ -BR^{-1}B^{T} & A \end{pmatrix}$$

and if the columns of $\begin{pmatrix} T_1 \\ T_3 \end{pmatrix}$ are chosen to be the eigenvectors corresponding to the eigenvalues of M which have negative real parts then the solution $P = T_1 T_3^{-1}$ is the unique positive definite solution of (3.3.2).

Example 3.3.3.

Consider the equation $XEX + DX + XF + G = 0$

where $$E = \begin{pmatrix} 1 & -1 \\ 1 & 0 \end{pmatrix} \qquad D = \begin{pmatrix} 3 & 1 \\ -1 & 2 \end{pmatrix} \qquad F = \begin{pmatrix} 3 & -1 \\ 1 & 2 \end{pmatrix}$$

$$G = \begin{pmatrix} -14 & 9 \\ 7 & -7 \end{pmatrix}$$

then $$M = \begin{pmatrix} 3 & 1 & -14 & 9 \\ -1 & 2 & 7 & -7 \\ -1 & 1 & -3 & 1 \\ -1 & 0 & -1 & -2 \end{pmatrix}$$

and $\det [\lambda I - M] = \lambda^4 - 23\lambda^2 - 28\lambda + 165$

$$= [\lambda^2 - 7\lambda + 11][\lambda^2 + 7\lambda + 15] \quad .$$

The eigenvalues of M are $\lambda_1 = \frac{7+\sqrt{5}}{2}$ $\lambda_2 = \frac{7-\sqrt{5}}{2}$

$\lambda_3 = \frac{-7+\sqrt{11}i}{2}$ $\lambda_4 = \frac{-7-\sqrt{11}i}{2}$.

The eigenvector associated with λ_3 is $\frac{1}{6}\begin{pmatrix} 4 - 2\sqrt{11}i \\ 1 + \sqrt{11}i \\ 5 - \sqrt{11}i \\ 6 \end{pmatrix}$

The eigenvector associated with λ_4 is $\frac{1}{6}\begin{pmatrix} 4 + 2\sqrt{11}i \\ 1 - \sqrt{11}i \\ 5 + \sqrt{11}i \\ 6 \end{pmatrix}$

$$\therefore \quad \begin{pmatrix} T_1 \\ T_3 \end{pmatrix} = \frac{1}{6}\begin{pmatrix} 4-2\sqrt{11}i & 4+2\sqrt{11}i \\ 1+\sqrt{11}i & 1-\sqrt{11}i \\ 5-\sqrt{11}i & 5+\sqrt{11}i \\ 6 & 6 \end{pmatrix}$$

and $X = T_1 T_3^{-1}$ $\therefore$ $X = \begin{pmatrix} 4 - 2\sqrt{11}i & 4+2\sqrt{11}i \\ 1+\sqrt{11}i & 1-\sqrt{11}i \end{pmatrix}\begin{pmatrix} 5-\sqrt{11}i & 5+\sqrt{11}i \\ 6 & 6 \end{pmatrix}$

$\therefore$ $X = \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}$ is the unique positive definite

solution.

Method 4.

This method is a variant of the eigenvector approach used in Method 3. The use of eigenvectors is often highly unsatisfactory from a numerical point of view and Laub [1979] describes a method which uses the Schur Canonical Form of a matrix and expresses the solution in terms of Schur vectors.

The following results and definitions are used:

(1) A matrix is orthogonal if $A^T = A^{-1}$.

(2) Let A be a matrix with eigenvalues $\lambda_1,\lambda_2,\ldots,\lambda_n$.

Then there exists an orthogonal matrix U such that $\tilde{A} = U^TAU$ where $\tilde{A}$ is upper triangular with diagonal elements $\lambda_1,\lambda_2,\ldots,\lambda_n$. $\tilde{A}$ is said to be the Schur Canonical Form of A.

The method is applied to the Matrix Riccati equation

$$PBR^{-1}B^TP - A^TP - PA - Q = 0 .$$

The matrix Z is formed

where $$Z = \begin{pmatrix} A & -BR^{-1}B^T \\ -Q & -A^T \end{pmatrix} .$$

Let S be the Schur Canonical Form of Z

then $Z = USU^T$ where U is orthogonal and S is upper triangular.

Since U is orthogonal $U^T = U^{-1}$

and hence $$ZU = US .$$

Let $$U = \begin{pmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{pmatrix} \quad \text{and} \quad S = \begin{pmatrix} S_{11} & S_{12} \\ 0 & S_{22} \end{pmatrix}$$

then $$ZU = \begin{pmatrix} AU_{11}-BR^{-1}B^TU_{21} & AU_{12}-BR^{-1}B^TU_{22} \\ -QU_{11}-A^TU_{21} & -QU_{12}-A^TU_{22} \end{pmatrix}$$

and $$US = \begin{pmatrix} U_{11}S_{11} & U_{11}S_{12}+U_{12}S_{22} \\ U_{21}S_{11} & U_{21}S_{12}+U_{22}S_{22} \end{pmatrix}$$

and setting ZU = US we can obtain the two matrix equations

(1) $$AU_{11} - BR^{-1}B^TU_{21} = U_{11}S_{11}$$

(2) $$-QU_{11} - A^TU_{21} = U_{21}S_{11} \quad .$$

Multiplying equation (1) on the left by U_{11}^{-1} gives

$$U_{11}^{-1}AU_{11} - U_{11}^{-1}BR^{-1}B^TU_{21} = S_{11} \quad .$$

Multiplying equation (2) on the left by U_{21}^{-1} gives

$$-U_{21}^{-1}QU_{11} - U_{21}^{-1}A^TU_{21} = S_{11} \quad .$$

Hence $$U_{11}^{-1}AU_{11} - U_{11}^{-1}BR^{-1}B^TU_{21} = -U_{21}^{-1}QU_{11} - U_{21}^{-1}A^TU_{21}$$

$$\therefore \quad U_{21}U_{11}^{-1}AU_{11} - U_{21}U_{11}^{-1}BR^{-1}B^TU_{21} = -QU_{11} - A^TU_{21}$$

$$\therefore \quad U_{21}U_{11}^{-1}A - U_{21}U_{11}^{-1}BR^{-1}B^TU_{21}U_{11}^{-1} = - Q - A^TU_{21}U_{11}^{-1}$$

$$\therefore \quad (U_{21}U_{11}^{-1})BR^{-1}B^T(U_{21}U_{11}^{-1}) - A^T(U_{21}U_{11}^{-1}) - (U_{21}U_{11}^{-1})A - Q = 0$$

$$\therefore \quad P = U_{21}U_{11}^{-1} \text{ is a solution of } PBR^{-1}B^TP - A^TP - PA - Q = 0.$$

Laub states that the Schur Vector approach is not designed for hand computation but he gives the following example to illustrate the method.

Example 3.3.4.

Consider the equation $PBR^{-1}B^TP - A^TP - PA - Q = 0$

where $A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ $B = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ $R = 1$ $Q = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$

then $Z = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ -1 & 0 & 0 & 0 \\ 0 & -2 & -1 & 0 \end{pmatrix}$

and $Z = USU^T$

where $S = \begin{pmatrix} -1 & 0 & 1 & -\frac{1}{2} \\ 0 & -1 & -1 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$

and U is the orthogonal matrix

$$U = \begin{pmatrix} \frac{1}{2} & -\frac{\sqrt{5}}{10} & -\frac{3\sqrt{5}}{10} & \frac{1}{2} \\ -\frac{1}{2} & -\frac{\sqrt{5}}{10} & -\frac{3\sqrt{5}}{10} & -\frac{1}{2} \\ \frac{1}{2} & -\frac{3\sqrt{5}}{10} & \frac{\sqrt{5}}{10} & -\frac{1}{2} \\ -\frac{1}{2} & -\frac{3\sqrt{5}}{10} & \frac{\sqrt{5}}{10} & \frac{1}{2} \end{pmatrix}$$

$$\therefore \qquad U_{11} = \begin{pmatrix} \frac{1}{2} & -\frac{\sqrt{5}}{10} \\ -\frac{1}{2} & -\frac{\sqrt{5}}{10} \end{pmatrix} \qquad U_{21} = \begin{pmatrix} \frac{1}{2} & -\frac{3\sqrt{5}}{10} \\ -\frac{1}{2} & -\frac{3\sqrt{5}}{10} \end{pmatrix}$$

and the unique positive definite solution of (3.3.2) is given by

$$P = U_{21}U_{11}^{-1}$$

$$\therefore \qquad P = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} .$$

Method 5.

This method uses the square root of a matrix in order to find the unique positive definite solution of the matrix Riccati equation. The application of the method, described by Incertis [1983], involves the computation of the square root of a matrix and the solution of a linear matrix Liapunov equation.

The equation to be considered is

$$PBR^{-1}B^{T}P - A^{T}P - PA - Q = 0 \quad .$$

Rearranging and multiplying on the left and right by P^{-1} the equation becomes

$$P^{-1}QP^{-1} + P^{-1}A^{T} + AP^{-1} - BR^{-1}B^{T} = 0 \ .$$

Multiplying on the left by Q gives

$$QP^{-1}QP^{-1} + QP^{-1}A^{T} + QAP^{-1} - QBR^{-1}B^{T} = 0$$

which is equivalent to

$$\left[QP^{-1}+A\right]\left[QP^{-1}+A^{T}\right] - \left[AQ-QA\right]P^{-1} - AA^{T} - QBR^{-1}B^{T} = 0 \quad .$$

Let S and D be the symmetric and skew symmetric components of the matrix A

$$\text{i.e. } S = \frac{A+A^T}{2} \qquad D = \frac{A-A^T}{2}$$

then substituting $A = S + D$ and $A^T = S - D$ the equation becomes

$$(QP^{-1}+S+D)(QP^{-1}+S-D) - (AQ-QA)P^{-1} = AA^T + QBR^{-1}B^T$$

$$\therefore \quad (QP^{-1}+S)^2 - (QP^{-1}+S)D + D(QP^{-1}+S) - (AQ-QA)P^{-1} = AA^T+QBR^{-1}B^T+D^2 \quad .$$

Now making the assumption that there exists a matrix T such that

$$AQ - QA = 2TQ$$

and letting $H = QP^{-1} + S$

the equation becomes

$$H^2 - HD + (D-2T)H = QBR^{-1}B^T + AA^T + D^2 - 2TS$$

and if the positive definite solution is being sought we must have the condition that $H - S \geq 0$.

Let $W = H - T$

then substituting $H = W + T$ in the equation gives

$$W^2 + (D-T)W - W(D-T) = QBR^{-1}B^T + (A-2T)A^T + (D-T)^2$$

or $$W^2 + \Phi W - W\Phi = F$$

where $$\Phi = D - T$$

and $$F = QBR^{-1}B^T + (A-2T)A^T + \Phi^2 \quad .$$

Now if F is a positive definite matrix then a positive definite square root exists and is unique.

Let $W = F^{\frac{1}{2}} + U$

then substituting for W we obtain

$$U^2 + (F^{\frac{1}{2}}+\Phi)U + U(F^{\frac{1}{2}}-\Phi) = F^{\frac{1}{2}}\Phi - \Phi F^{\frac{1}{2}}$$

and the original quadratic problem has been reduced to finding a matrix U which satisfies this equation and the condition $H - S \geq 0$ becomes

$$F^{\frac{1}{2}} + T - S + U \geq 0 \quad .$$

The technique which now follows depends upon whether the matrix $F^{\frac{1}{2}}$ commutes with Φ. If $F^{\frac{1}{2}}\Phi - \Phi F^{\frac{1}{2}} = 0$ then it can be proved that the unique solution U which fulfils the positive definiteness condition is the null matrix $U = 0$. In this case $W = F^{\frac{1}{2}}$ and substituting the definitions of H, W, Φ into the intermediate equation $H = QP^{-1} + S$ we obtain

$$QP^{-1} = F^{\frac{1}{2}} - A^T - \Phi \quad .$$

Also, multiplying the rearrangement of the Riccati equation by 2 we obtain

$$2P^{-1}QP^{-1} + 2P^{-1}A^T + 2AP^{-1} = 2BR^{-1}B^T$$

which factorizes as

$$[2A+P^{-1}Q]P^{-1} + P^{-1}[2A^T+QP^{-1}] = 2BR^{-1}B^T \quad .$$

This can be written as the Liapunov equation

$$A_\alpha^T P^{-1} + P^{-1}A_\alpha = 2BR^{-1}B^T$$

where $$A_\alpha = 2A^T + QP^{-1}$$

$$= F^{\frac{1}{2}} + A^T - \Phi \quad .$$

Hence the unique positive definite solution of (3.3.2) is given by $P = Y^{-1}$ where Y is the solution of

$$A_\alpha^T Y + YA_\alpha = 2BR^{-1}B^T \quad .$$

The technique is not quite so straightforward when the commutativity condition $F^{\frac{1}{2}}\Phi = \Phi F^{\frac{1}{2}}$ is not satisfied. In this case there is no trivial solution U for the quadratic equation in U. When this happens the matrix W is obtained by forming an iterative procedure from the intermediate equation

$$W^2 + \Phi W - W\Phi = F.$$

This is $$W_{k+1} = \left[F + W_k\Phi - \Phi W_k\right]^{\frac{1}{2}} \quad \text{for } k = 0,1,2,....$$

and taking $W_o = F^{\frac{1}{2}}$ it can be shown that a sufficient convergence condition of the sequence $\{W_k\}$ is that $F + \Phi^2$ is to be a positive definite matrix.

Having obtained the matrix W then substituting in the intermediate equation $H = QP^{-1} + S$ we obtain

$$QP^{-1} = W - \Phi - A^T$$

and hence $$A_\alpha = 2A^T + QP^{-1}$$

becomes $$A_\alpha = W + A^T - \Phi \quad .$$

The solution P is then obtained by solving the Liapunov equation

$$A_\alpha^T Y + YA_\alpha = 2BR^{-1}B^T$$

and finding $P = Y^{-1}$ as before.

The method is illustrated in the following example given by Incertis. In this case the commutativity condition $F^{\frac{1}{2}}\Phi = \Phi F^{\frac{1}{2}}$ is satisfied.

Example 3.3.5.

Consider the equation $PBR^{-1}B^TP - A^TP - PA - Q = 0$

where $A = \begin{pmatrix} -1 & 0 \\ 1 & 2 \end{pmatrix}$ $B = \frac{1}{\sqrt{11}} \begin{pmatrix} \sqrt{7} & 9 \\ 0 & 11 \end{pmatrix}$ $R = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

$$Q = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} .$$

Following the definitions for S and D we obtain

$$S = \begin{pmatrix} -1 & 0.5 \\ 0.5 & 2 \end{pmatrix} \qquad D = \begin{pmatrix} 0 & -0.5 \\ 0.5 & 0 \end{pmatrix} .$$

We now seek to find a matrix T such that

$$AQ - QA = 2TQ$$

this gives $TQ = \begin{pmatrix} 0 & 0 \\ 0.5 & 0 \end{pmatrix}$

and since Q is singular T can be any matrix of the form

$$T = \begin{pmatrix} 0 & a \\ 0.5 & b \end{pmatrix} .$$

Since $\Phi = D - T$ computation is simplified if we choose

$$T = \begin{pmatrix} 0 & -0.5 \\ 0.5 & 0 \end{pmatrix}$$

so that $D = T$ and hence $\Phi = 0$.

It then follows that since $F^{\frac{1}{2}}\Phi - \Phi F^{\frac{1}{2}} = 0$ then $U = 0$.

Hence $A_\alpha = F^{\frac{1}{2}} + A^T$

where $F = QBR^{-1}B^T + (A-2T)A^T$

$\therefore$ $F = \begin{pmatrix} 9 & 10 \\ 0 & 4 \end{pmatrix}$ and the unique positive definite

square root of F is $F^{\frac{1}{2}} = \begin{pmatrix} 3 & 2 \\ 0 & 2 \end{pmatrix}$.

Hence $A_\alpha = \begin{pmatrix} 3 & 2 \\ 0 & 2 \end{pmatrix} + \begin{pmatrix} -1 & 1 \\ 0 & 2 \end{pmatrix} = \begin{pmatrix} 2 & 3 \\ 0 & 4 \end{pmatrix}$

$\therefore \quad A_\alpha^T Y + Y A_\alpha = 2BR^{-1}B^T$

becomes $\begin{pmatrix} 2 & 0 \\ 3 & 4 \end{pmatrix} Y + Y \begin{pmatrix} 2 & 3 \\ 0 & 4 \end{pmatrix} = \begin{pmatrix} 16 & 18 \\ 18 & 22 \end{pmatrix}$

$\therefore \quad Y = \begin{pmatrix} 4 & 1 \\ 1 & 2 \end{pmatrix}$

and hence the positive definite solution of the original Riccati equation is

$$P = \begin{pmatrix} 4 & 1 \\ 1 & 2 \end{pmatrix}^{-1}$$

$\therefore \quad P = \frac{1}{7} \begin{pmatrix} 2 & -1 \\ -1 & 4 \end{pmatrix}$.

3.4 CONCLUSION.

In this chapter a selection of methods of solution produced over the last 50 years has been considered. It is interesting to note that many methods make use of the fact that every characteristic root of a solution X of the unilateral equation (3.2.1) is also a root of det $[A(\lambda)] = 0$ where $A(\lambda) = \sum_{o}^{n} A_i \lambda^{n-i}$.

This fact was established in the last century and shows how recent developments are built on the foundations established long ago.

With the development of computer techniques, efforts have concentrated on methods which can be adapted to computer methods of solution rather than purely algebraic techniques.

In Section 3.2 Methods 1 and 2 require the solution of $\det [A(\lambda)] = 0$. Dennis, Traub & Weber [1976] have shown that if $A_o = I$ the required roots can be obtained by forming the block Companion Matrix C

where
$$C = \begin{pmatrix} 0 & 0 \ldots\ldots & -A_n \\ I & 0 \ldots\ldots & -A_{n-1} \\ \ldots\ldots\ldots\ldots & & \\ 0 & 0 \ldots I & -A_1 \end{pmatrix}$$

and the characteristic roots of C are equivalent to the roots of $\det [A(\lambda)] = 0$. Hence computer methods for determining characteristic roots may be applied.

Method 3 in this section describes an algebraic method of solution. Since it was written in 1941 it is presumably intended to be applied without computer assistance, but even in the 2×2 case it involves practical problems of computation as can be seen from the worked example.

Method 4 is interesting as it deals with the case where the coefficient matrices are not square. Obtaining the general solution, however, in the cases where X has arbitrary characteristic roots would be difficult for matrices of large order.

The methods in Section 3.3 are mainly concerned with the solution of the matrix Riccati equation. Methods 1 and 2 share the disadvantage of requiring the characteristic polynomial of a $2m \times 2m$ matrix to be factorized into 2 polynomials of degree m. This would involve problems in computation for matrices of order greater than 3×3.

Methods 3 and 4 are most suitable for application of computer methods since the solutions are expressed in terms of the eigenvectors or Schur vectors of a $2m \times 2m$ matrix.

Method 5 uses a completely different approach in that no $2m \times 2m$ matrix is introduced. The ease of computation displayed in the example given by Incertis is illusory however since this example is one of a small class of equations for which the commutativity condition is fulfilled. When it is not fulfilled, the iterative procedure given involves the computation of the square root of a matrix at each stage. This would be difficult if the matrices were of large order.

All the methods described have certain disadvantages with any purely algebraic methods suffering from problems of computation with large order matrices.

CHAPTER 4

The Solution of Matrix Equations Using the Characteristic Equation and Elimination Methods

4.1 THE SCALAR EQUATION.

The general equation $f(X) = 0$ where $f(\lambda)$ is a polynomial function of λ is known as a scalar equation.

The solutions of a scalar matrix equation are divided into sets of similar matrices. If $X = TJT^{-1}$ where J is the Jordan Normal Form of X then

$$f(X) = f[TJT^{-1}] = T\, f(J)T^{-1} \quad .$$

Hence if $f(X) = 0$

then $f(J) = 0$ since T is non-singular.

For example, if $f(X) = a_oX^n + a_1X^{n-1} + \ldots . a_nI$ where the a_i are scalars, then $f[TJT^{-1}] = T[a_oJ^n + a_1J^{n-1} + \ldots . a_nI]T^{-1}$

$$= T\, f(J)T^{-1} \quad .$$

Hence to find solutions of $f(X) = 0$ it is sufficient to find solutions of $f(J) = 0$ and X is then any matrix such that $X = TJT^{-1}$.

Due to the special form of the Jordan matrix then f(J) also has a special form.

Let $J = \begin{pmatrix} J_1 & 0 & \ldots . & 0 \\ 0 & J_2 & \ldots . & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots . & J_n \end{pmatrix}$ where the J_i are Jordan Blocks.

then $$f(J) = \begin{pmatrix} f(J_1) & 0 & \cdots & 0 \\ 0 & f(J_2) & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & & f(J_n) \end{pmatrix}$$

$\therefore$ If $f(J) = 0$ then $f(J_i) = 0 \quad i = 1,2,\ldots,n$.

If J_i is the Jordan Block $\begin{pmatrix} \lambda_i & 1 & 0 & \cdots & 0 \\ 0 & \lambda_i & 1 & \cdots & 0 \\ 0 & 0 & \lambda_i & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & & \lambda_i \end{pmatrix}$

then provided the derivatives $f^j(\lambda_i)$ $j = 1,2,\ldots,(n-1)$ exist

for all λ_i

$$f(J_i) = \begin{pmatrix} f(\lambda_i) & f'(\lambda_i) & \frac{f''(\lambda_i)}{2!} & \cdots & \frac{f^{n-1}(\lambda_i)}{(n-1)!} \\ 0 & f(\lambda_i) & f'(\lambda_i) & \cdots & \\ 0 & 0 & f(\lambda_i) & \cdots & \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & f(\lambda_i) \end{pmatrix}$$

then $f(J_i) = 0 \Leftrightarrow f(\lambda_i) = 0,\ f'(\lambda_i) = 0 \ldots\ f^{n-1}(\lambda_i) = 0.$

Example 4.1.1.

Consider the equation $f(X) = 0$ where $f(X) = X^2-3X+2I$ and X is a 2×2 matrix.

Then $f(\lambda) = \lambda^2-3\lambda+2$

$f'(\lambda) = 2\lambda-3$

$\therefore$ there is no λ for which $f(\lambda) = 0$ and $f'(\lambda) = 0$.

Hence the solutions of the matrix equation $f(X) = 0$ are

$$X = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad X = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} \quad \text{and} \quad X = T \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} T^{-1}$$

where T is an arbitrary non-singular matrix.

Example 4.1.2.

Consider the equation $f(X) = 0$ where $f(X) = X^2 - 2aX + a^2 I$

then $f(\lambda) = \lambda^2 - 2a\lambda + a^2$

$f'(\lambda) = 2\lambda - 2a$.

Hence $f(\lambda) = 0$ and $f'(\lambda) = 0$ both have the root $\lambda = a$.

Solutions of the matrix equation are therefore

$$X = \begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix} \quad \text{and} \quad X = T \begin{pmatrix} a & 1 \\ 0 & a \end{pmatrix} T^{-1} \quad .$$

Example 4.1.3.

Consider the equation $f(X) = 0$ where $f(X) = X^2 - 3X + 2I$ and X is a 3×3 matrix

then $f(\lambda) = \lambda^2 - 3\lambda + 2$

$f'(\lambda) = 2\lambda - 3$

$f''(\lambda) = 2$

$\therefore$ $f(\lambda) = 0$ has roots $\lambda_1 = 1$ and $\lambda_2 = 2$

$f'(\lambda) = 0$ has the root $\lambda = \frac{3}{2}$.

Since $f(\lambda) = 0$ and $f'(\lambda) = 0$ have no common solutions the only matrix solutions are

$$X = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \qquad X = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix} \qquad X = T \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix} T^{-1}$$

$$X = T \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix} T^{-1}$$

Example 4.1.4.

Consider the equation $f(X) = 0$ where $f(X) = X^2-2aX+a^2I$

then $f(\lambda) = \lambda^2-2a\lambda+a^2$

$f'(\lambda) = 2\lambda-2a$

$f''(\lambda) = 2$

$\therefore$ $f(\lambda) = 0$ has the single root $\lambda = a$

$f'(\lambda) = 0$ has the root $\lambda = a$

$f''(\lambda) = 0$ has no roots .

Since $f(\lambda) = 0$ and $f'(\lambda) = 0$ have a common solution, the solutions of the matrix equation are

$$X = \begin{pmatrix} a & 0 & 0 \\ 0 & a & 0 \\ 0 & 0 & a \end{pmatrix} \quad \text{or} \quad X = T \begin{pmatrix} a & 1 & 0 \\ 0 & a & 0 \\ 0 & 0 & a \end{pmatrix} T^{-1}$$

A very important scalar equation connected with any square matrix X is its characteristic equation $\det [\lambda I-X] = 0$.

The characteristic polynomial is an example of an annihilating

polynomial of X since X satisfies the equation $\det[\lambda I - X] = 0$.

The solution of a scalar matrix equation $f(X) = 0$ has been shown to be quite straightforward. The corresponding scalar equation $f(\lambda) = 0$ is solved. If the roots of $f(\lambda) = 0$ are all distinct, $\lambda_1 \lambda_2, \ldots, \lambda_n$ then solutions of the matrix equation $f(X) = 0$ can be written as

$$X = T \begin{pmatrix} \lambda_1 & 0 & 0 & \ldots & 0 \\ 0 & \lambda_2 & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & & \lambda_n \end{pmatrix} T^{-1}$$

If $f(\lambda) = 0$ has a repeated factor $(\lambda - a)$ then

$$X = T \begin{pmatrix} J_1 & 0 & \ldots & 0 \\ 0 & J_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & J_r \end{pmatrix}$$

where the J_i are Jordan Blocks where at least one is of the form

$$\begin{pmatrix} a & 1 & 0 & \ldots & 0 \\ 0 & a & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & & a \end{pmatrix}.$$

The solutions of a scalar matrix equation can therefore be divided into sets of similar matrices.

4.2 OBTAINING THE CHARACTERISTIC EQUATION FROM $\det [A_o \lambda^n + A_1 \lambda^{n-1} + \ldots A_n]$.

Let X be a solution of the unilateral equation

$$A_o X^n + A_1 X^{n-1} + \ldots A_n = 0 \qquad (4.2.1)$$

and let $A(\lambda) = A_o\lambda^n + A_1\lambda^{n-1} + \ldots. A_{n-1}\lambda + A_n$.

As shown in 3.2 several methods of solution make use of the fact that every characteristic root of a solution X is also a root of det $[A(\lambda)] = 0$ and hence the characteristic polynomial of a solution X is a factor of det $[A(\lambda)]$.

This gives a method of determining possible characteristic polynomials of a solution X of the unilateral equation. By forming the determinant of $A(\lambda)$ and factorizing into irreducible factors, then if X is an $m \times m$ matrix, any factor of degree m is a possible characteristic polynomial for a solution X.

Sylvester [1884] suggested that solutions of the unilateral matrix equation could be obtained by choosing a factor $\phi(\lambda)$ of degree m from det $[A(\lambda)]$ and by combining $\phi(X) = 0$ with equation (4.2.1), higher powers of X could be eliminated until a linear equation in X is obtained, from which the solution could be found. He does not appear however, to have given details of how this might be carried out.

The elimination methods used for a system of polynomial equations described in 2.4 may be applied to the two equations $\phi(X) = 0$ and (4.2.1) to obtain the linear equation in X.

The method is illustrated in the following examples.

Example 4.2.1.

Consider the equation $A_oX^2 + A_1X + A_2 = 0$

where $A_o = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ $A_1 = \begin{pmatrix} -2 & 1 \\ 1 & 0 \end{pmatrix}$ $A_2 = \begin{pmatrix} 1 & -1 \\ 1 & -1 \end{pmatrix}$

then $A(\lambda) = A_o\lambda^2 + A_1\lambda + A_2 = \begin{pmatrix} \lambda^2-2\lambda+1 & \lambda-1 \\ \lambda+1 & \lambda^2-1 \end{pmatrix}$

and $\det [A(\lambda)] = \lambda^4-2\lambda^3-\lambda^2+2\lambda$

$= \lambda(\lambda-1)(\lambda+1)(\lambda-2)$.

Since X is a 2×2 matrix its characteristic polynomial is of degree 2 and since it must be a factor of det [A(λ)] there are six possible characteristic polynomials for X

$\phi_1(\lambda) = \lambda(\lambda-1)$ $\phi_3(\lambda) = \lambda(\lambda-2)$ $\phi_5(\lambda) = (\lambda-1)(\lambda-2)$

$\phi_2(\lambda) = \lambda(\lambda+1)$ $\phi_4(\lambda) = (\lambda-1)(\lambda+1)$ $\phi_6(\lambda) = (\lambda+1)(\lambda-2)$.

Taking $\phi_1(\lambda)$ as the characteristic polynomial of a solution means that X satisfies the two equations

(1) $X^2 - X = 0$

(2) $A_oX^2 + A_1X + A_2 = 0$.

Multiplying (1) on the left by A_o and subtracting from (2) gives

$[A_1+A_o]X + A_2 = 0$

$\therefore \quad X = [A_1+A_o]^{-1}[-A_2]$

$\therefore \quad X = \begin{pmatrix} -1 & 1 \\ 1 & 1 \end{pmatrix}^{-1} \begin{pmatrix} -1 & 1 \\ -1 & 1 \end{pmatrix}$

$\therefore \quad X_1 = \begin{pmatrix} 0 & 0 \\ -1 & 1 \end{pmatrix}$

Taking $\phi_2(\lambda)$ as the characteristic polynomial of a solution means that X satisfies the two equations

(1) $X^2 + X = 0$

(2) $A_o X^2 + A_1 X + A_2 = 0$.

Eliminating X^2 as before gives

$$[A_1 - A_o]X + A_2 = 0$$

$$X = [A_1 - A_o]^{-1}[-A_2]$$

$$\therefore \quad X_2 = \begin{pmatrix} 1 & -1 \\ 2 & -2 \end{pmatrix} .$$

The other four solutions may be found in a similar manner.

Taking $\phi_3(\lambda) = \lambda(\lambda-2)$ as the characteristic polynomial

gives $X_3 = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$.

Taking $\phi_4(\lambda) = (\lambda-1)(\lambda+1)$ as the characteristic polynomial

gives $X_4 = \begin{pmatrix} -1 & 0 \\ -4 & 1 \end{pmatrix}$.

Taking $\phi_5(\lambda) = (\lambda-1)(\lambda-2)$ as the characteristic polynomial

gives $X_5 = \begin{pmatrix} 2 & 0 \\ -1 & 1 \end{pmatrix}$.

Taking $\phi_6(\lambda) = (\lambda+1)(\lambda-2)$ as the characteristic polynomial

gives $X_6 = \begin{pmatrix} 1 & -1 \\ -2 & 0 \end{pmatrix}$.

There are therefore 6 solutions for this matrix equation. The result generalizes as follows. If the matrices involved in the equation (4.2.1) are m×m then the degree of the polynomial det [A(λ)] is 2m. Since the characteristic polynomial of X is of degree m, there are ${}^{2m}C_m$ possible choices of factor and hence ${}^{2m}C_m$ possible solutions for the matrix equation.

Example 4.2.2.

This example illustrates the method when applied to the unilateral quadratic equation involving 3×3 matrices.

Consider the equation $X^2 + A_1X + A_2 = 0$

where $$A_1 = \begin{pmatrix} 0 & 1 & -1 \\ 1 & 0 & 0 \\ 2 & 0 & -1 \end{pmatrix} \qquad A_2 = \begin{pmatrix} -3 & 2 & 1 \\ -1 & 1 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

then $$A(\lambda) = \begin{pmatrix} \lambda^2-3 & \lambda+2 & -\lambda+1 \\ \lambda-1 & \lambda^2+1 & 0 \\ 2\lambda & 1 & \lambda^2-\lambda \end{pmatrix}$$

and $$\det [A(\lambda)] = (\lambda-1)^3(\lambda+1)(\lambda^2+\lambda+1) \quad .$$

Since X is a 3×3 matrix its characteristic polynomial will be of degree 3. As there is a repeated factor in det [A(λ)] and $(\lambda^2+\lambda+1)$ is irreducible over the real numbers there are only 4 possible choices for the characteristic polynomial of X.

$$\phi_1(\lambda) = \lambda^3 - 3\lambda^2 + 3\lambda - 1$$

$$\phi_2(\lambda) = \lambda^3 - \lambda^2 - \lambda + 1$$

95

$$\phi_3(\lambda) = \lambda^3 + 2\lambda^2 + 2\lambda + 1$$

$$\phi_4(\lambda) = \lambda^3 - 1 \quad .$$

Taking $\phi_1(\lambda)$ as the characteristic polynomial of a solution X means that X satisfies the two equations

(1) $X^3 - 3X^2 + 3X - I = 0$

(2) $X^2 + A_1X + A_2 = 0$

Multiplying equation (2) on the right by X and subtracting equation (1) from it gives

(3) $[A_1+3I]X^2 + [A_2-3I]X + I = 0$.

Multiplying equation (2) on the left by $[A_1+3I]$ and subtracting equation (3) from it gives

$$[A_1^2+3A_1-A_2+3I]X + A_1A_2 + 3A_2 - I = 0$$

$$\therefore \quad \begin{pmatrix} 5 & 1 & -3 \\ 4 & 3 & -1 \\ 4 & 1 & -1 \end{pmatrix} X + \begin{pmatrix} -11 & 6 & 3 \\ -6 & 4 & 1 \\ -6 & 6 & 1 \end{pmatrix} = 0$$

$$\therefore \quad X_1 = \begin{pmatrix} 1 & -2 & 0 \\ 0 & 1 & 0 \\ -2 & -1 & 1 \end{pmatrix}$$

Taking $\phi_2(\lambda)$ as the characteristic polynomial of a solution X means that X satisfies the two equations

(1) $X^3 - X^2 - X + I = 0$

(2) $X^2 + A_1X + A_2 = 0$

Eliminating X^3 gives

$$(3) \qquad [A_1+I]X^2 + [A_2+I]X - I = 0$$

Eliminating X^2 gives

$$(4) \qquad [A_1^2+A_1-A_2-I]X + A_1A_2 + A_2 + I = 0$$

$$\therefore \qquad \begin{pmatrix} 1 & -1 & -1 \\ 2 & -1 & -1 \\ 0 & 1 & -3 \end{pmatrix} X + \begin{pmatrix} -3 & 2 & 1 \\ -4 & 4 & 1 \\ -6 & 4 & 3 \end{pmatrix} = 0$$

$$\therefore \qquad X_2 = \begin{pmatrix} 1 & -2 & 0 \\ 0 & -1 & 0 \\ -2 & 1 & 1 \end{pmatrix}$$

Taking $\phi_3(\lambda) = \lambda^3 + 2\lambda^2 + 2\lambda + 1$ as the characteristic polynomial and applying elimination methods leads to

$$X_3 = \begin{pmatrix} -2 & 1 & 0 \\ -\frac{9}{7} & \frac{5}{7} & -\frac{6}{7} \\ -\frac{11}{7} & \frac{10}{7} & -\frac{5}{7} \end{pmatrix} .$$

Taking $\phi_4(\lambda) = \lambda^3 - 1$ as the characteristic polynomial leads to

$$X_4 = \begin{pmatrix} -2 & 1 & 0 \\ -3 & 1 & 0 \\ -5 & 2 & 1 \end{pmatrix}$$

Example 4.2.3.

This example shows how the elimination method may be applied to a unilateral cubic matrix equation involving 2×2 matrices.

Consider the equation $X^3 + A_1X^2 + A_2X + A_3 = 0$

where $A_1 = \begin{pmatrix} -6 & 6 \\ -3 & -15 \end{pmatrix}$ $A_2 = \begin{pmatrix} 2 & -42 \\ 21 & 65 \end{pmatrix}$ $A_3 = \begin{pmatrix} 18 & 66 \\ -33 & -81 \end{pmatrix}$

then $A(\lambda) = \begin{pmatrix} \lambda^3-6\lambda^2+2\lambda+18 & 6\lambda^2-42\lambda+66 \\ -3\lambda^2+21\lambda-33 & \lambda^3-15\lambda^2+65\lambda-81 \end{pmatrix}$

and $\det [A(\lambda)] = \lambda^6 - 21\lambda^5 + 175\lambda^4 - 735\lambda^3 + 1624\lambda^2 - 1764\lambda + 720$

$= (\lambda-1)(\lambda-2)(\lambda-3)(\lambda-4)(\lambda-5)(\lambda-6)$.

Since X is a 2×2 matrix, its characteristic polynomial will be of degree 2. Hence any quadratic factor of $\det [A(\lambda)]$ is a possible characteristic polynomial of a solution X and since $\det [A(\lambda)]$ has six linear factors the maximum possible number of solutions is ${}^6C_2 = 15$.

Choosing $\phi_1(\lambda) = (\lambda-1)(\lambda-2) = \lambda^2-3\lambda+2$ as a characteristic polynomial

then X satisfies the two equations

(1) $X^3 + A_1X^2 + A_2X + A_3 = 0$

(2) $X^2 - 3X + 2I = 0$.

Multiplying equation (2) on the right by X and subtracting from equation (1) gives

(3) $[A_1+3I]X^2 + [A_2-2I]X + A_3 = 0$.

Multiplying equation (2) on the left by $[A_1+3I]$ and subtracting from equation (3) gives

$$[A_2+3A_1+7I]X + [A_3-2A_1-6I] = 0$$

$$\therefore \quad \begin{pmatrix} -9 & -24 \\ 12 & 27 \end{pmatrix} X + \begin{pmatrix} 24 & 54 \\ -27 & -57 \end{pmatrix} = 0$$

$$\therefore \quad X_1 = \begin{pmatrix} 0 & -2 \\ 1 & 3 \end{pmatrix} .$$

Other choices of quadratic factors from det $[A(\lambda)]$ lead to other solutions.

4.3 SOLUTION OF THE UNILATERAL MATRIX QUADRATIC EQUATION FOR m×m MATRICES.

The elimination method for the unilateral quadratic matrix equation may be extended for matrices of any order. The first step is to make the equation $A_oX^2 + A_1X + A_2 = 0$ monic by multiplying on the left by A_o^{-1}. The method clearly cannot be applied if A_o is singular.

The determinant of the matrix $[\lambda^2 I+A_1\lambda+A_2]$ is formed. If X is a square matrix of order m this determinant will be a polynomial of degree 2m. Factors of degree m are then chosen as possible characteristic polynomials of X. For each choice, a possible solution may be sought by elimination methods. If the number of solutions is finite it will be at most $^{2m}C_m$.

2×2 Case.

Let the factor of det $[A(\lambda)]$ chosen for the possible characteristic polynomial be $[\lambda^2+a_1\lambda+a_2]$.

then X satisfies the two equations

$$X^2 + a_1X + a_2I = 0$$

and $$X^2 + A_1X + A_2 = 0$$

then $X = - J_1^{-1} K_1$ where $J_1 = a_1I - A_1$

$$K_1 = a_2I - A_2 \quad .$$

3×3 Case.

Let the possible characteristic equation be

$$\lambda^3 + a_1\lambda^2 + a_2\lambda + a_3 = 0$$

then X satisfies

$$X^3 + a_1X^2 + a_2X + a_3I = 0$$

and $$X^2 + A_1X + A_2 = 0 \quad .$$

By elimination $X = - J_2^{-1} K_2$ where $J_1 = a_1I - A_1$

$$J_2 = K_1 - J_1A_1$$

$$K_1 = a_2I - A_2$$

$$K_2 = a_3I - J_1A_2 \quad .$$

4×4 Case.

Let the possible characteristic equation be

$$\lambda^4 + a_1\lambda^3 + a_2\lambda^2 + a_3\lambda + a_4 = 0$$

then X satisfies

$$X^4 + a_1X^3 + a_2X^2 + a_3X + a_4I = 0$$

and $$X^2 + A_1X + A_2 = 0 \quad .$$

By elimination $$X = - J_3^{-1} K_3$$

where $$J_1 = a_1I - A_1 \qquad K_1 = a_2I - A_2$$

$$J_2 = K_1 - J_1A_1 \qquad K_2 = a_3I - J_1A_2$$

$$J_3 = K_2 - J_2A_1 \qquad K_3 = a_4I - J_2A_2 \quad .$$

m×m Case.

Let the possible characteristic equation be

$$\lambda^m + a_1\lambda^{m-1} + a_2\lambda^{m-2} + ... + a_{m-1}\lambda + a_m = 0$$

then X satisfies

$$X^m + a_1X^{m-1} + a_2X^{m-2} + ... + a_{m-1}X + a_mI = 0$$

and $$X^2 + A_1X + A_2 = 0 \quad .$$

By elimination $$X = - J_{m-1}^{-1} K_{m-1} \quad .$$

where the J_{m-1} and K_{m-1} are obtained from the recurrence relations

$$J_1 = a_1I - A_1 \qquad K_1 = a_2I - A_2$$

$$J_{i+1} = K_i - J_iA_1 \qquad K_{i+1} = a_{i+2}I - J_iA_2$$

Choice of factor which may not lead to a solution.

Since the solution X at the final stage is obtained by the inversion of J_{m-1} it can be seen that a solution may not be

obtained if the matrix J_{m-1} is singular. When this happens it means that the final linear equation

$$J_{m-1}X + K_{m-1} = 0$$

either (1) has no solution

or (2) has an infinite number of solutions.

This can occur if the chosen factor of det $[A(\lambda)]$ is not in fact the characteristic polynomial of a solution X. Though the characteristic polynomial of a solution X must be a factor of det $[A(\lambda)]$ it is not necessarily true that every factor of degree m is the characteristic polynomial of a solution X.

It is also true that J_{m-1} can be singular even when the chosen factor is the characteristic polynomial of a solution. The two cases are illustrated in the following examples.

Example 4.3.1.

Consider the equation $X^2 + A_1X + A_2 = 0$

where $A_1 = \begin{pmatrix} -1 & -6 \\ 2 & -9 \end{pmatrix}$ $A_2 = \begin{pmatrix} 0 & 12 \\ -2 & 14 \end{pmatrix}$

$$A(\lambda) = \begin{pmatrix} \lambda^2-\lambda & -6\lambda+12 \\ 2\lambda-2 & \lambda^2-9\lambda+14 \end{pmatrix}.$$

$\therefore$ $$\det [A(\lambda)] = (\lambda-1)(\lambda-2)(\lambda-3)(\lambda-4) .$$

Choosing $\phi_1(\lambda) = (\lambda-3)(\lambda-4)$ as the characteristic polynomial of a solution leads to the final linear equation

$$\begin{pmatrix} 6 & -6 \\ 2 & -2 \end{pmatrix} X = \begin{pmatrix} 12 & -12 \\ 2 & -2 \end{pmatrix} .$$

This linear equation has no solution and hence there is no solution of $X^2 + A_1X + A_2 = 0$ which has $(\lambda-3)(\lambda-4)$ as its characteristic polynomial.

Choosing $\phi_2(\lambda) = (\lambda-1)(\lambda-2)$ as the characteristic polynomial of a solution leads to the final linear equation

$$\begin{pmatrix} 2 & -6 \\ 2 & -6 \end{pmatrix} X = \begin{pmatrix} 2 & -12 \\ 2 & -12 \end{pmatrix} .$$

This linear equation has an infinite number of solutions. In parametric form the solution of the final linear equation is

$$X = \begin{pmatrix} 1+3a & 3b-6 \\ a & b \end{pmatrix} .$$

Substituting this in $X^2 + A_1X + A_2 = 0$

gives $$\begin{pmatrix} 3a(3a+b-3) & (3b-6)(3a+b-2) \\ a(3a+b-2) & (b-2)(3a+b-1) \end{pmatrix} = 0 .$$

Since $(3a+b-3)$, $(3a+b-2)$, $(3a+b-1)$ cannot be simultaneously zero, then $a = 0$ and $b = 2$

$\therefore$ the solution of the equation $X^2 + A_1X + A_2 = 0$ which has $(\lambda-1)(\lambda-2)$ as its characteristic polynomial is

$$X = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$$

Example 4.3.2.

This example shows that though a particular choice of factor

$\phi(\lambda)$ may lead to a linear equation which has an infinite number of solutions, it is not necessarily true that $\phi(\lambda)$ is the characteristic polynomial of a solution X.

Consider the equation $X^2 + A_1X + A_2 = 0$

where $A_1 = \begin{pmatrix} 5 & -1 \\ 1 & 3 \end{pmatrix}$ $A_2 = \begin{pmatrix} 0 & 3 \\ -3 & 6 \end{pmatrix}$

then $A(\lambda) = \begin{pmatrix} \lambda^2+5\lambda & -\lambda+3 \\ \lambda-3 & \lambda^2+3\lambda+6 \end{pmatrix}$

and $\det [A(\lambda)] = (\lambda+1)^2(\lambda+3)^2$.

Choosing $\phi_1(\lambda) = (\lambda+1)(\lambda+3)$ as the characteristic polynomial of a solution X leads to the final linear equation

$$\begin{pmatrix} 1 & -1 \\ 1 & -1 \end{pmatrix} X + \begin{pmatrix} -3 & 3 \\ -3 & 3 \end{pmatrix} = 0 \quad .$$

This linear equation has an infinite number of solutions. In parametric form the solution of the final linear equation is

$$X = \begin{pmatrix} a+3 & b-3 \\ a & b \end{pmatrix} .$$

Substituting this in $X^2 + A_1X + A_2 = 0$

gives $$\begin{pmatrix} a^2+7a+ab+24 & b^2+ab-3a+4b-21 \\ a^2+7a+ab & b^2+ab-3a+4b+3 \end{pmatrix} = 0 \quad .$$

Since there are no values of a and b which would make these four elements simultaneously zero then there is no solution X of $X^2 + A_1X + A_2 = 0$ which has $(\lambda+1)(\lambda+3)$ as its characteristic polynomial.

It is possible that equation (4.2.1) may have an infinite number of solutions. In this case the elimination method may lead to a particular solution or may again lead to the situation where J_{m-1} is singular. The two cases are illustrated in the following examples.

Example 4.3.3.

Consider the equation $A_o X^2 + A_1 X + A_2 = 0$

where $A_o = \begin{pmatrix} 1 & 2 & 0 \\ -1 & 0 & 1 \\ 3 & 1 & 2 \end{pmatrix}$ $A_1 = \begin{pmatrix} -4 & -8 & 0 \\ 4 & 0 & -4 \\ -12 & -4 & -8 \end{pmatrix}$

$A_2 = \begin{pmatrix} 3 & 6 & 0 \\ -3 & 0 & 3 \\ 9 & 3 & 6 \end{pmatrix}$

$$A(\lambda) = \begin{pmatrix} \lambda^2-4\lambda+3 & 2\lambda^2-8\lambda+6 & 0 \\ -\lambda^2+4\lambda-3 & 0 & \lambda^2-4\lambda+3 \\ 3\lambda^2-12\lambda+9 & \lambda^2-4\lambda+3 & 2\lambda^2-8\lambda+6 \end{pmatrix}$$

and $\det [A(\lambda)] = 9(\lambda-1)^3(\lambda-3)^3$.

Choosing $(\lambda-1)^3$ as the characteristic polynomial of X, then X satisfies the two equations

(1) $X^3 - 3X^2 + 3X - I = 0$

(2) $A_o X^2 + A_1 X + A_2 = 0$.

In this case it is necessary to make equation (2) monic by

multiplying on the left by A_o^{-1}. Eliminating X^3 and X^2 leads to the final linear equation

$$J_2X + K_2 = 0$$

where $$J_2 = A_1A_o^{-1}A_1 + 3A_1 - A_2 + 3A_o$$

$$K_2 = A_1A_o^{-1}A_2 + 3A_2 - A_o$$

$$\therefore \quad \begin{pmatrix} 4 & 8 & 0 \\ -4 & 0 & 4 \\ 12 & 4 & 8 \end{pmatrix} X = \begin{pmatrix} 4 & 8 & 0 \\ -4 & 0 & 4 \\ 12 & 4 & 8 \end{pmatrix}$$

$$\therefore \quad X = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Choosing $(\lambda-3)^3$ as the characteristic polynomial similar calculations lead to the solution

$$X = \begin{pmatrix} 3 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 3 \end{pmatrix} .$$

Choosing $(\lambda-1)^2(\lambda-3)$ as the characteristic polynomial however leads to the situation where J_2 is singular.

If X satisfies the two equations

(1) $$X^3 - 5X^2 + 7X - 3I = 0$$

(2) $$A_oX^2 + A_1X + A_2 = 0 \quad .$$

Eliminating X^3 and X^2 leads to

$$J_2X + K_2 = 0$$

where $$J_2 = A_2 - 7A_o - A_1A_o^{-1}A_1 - 5A_1$$

$$K_2 = 3A_o - A_1A_o^{-1}A_2 - 5A_2$$

But $$J_2 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad K_2 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

Hence the method fails at this stage. There are in fact an infinite number of solutions of the equation $A_oX^2 + A_1X + A_2 = 0$ and any matrix with characteristic polynomial $(\lambda-1)^2(\lambda-3)$ is a solution. The infinite set of solutions is therefore the set

$$T\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{pmatrix}T^{-1}$$

where T is any non-singular matrix.

Example 4.3.4.

Consider the equation $X^2 + A_2 = 0$

where $$A_2 = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -4 \end{pmatrix}$$

then $$A(\lambda) = \begin{pmatrix} \lambda^2-1 & 0 & 0 \\ 0 & \lambda^2-1 & 0 \\ 0 & 0 & \lambda^2-4 \end{pmatrix}$$

and $$\det [A(\lambda)] = (\lambda-1)^2(\lambda+1)^2(\lambda-2)(\lambda+2) .$$

Choosing $(\lambda-1)^2(\lambda-2)$ as the characteristic polynomial of X then X satisfies the two equations

$$(1) \qquad X^3 - 4X^2 + 5X - 2I = 0$$

$$(2) \qquad X^2 + A_2 = 0 \quad .$$

Eliminating X^2 and X^3 leads to the final linear equation

$$[A_2-5I]X + [2I-4A_2] = 0$$

$$\therefore \qquad \begin{pmatrix} -6 & 0 & 0 \\ 0 & -6 & 0 \\ 0 & 0 & -9 \end{pmatrix} X = \begin{pmatrix} -6 & 0 & 0 \\ 0 & -6 & 0 \\ 0 & 0 & -18 \end{pmatrix}$$

$$\therefore \qquad X = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix} .$$

The elimination method in this case had led to a particular solution when in fact there are an infinite number of solutions with characteristic polynomial $(\lambda-1)^2(\lambda-2)$ and any matrix of the form $\begin{pmatrix} a & \frac{1-a^2}{b} & 0 \\ b & -a & 0 \\ 0 & 0 & 2 \end{pmatrix}$ where $b \neq 0$ is a solution of $X^2 + A_2 = 0$.

These examples illustrate the different problems which can occur in applying the elimination method. If J_{m-1} is singular and the linear equation $J_{m-1}X + K_{m-1} = 0$ has no solution then it can be stated that there is no solution of the matrix equation corresponding to that choice of factor for $\phi(\lambda)$.

108

If however the matrix J_{m-1} is singular and the linear equation $J_{m-1}X + K_{m-1} = 0$ has an infinite number of solutions then it is possible that the original matrix equation may or may not have a solution corresponding to that choice of factor or in fact the original matrix equation may have an infinite number of solutions.

4.4 SOLUTION OF THE MATRIX RICCATI EQUATION BY ELIMINATION.

The elimination method may be extended to finding solutions of the Riccati equation

$$XEX + DX + XF + G = 0.$$

This equation may be made unilateral, provided the matrix E is non-singular, by means of the substitution

$$X = (Z-D)E^{-1} .$$

The equation then becomes

$$Z^2 + ZP + Q = 0$$

where $$P = E^{-1}FE - D$$

$$Q = GE - DE^{-1}FE \quad .$$

The method previously described may be used to obtain a solution Z and hence a solution X. The method is illustrated in the following examples.

Example 4.4.1.

Consider the equation $XEX + DX + XF + G = 0$

where $$E = \begin{pmatrix} 1 & -1 \\ 1 & 0 \end{pmatrix} \quad D = \begin{pmatrix} 3 & 1 \\ -1 & 2 \end{pmatrix} \quad F = \begin{pmatrix} 3 & -1 \\ 1 & 2 \end{pmatrix}$$

$$G = \begin{pmatrix} -14 & 9 \\ 7 & -7 \end{pmatrix} .$$

This is equivalent to $Z^2 + ZP + Q = 0$

where $$P = \begin{pmatrix} 0 & -2 \\ 2 & 0 \end{pmatrix} \quad Q = \begin{pmatrix} -15 & 15 \\ 1 & -12 \end{pmatrix}$$

$$A(\lambda) = [\lambda^2+\lambda P+Q] = \begin{pmatrix} \lambda^2-15 & -2\lambda+15 \\ 2\lambda+1 & \lambda^2-12 \end{pmatrix}$$

$\therefore$ $$\det [A(\lambda)] = (\lambda^2-7\lambda+11)(\lambda^2+7\lambda+15) .$$

Since $\lambda^2+7\lambda+15$ is irreducible over the real numbers, there are only 2 possible choices for the characteristic polynomial of a solution Z

$$\phi_1(\lambda) = \lambda^2 - 7\lambda + 11$$

$$\phi_2(\lambda) = \lambda^2 + 7\lambda + 15 \quad .$$

Choosing $\phi_1(\lambda) = \lambda^2-7\lambda+11$ as the characteristic polynomial of Z then Z satisfies the two equations

$$Z^2 - 7Z + 11I = 0$$

and $$Z^2 + ZP + Q = 0 \quad .$$

Eliminating Z^2 gives

$$Z[P+7I] = 11I - Q$$

$\therefore$ $$Z = \begin{pmatrix} 26 & -15 \\ -1 & 23 \end{pmatrix} \begin{pmatrix} 7 & -2 \\ 2 & 7 \end{pmatrix}^{-1}$$

$\therefore$ $$Z = \begin{pmatrix} 4 & -1 \\ -1 & 3 \end{pmatrix}$$

and hence $$X = [Z-D]E^{-1} = \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}$$

Choosing $\phi_2(\lambda) = \lambda^2+7\lambda+15$ as the characteristic polynomial of Z then Z satisfies the two equations

$$Z^2 + 7Z + 15I = 0$$

and $$Z^2 + ZP + Q = 0$$

$$\therefore \quad Z[P-7I] = 15I - Q$$

$$\therefore \quad Z = \begin{pmatrix} 30 & -15 \\ -1 & 27 \end{pmatrix} \begin{pmatrix} -7 & -2 \\ 2 & -7 \end{pmatrix}^{-1}$$

$$\therefore \quad z = \frac{1}{53} \begin{pmatrix} -180 & 165 \\ -47 & -191 \end{pmatrix}$$

Hence $$X = \frac{1}{53} \begin{pmatrix} -112 & -227 \\ 297 & -291 \end{pmatrix}$$

Example 4.4.2.

This example is the same problem as the one used in Method 5 of 3.3 where the method of solution was the one described by Incertis [1983]. This led only to the positive definite solution, whereas the elimination method can be used to obtain six solutions.

Consider the equation $PBR^{-1}B^TP - A^TP - PA - Q = 0$

where $$A = \begin{pmatrix} -1 & 0 \\ 1 & 2 \end{pmatrix} \qquad B = \frac{1}{\sqrt{11}} \begin{pmatrix} \sqrt{7} & 9 \\ 0 & 11 \end{pmatrix} \qquad R = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$Q = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$$

and P is the matrix to be determined.

Let $D = BR^{-1}B^T = \begin{pmatrix} 8 & 9 \\ 9 & 11 \end{pmatrix}$ and let $Z = PD - A^T$

then the original equation is equivalent to the unilateral equation

$$Z^2 + ZF + G = 0$$

where $F = - D^{-1}AD + A^T$

$G = - QD - A^TD^{-1}AD$

$$A(\lambda) = \lambda^2 I + \lambda F + G = \begin{pmatrix} \lambda^2+45\lambda-94 & 55\lambda-110 \\ -40\lambda-80 & \lambda^2-45\lambda-94 \end{pmatrix}$$

and $\det [A(\lambda)] = (\lambda-3)(\lambda+3)(\lambda-2)(\lambda+2)$.

There are therefore 6 possible choices for the characteristic polynomial of Z

$\phi_1(\lambda) = \lambda^2 - 5\lambda + 6$ $\qquad$ $\phi_4(\lambda) = \lambda^2 + \lambda - 6$

$\phi_2(\lambda) = \lambda^2 - 9$ $\qquad$ $\phi_5(\lambda) = \lambda^2 + 5\lambda + 6$

$\phi_3(\lambda) = \lambda^2 - \lambda - 6$ $\qquad$ $\phi_6(\lambda) = \lambda^2 - 4$.

Choosing $\phi_1(\lambda)$ Z satisfies the two equations

$$Z^2 - 5Z + 6I = 0$$

and $Z^2 + ZF + G = 0$.

Eliminating Z^2 gives

$$Z = [6I-G][F+5I]^{-1}$$

$\therefore$ $Z = \begin{pmatrix} 2 & 0 \\ 4 & 3 \end{pmatrix}$

and from $P = [Z+A^T]D^{-1}$ we obtain $P = \frac{1}{7}\begin{pmatrix} 2 & -1 \\ -1 & 4 \end{pmatrix}$

Choosing $\phi_2(\lambda) = \lambda^2-9$ then Z satisfies

$$Z^2 - 9I = 0$$

and $$Z^2 + ZF + G = 0 \quad .$$

Eliminating Z^2 gives

$$Z = \frac{1}{7}\begin{pmatrix} 23 & 11 \\ -8 & -23 \end{pmatrix}$$

and hence $$P = \frac{1}{7}\begin{pmatrix} 2 & 0 \\ -1 & 0 \end{pmatrix}$$

Choosing $\phi_3(\lambda) = \lambda^2-\lambda-6$ then Z satisfies

$$Z^2 - Z - 6I = 0$$

and $$Z^2 + ZF + G = 0 \quad .$$

Eliminating Z^2 gives

$$Z = \frac{1}{4}\begin{pmatrix} 12 & 5 \\ 0 & -8 \end{pmatrix}$$

and hence $$P = \frac{1}{4}\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$$

Choosing $\phi_4(\lambda) = \lambda^2+\lambda-6$ leads to

$$Z = \frac{1}{11}\begin{pmatrix} 22 & 0 \\ -10 & -33 \end{pmatrix}$$

and hence $$P = \frac{1}{7}\begin{pmatrix} 2 & -1 \\ -1 & \frac{2}{11} \end{pmatrix}$$

Choosing $\phi_5(\lambda) = \lambda^2+5\lambda+6$ leads to

$$Z = \frac{1}{2}\begin{pmatrix} -6 & -11 \\ 0 & -4 \end{pmatrix}$$

and hence $$P = \frac{1}{2}\begin{pmatrix} -1 & 0 \\ 0 & 0 \end{pmatrix}$$

Choosing $\phi_6(\lambda) = \lambda^2-4$ leads to

$$Z = \begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix}$$

and hence $$P = \frac{1}{7}\begin{pmatrix} 2 & -1 \\ 0 & 0 \end{pmatrix}$$

Example 4.4.3.

This example shows how the method can be used to solve the matrix Riccati equation involving 3×3 matrices.

Consider the equation $XEX + DX + XF + G = 0$

where $$E = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \qquad D = \begin{pmatrix} 2 & 0 & 1 \\ -1 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix}$$

$$F = \begin{pmatrix} 2 & -1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & -1 \end{pmatrix} \qquad G = \begin{pmatrix} -6 & 5 & -3 \\ 2 & -3 & 2 \\ -5 & 2 & 0 \end{pmatrix}$$

Using the substitution $Z = D + XE$ the equation is equivalent to the unilateral equation

$$Z^2 + ZP + Q = 0$$

where $$P = \begin{pmatrix} -2 & -1 & 0 \\ 2 & 1 & -2 \\ 1 & 0 & 1 \end{pmatrix} \qquad Q = \begin{pmatrix} -2 & 7 & -13 \\ -2 & -6 & 8 \\ -2 & 2 & -5 \end{pmatrix}$$

$$A(\lambda) = \lambda^2 I + \lambda P + Q = \begin{pmatrix} \lambda^2-2\lambda-2 & -\lambda+7 & \lambda-13 \\ 2\lambda-2 & \lambda^2+\lambda-6 & -2\lambda+8 \\ \lambda-2 & 2 & \lambda^2+\lambda-5 \end{pmatrix}$$

and $\det [A(\lambda)] = [\lambda^3-5\lambda^2+6\lambda-1][\lambda^3+5\lambda^2+4\lambda+2]$.

Since Z is a 3×3 matrix its characteristic polynomial will be of degree 3.

Possible choices are $\phi_1(\lambda) = \lambda^3 - 5\lambda^2 + 6\lambda - 1$

$\phi_2(\lambda) = \lambda^3 + 5\lambda^2 + 4\lambda + 2$.

The solution Z is given by

$Z = - K_2 J_2^{-1}$ where $J_1 = a_1 I - P$ $\qquad K_1 = a_2 I - Q$

$J_2 = K_1 - PJ_1$ $\qquad K_2 = a_3 I - QJ_1$

and a_1, a_2, a_3 are the coefficients of the characteristic polynomial of Z.

Choosing $\phi_1(\lambda) = \lambda^3-5\lambda^2+6\lambda-1$ as the characteristic polynomial of Z

then $a_1 = -5$ $\qquad a_2 = 6$ $\qquad a_3 = -1$

and $$J_1 = a_1 I - P = \begin{pmatrix} -3 & 1 & -1 \\ -2 & -6 & 2 \\ -1 & 0 & -6 \end{pmatrix} \qquad K_1 = a_2 I - Q = \begin{pmatrix} 8 & -7 & 13 \\ 2 & 12 & -8 \\ 2 & -2 & 11 \end{pmatrix}$$

$$J_2 = K_1 - PJ_1 = \begin{pmatrix} 1 & -11 & 19 \\ 8 & 16 & -20 \\ 6 & -3 & 18 \end{pmatrix} \qquad K_2 = a_3 I - QJ_1 = \begin{pmatrix} -6 & 44 & -94 \\ -10 & -35 & 58 \\ -7 & 14 & -37 \end{pmatrix}$$

$$\text{then } Z = -K_2 J_2^{-1} = \begin{pmatrix} 2 & -1 & 2 \\ 0 & 2 & -1 \\ 1 & 0 & 1 \end{pmatrix}$$

$$\text{and } X = [Z-D]E^{-1} = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ 1 & 0 & 1 \end{pmatrix}$$

Choosing $\phi_2(\lambda) = \lambda^3+5\lambda^2+4\lambda+2$ as the characteristic polynomial of Z

then $a_1 = 5 \qquad a_2 = 4 \qquad a_3 = 2$

$$\therefore \quad J_1 = \begin{pmatrix} 7 & 1 & -1 \\ -2 & 4 & 2 \\ -1 & 0 & 4 \end{pmatrix} \qquad K_1 = \begin{pmatrix} 6 & -7 & 13 \\ 2 & 10 & -8 \\ 2 & -2 & 9 \end{pmatrix}$$

$$J_2 = \begin{pmatrix} 19 & -1 & 9 \\ -12 & 4 & 0 \\ -4 & -3 & 6 \end{pmatrix} \qquad K_2 = \begin{pmatrix} 17 & -26 & 36 \\ 10 & 28 & -22 \\ 13 & -6 & 16 \end{pmatrix}$$

$$\therefore \quad Z = -K_2 J_2^{-1} = \frac{1}{852} \begin{pmatrix} -408 & 2061 & -4500 \\ -1112 & -2648 & 4792 \\ -712 & 197 & -1204 \end{pmatrix}$$

$$\text{and} \quad X = [Z-D]E^{-1} = \frac{1}{852} \begin{pmatrix} -4173 & 2061 & -1384 \\ 3240 & -3500 & 1552 \\ -909 & 197 & 557 \end{pmatrix}$$

$$\therefore \quad X \approx \begin{pmatrix} -4.898 & 2.419 & -1.624 \\ 3.803 & -4.108 & 1.822 \\ -1.067 & 0.231 & 0.654 \end{pmatrix}$$

The method may be extended to the solution of the Riccati equation for m×m matrices.

The equation $XEX + DX + XF + G = 0$

has a solution $X = [Z-D]E^{-1}$

where Z is the solution of the unilateral equation

$$Z^2 + ZP + Q = 0 \quad \text{where} \quad P = E^{-1}FE - D \quad \text{and} \quad Q = GE - DE^{-1}FE .$$

Then $Z = -K_{m-1} J_{m-1}^{-1}$ where K_{m-1} and J_{m-1} may be obtained from the recurrence relations

$$J_1 = a_1 I - P$$

$$K_1 = a_2 I - Q$$

$$J_{i+1} = K_i - PJ_i \qquad K_{i+1} = a_{i+2} I - QJ_i \qquad i = 1 \text{ to } m-2$$

and the a_i are the coefficients of the characteristic polynomial of Z.

As in the case of the unilateral quadratic equation, certain choices of factor from $\det [\lambda^2 I + \lambda P + Q]$ may not lead to a solution Z of the equation $Z^2 + ZP + Q = 0$.

4.5 CONCLUSION.

This chapter has shown how solutions of matrix equations may be obtained by using two properties

a) that any matrix satisfies its own characteristic equation

b) that the characteristic polynomial of any solution X of the unilateral matrix equation is a factor of det [$A(\lambda)$].

The usefulness of this method as a practical means of finding solutions can only be decided by comparing it with established methods.

The elimination method compares favourably with four of the methods of solution described in Chapter 3. They are Methods I and II of 3.2 described by Gantmacher [1959] and Dennis Traub & Weber [1976], and Methods I and III of 3.3 described by Roth [1950] and Potter [1966].

The first stage of each of these four methods and also of the elimination method is virtually the same. All involve forming a polynomial of degree 2m and finding m roots. In each case this is equivalent to finding the eigenvalues of a 2m×2m matrix since the roots of the equation det $[\lambda^2 I+A_1\lambda+A_2] = 0$ are the same as the eigenvalues of the Block Companion Matrix $\begin{pmatrix} 0 & -A_2 \\ I & -A_1 \end{pmatrix}$.

Having found at least m roots, the Roth method involves forming the degree m polynomial $f(\lambda) = (\lambda-\lambda_1)(\lambda-\lambda_2) \ldots\ldots (\lambda-\lambda_m)$ where the λ_i are m eigenvalues chosen from the eigenvalues of a 2m×2m matrix. This is also the second stage in the elimination method.

In the Roth method, when $f(\lambda)$ is formed then $f(R)$ must be evaluated where R is a 2m×2m matrix. For the Potter method, eigenvectors of a 2m×2m matrix must be found and in the Gantmacher method a set of m^2 linear equations in the elements of the

transforming matrix must be solved. The number of operations involved in the elimination method therefore compares very favourably with the four established methods.

All the methods involve the inversion of an m×m matrix to obtain the solution X.

In the elimination method the final solution is obtained from $X = - J_{m-1}^{-1} K_{m-1}$ and since the J_{m-1} and K_{m-1} are obtainable from recurrence relations, this makes their computation easily programmable for the computer.

The four established methods of solution all fail to lead to solutions in certain cases. The Gantmacher method fails if the transforming matrix T is singular. The Dennis Traub and Weber method fails if the latent vectors are linearly dependent. The Roth method fails if both M and U are singular when $f(R) = \begin{pmatrix} U & M \\ V & N \end{pmatrix}$ and the Potter method fails if the eivenvectors of the matrix M are linearly dependent.

The elimination method may fail to lead to a solution if J_{m-1} is singular. This happens if the choice of factor of degree m from det $[A(\lambda)]$ is not the characteristic polynomial of a solution. However, it is possible for J_{m-1} to be singular even when the choice of factor is the characteristic polynomial of a solution as illustrated in examples 4.3.1 and 4.3.2. In spite of this the elimination method does appear to offer some advantages for the solution of the matrix quadratic equation.

CHAPTER 5

Iterative Methods Applied to the Solution of Matrix Equations

5.1 INTRODUCTION.

Iterative methods for the solution of a polynomial equation in a single variable are well known. In this chapter an attempt is made to apply some of these methods directly to the matrix equation. In doing this the usual problems of non commutativity and singularity arise. Some methods also require the derivative of f(X) where X is a matrix and f is a matrix valued function. The Fréchet derivative of f is an operator and is described by what it does to a typical matrix.

The derivative operator is defined as the coefficient of Y in f(X+Y) - f(X).

For example if g'(X) is required where g(X) is the matrix valued function defined by

$$g(X) = A_0X^2 + A_1X + A_2$$

where X and A_i, $i = 0,1,2$ are m×m matrices.

Then

$$g(X+Y) - g(X) = A_0[X+Y]^2 + A_1[X+Y] + A_2 - [A_0X^2+A_1X+A_2]$$

$$= A_0XY + A_0YX + Y^2 + A_1Y$$

and the derivative operator in this case is

$$g'(X) = [A_0X+A_1][\ ] + A_0[\ ]X$$

where the square brackets are replaced by the matrix which the

derivative operator g'(X) is operating on.

Hence g'(X)H means "the operator g'(X) applied to H" and

$$g'(X)H = [A_o X + A_1]H + A_o HX \quad .$$

The derivative can also be expressed as an $m^2 \times m^2$ matrix. In this case

$$g'(X)H = [(A_o X + A_1) \otimes I + A_o \otimes X^T]\underline{h}$$

where $\otimes$ is the Kronecker product and $\underline{h}$ is the column vector composed of the transposed rows of H, taken in order.

i.e. $\underline{h} = [h_{11}, h_{12}, \ldots, h_{1m}, h_{21}, h_{22}, \ldots, h_{2m}, \ldots, h_{m1}, h_{m2}, \ldots, h_{mm}]^T$.

It is shown in 5.4 that this $m^2 \times m^2$ matrix is the same as the Jacobian matrix of the functions $g_{ij}(\underline{x})$ i = 1...m, j = 1...m, where the $g_{ij}(\underline{x}) = 0$ are the constituent equations obtained from the matrix equation g(X) = 0.

The Newton-Raphson method for the solution of n polynomial equations in n variables which can be applied to the constituent equations can also be applied directly to the matrix equation by use of the derivative operator. Though the two versions of the Newton method are equivalent, the computation involved in applying it directly to the matrix equation is often simpler since the formation of the constituent equations is not required.

In this chapter, section 2 deals with the method of simple iteration. This is a well known iterative process for the solution of a scalar polynomial equation. The equation F(x) = 0 is rearranged in the form x = f(x) and the iterative function $x_{i+1} = f(x_i)$ is defined.

The method of simple iteration can be extended to a system of equations $\underline{F}(\underline{x}) = 0$

where $\underline{F}(\underline{x}) = (F_1(\underline{x}), F_2(\underline{x}) \ldots F_n(\underline{x}))^T$

and $\underline{x} = (x_1 x_2, \ldots, x_n)$.

The system of equations is rearranged in the form $\underline{x} = \underline{f}(\underline{x})$ and the iterative function is then $\underline{x}_{i+1} = \underline{f}(\underline{x}_i)$. In section 2 an attempt is made to apply the method directly to the matrix equation.

Section 5.3 applies the Bernoulli algorithm for the solution of a scalar polynomial equation to the unilateral matrix equation. An algorithm described by Dennis, Traub and Weber [1978] is illustrated as a means of increasing the rate of convergence of the Bernoulli iteration.

In section 5.4 the Newton iterative method is applied to both the constituent equations and directly to the matrix equation. The difficulty in finding conditions for convergence is illustrated by several numerical examples.

The chapter ends by showing an iterative method based on the elimination method of Chapter 4. The main advantage of this method is that only m initial values are required when the matrices involved are m×m, instead of the m^2 values required for most iterative methods.

5.2 THE METHOD OF SIMPLE ITERATION.

Given the equation $x = f(x)$, an iterative process can be set up and denoted by $x_{i+1} = f(x_i)$ $i = 0,1,2,\ldots$ with x_o a given starting value.

To find a solution of $F(x) = 0$, the equation is rearranged

in the form $x = f(x)$. If $|f'(a)| < 1$ for a solution $x = a$ of $F(x) = 0$, then the sequence of values x_i obtained from $x_{i+1} = f(x_i)$ will converge to the solution $x = a$ for a starting value x_o near a. The distance of x_o from the solution necessary for convergence depends upon the function $f(x)$.

For example, the equation $x^2 - 4x - 5 = 0$ has two solutions $x = -1$ and $x = 5$.

The equation may be rearranged in the form

$$x = \frac{x^2-5}{4}$$

and the iterative process $x_{i+1} = f(x_i)$ defined where

$$f(x) = \frac{x^2-5}{4} \quad .$$

In this case $f'(x) = \frac{x}{2}$ and $|f'(-1)| = 0.5 \qquad |f'(5)| = 2.5$.

Hence the iteration can be expected to converge to the solution $x = -1$ for a starting value near $x = -1$ but will not converge to the solution $x = 5$ for a starting value near $x = 5$. In fact the iteration converges to the root $x = -1$ for all $x \in]-5.\ 5[$.

The same equation may be rearranged in the form $x = \frac{5}{x} + 4$ and the iterative process $x_{i+1} = f(x_i)$ defined where $f(x) = \frac{5}{x} + 4$.

In this case

$$f'(x) = -\frac{5}{x^2} \quad \text{and} \quad |f'(-1)| = 5 \qquad |f'(5)| = 0.2 \ .$$

Hence this iteration will converge to the root $x = 5$ for a starting value near $x=5$ but will not converge to the solution $x = -1$. In fact the iteration converges to $x = 5$ for all x except $x = 0$.

In this section an attempt is made to apply the method of

simple iteration to the unilateral matrix quadratic equation.

Consider the equation

$$X^2 + A_1X + A_2 = 0$$

where X, A_1. A_2 are m×m matrices. The equation may be rearranged in the form

$$X = - A_1^{-1}[A_2 + X^2] \quad , \quad A_1 \text{ nonsingular .}$$

and from this the iterative process $X_{i+1} = f(X_i)$ defined where

$$f(X) = - A_1^{-1}[A_2 + X^2].$$

The process is applied in the following examples.

Example 5.2.1.

Consider the equation $X^2 + A_1X + A_2 = 0$

where $A_1 = \begin{pmatrix} 4 & -4 \\ 3 & 5 \end{pmatrix}$ $A_2 = \begin{pmatrix} -9 & -4 \\ 2 & 4 \end{pmatrix}$.

Taking $X_o = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ as the initial matrix, the sequence

of matrices obtained from $X_{i+1} = - A_1^{-1}[A_2 + X_i^2]$ is

$$X_1 = \begin{pmatrix} 0.59375 & -0.4375 \\ -1.15625 & -0.9375 \end{pmatrix}$$

$$X_2 = \begin{pmatrix} 0.9724426 & -0.0715942 \\ -1.0629578 & -1.0339966 \end{pmatrix}$$

$$X_3 = \begin{pmatrix} 0.9884235 & -0.0174677 \\ -1.00614 & -1.0185695 \end{pmatrix}$$

$$X_4 = \begin{pmatrix} 0.9970592 & -0.0069646 \\ -1.0043018 & -1.006833 \end{pmatrix}$$

$$X_5 = \begin{pmatrix} 0.9985978 & -0.002599 \\ -1.0011218 & -1.002582 \end{pmatrix}$$

$$X_6 = \begin{pmatrix} 0.9995327 & -0.00097318 \\ -1.0005174 & -1.0009706 \end{pmatrix}$$

$$X_7 = \begin{pmatrix} 0.999814 & -0.00036469 \\ -1.0001762 & -1.0003643 \end{pmatrix} .$$

The sequence is apparently converging to $X = \begin{pmatrix} 1 & 0 \\ -1 & -1 \end{pmatrix}$ which is a solution.

By using the method described in Chapter 4 it can be shown that there are exactly 2 real solutions of the equation

$$X^2 + A_1X + A_2 = 0.$$

The other solution is $X = \begin{pmatrix} -4.875 & 2.5 \\ -3.15625 & -4.125 \end{pmatrix}$ where the elements are exact values.

Taking as an initial matrix $X_o = \begin{pmatrix} -4.9 & 2.4 \\ -3.2 & -4.1 \end{pmatrix}$ which is very close to the second solution, the sequence of matrices obtained is

$$X_1 = \begin{pmatrix} -4.9953125 & 2.35875 \\ -3.1628125 & -4.04125 \end{pmatrix}$$

$$X_2 = \begin{pmatrix} -5.149629 & 2.3465404 \\ -3.0264132 & -3.9822076 \end{pmatrix}$$

$$X_3 = \begin{pmatrix} -5.3322546 & 2.3786128 \\ -2.7279791 & -3.9784429 \end{pmatrix}$$

$$X_4 = \begin{pmatrix} -5.4474443 & 2.4179974 \\ -2.211411 & -4.1186386 \end{pmatrix}$$

$$X_5 = \begin{pmatrix} -5.2892337 & 2.2871815 \\ -1.457368 & -4.4955085 \end{pmatrix}$$

$$X_6 = \begin{pmatrix} -4.4766725 & 1.5122524 \\ -0.5659905 & -0.1364652 \end{pmatrix}$$

$$X_7 = \begin{pmatrix} -2.1677296 & 1.3196978 \\ 0.3784394 & -1.4243592 \end{pmatrix}$$

$$X_8 = \begin{pmatrix} 0.5139114 & 0.5496706 \\ -0.4364693 & -1.6354473 \end{pmatrix}$$

$$X_9 = \begin{pmatrix} 1.0912807 & -0.0830224 \\ -1.1526716 & -1.2371413 \end{pmatrix}$$

$$X_{10} = \begin{pmatrix} 0.934204 & -0.0801692 \\ -0.9941483 & -1.0771417 \end{pmatrix}$$

$$X_{11} = \begin{pmatrix} 0.989669 & -0.0317823 \\ -1.0222217 & -1.0289175 \end{pmatrix}$$

$$X_{12} = \begin{pmatrix} 0.9931203 & -0.011594 \\ -1.0038963 & -1.011278 \end{pmatrix}$$

$$X_{13} = \begin{pmatrix} 0.9980453 & -0.0043232 \\ -1.0024729 & -1.0042705 \end{pmatrix}$$

$$X_{14} = \begin{pmatrix} 0.999153 & -0.0016158 \\ -1.0007399 & -1.0016091 \end{pmatrix}$$

The sequence is clearly converging to the solution $X = \begin{pmatrix} 1 & 0 \\ -1 & -1 \end{pmatrix}$ even though the initial matrix was very close to the other solution.

This shows similarities to the scalar case when the rearrangement $x = \frac{x^2-5}{4}$ of the equation $x^2 - 4x - 5 = 0$ converges to the solution $x = -1$ even when the starting value is $x_o = 4.9$ which is very close to the other solution $x = 5$.

Example 5.2.2.

Consider the equation $X^2 + A_1X + A_2 = 0$

where $A_1 = \begin{pmatrix} 8 & -6 \\ 6 & 8 \end{pmatrix}$ $\qquad A_2 = \begin{pmatrix} -14 & -19 \\ 1 & 7 \end{pmatrix}$.

Taking $X_o = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ the sequence of matrices obtained from

the iterative function $X_{i+1} = - A_1^{-1}[A_2 + X_i^2]$ is

$$X_1 = \begin{pmatrix} 0.98 & 1.04 \\ -0.86 & -1.78 \end{pmatrix}$$

$$X_2 = \begin{pmatrix} 1.01344 & 1.03012 \\ -0.97108 & -1.93184 \end{pmatrix}$$

$$X_3 = \begin{pmatrix} 1.0043511 & 1.0117844 \\ -0.9897433 & -1.9752979 \end{pmatrix}$$

$$X_4 = \begin{pmatrix} 1.0017556 & 1.0045673 \\ -0.9964402 & -1.9909749 \end{pmatrix}$$

$$X_5 = \begin{pmatrix} 1.0006563 & 1.0017196 \\ -0.9987045 & -1.9966634 \end{pmatrix}$$

$$X_6 = \begin{pmatrix} 1.0002457 & 1.000643 \\ -0.9995239 & -1.9987626 \end{pmatrix}.$$

The sequence is converging quite quickly to the matrix $X = \begin{pmatrix} 1 & 1 \\ -1 & -2 \end{pmatrix}$ which is a solution.

Again, by use of the method of Chapter 4 it can be shown that the equation has exactly two real solutions. The other solution is

$$\begin{pmatrix} -\frac{657}{85} & \frac{299}{85} \\ -\frac{551}{85} & -\frac{618}{85} \end{pmatrix} \approx \begin{pmatrix} -7.729 & 3.518 \\ -6.482 & -7.271 \end{pmatrix}$$

Taking $X_o = \begin{pmatrix} -8 & 4 \\ -6 & -7 \end{pmatrix}$ as an initial matrix which is close to this second solution the sequence of matrices obtained is

$$X_1 = \begin{pmatrix} -7.54 & 4.4 \\ -5.72 & -7.3 \end{pmatrix}$$

$$X_2 = \begin{pmatrix} -3.7384 & 4.63636 \\ -5.809768 & -7.86752 \end{pmatrix}$$

$$X_3 = \begin{pmatrix} -1.9488189 & 3.3070361 \\ -7.0918487 & -7.725489 \end{pmatrix}$$

$$X_4 = \begin{pmatrix} -1.4841153 & 1.4856519 \\ -7.5880045 & -6.5180116 \end{pmatrix}$$

$$X_5 = \begin{pmatrix} -1.8575677 & 0.1783895 \\ -6.3218461 & -4.9102099 \end{pmatrix}$$

$$X_6 = \begin{pmatrix} -1.6929126 & -0.1823606 \\ -4.2034215 & -3.6110308 \end{pmatrix}$$

$$X_7 = \begin{pmatrix} -0.5682819 & 0.1942567 \\ -2.4856273 & -2.7464528 \end{pmatrix}$$

$$X_8 = \begin{pmatrix} 0.5784407 & 0.7279035 \\ -1.5887299 & -2.3034468 \end{pmatrix}$$

129

$$X_9 = \begin{pmatrix} 0.9613137 & 0.9514855 \\ -1.1885564 & -2.1072923 \end{pmatrix}$$

$$X_{10} = \begin{pmatrix} 0.994818 & 0.9886434 \\ -1.041371 & -2.0302058 \end{pmatrix}$$

$$X_{11} = \begin{pmatrix} 0.9984971 & 0.9963589 \\ -1.0086507 & -2.008793 \end{pmatrix}$$

$$X_{12} = \begin{pmatrix} 0.9994963 & 0.9987131 \\ -1.0020017 & -2.0028187 \end{pmatrix}$$

As in example 5.2.1, though the initial value is chosen near one solution, the sequence is converging to the other solution.

In attempting to find reasons for convergence to a particular solution, the iterative process

$$X_{i+1} = - A_1^{-1}[A_2 + X_i^2]$$

may be considered in vector form.

In example 5.2.1 where $A_1 = \begin{pmatrix} 4 & -4 \\ 3 & 5 \end{pmatrix}$ and $A_2 = \begin{pmatrix} -9 & -4 \\ 2 & 4 \end{pmatrix}$

if X_i is taken to be the matrix $\begin{pmatrix} x_1 & x_2 \\ x_3 & x_4 \end{pmatrix}$ then

$$- A_1^{-1}[A_2+X_i^2] = \begin{pmatrix} -\frac{5}{32} & -\frac{4}{32} \\ -\frac{3}{32} & \frac{4}{32} \end{pmatrix} \begin{pmatrix} -9 + x_1^2 + x_2x_3 & -4 + x_1x_2 + x_2x_4 \\ 2 + x_1x_3 + x_3x_4 & 4 + x_2x_3 + x_4^2 \end{pmatrix}$$

$$= \begin{pmatrix} f_1(x_1,x_2,x_3,x_4) & f_2(x_1,x_2,x_3,x_4) \\ f_3(x_1,x_2,x_3,x_4) & f_4(x_1,x_2,x_3,x_4) \end{pmatrix}$$

where $f_1(x_1,x_2,x_3,x_4) = \frac{1}{32}[-5x_1^2 - 5x_2x_3 - 4x_1x_3 - 4x_3x_4 + 37]$

$f_2(x_1,x_2,x_3,x_4) = \frac{1}{32}[-5x_1x_2 - 5x_2x_4 - 4x_2x_3 - 4x_4^2 + 4]$

$f_3(x_1,x_2,x_3,x_4) = \frac{1}{32}[-3x_1^2 - 3x_2x_3 + 4x_1x_3 + 4x_3x_4 + 35]$

$f_4(x_1,x_2,x_3,x_4) = \frac{1}{32}[-3x_1x_2 - 3x_2x_4 - 4x_2x_3 + 4x_4^2 + 28]$.

Hence $X^{i+1} = -A_1^{-1}[A_2 + (X^i)^2]$ may be written as

$$\begin{pmatrix} x_1^{i+1} & x_2^{i+1} \\ x_3^{i+1} & x_4^{i+1} \end{pmatrix} = \begin{pmatrix} f_1(x_1^i,x_2^i,x_3^i,x_4^i) & f_2(x_1^i,x_2^i,x_3^i,x_4^i) \\ f_3(x_1^i,x_2^i,x_3^i,x_4^i) & f_4(x_1^i,x_2^i,x_3^i,x_4^i) \end{pmatrix}$$

or in vector form

$$\begin{pmatrix} x_1^{i+1} \\ x_2^{i+1} \\ x_3^{i+1} \\ x_4^{i+1} \end{pmatrix} = \begin{pmatrix} f_1(x_1^i,x_2^i,x_3^i,x_4^i) \\ f_2(x_1^i,x_2^i,x_3^i,x_4^i) \\ f_3(x_1^i,x_2^i,x_3^i,x_4^i) \\ f_4(x_1^i,x_2^i,x_3^i,x_4^i) \end{pmatrix}$$

The following result holds, see for example, Morris [1983] for the iterative process

$$\underline{x}^{i+1} = \underline{f}(\underline{x}^i) \quad .$$

Let R denote the region $a_i \leq x_i \leq b_i$ $i = 1,2,\ldots,n$ and let the functions f_i, $i = 1$ to n, satisfying the following conditions

(i) f_i is defined and continuous on R,

(ii) for each $\underline{x} \in R$ the point $f_i(\underline{x})$ also lies in R,

(iii) if J_f is the Jacobian matrix of $\underline{f}(\underline{x})$ then

$$\| J_f(\underline{x}) \| < 1 \quad \text{for some matrix norm} \quad .$$

Then the equation $\underline{x} = \underline{f}(\underline{x})$ has precisely one solution in R and for any choice $\underline{x}_o$ in R the sequence $\{\underline{x}^{(i)}\}$ given by $\underline{x}^{(i+1)} = \underline{f}(\underline{x}^{(i)})$ is defined and converges to the solution $\underline{x}$.

The Jacobian matrix in this example is

$$J = \frac{1}{32} \begin{pmatrix} -10x_1-4x_3 & -5x_3 & -5x_2-4x_4-4x_1 & -4x_3 \\ -5x_2 & -5x_1-5x_4-4x_3 & -4x_2 & -5x_2-8x_4 \\ 6x_1-4x_3 & 3x_3 & 3x_2-4x_4-4x_1 & -4x_3 \\ 3x_2 & -4x_3+3x_1+3x_4 & -4x_2 & 3x_2-8x_4 \end{pmatrix} .$$

At the solution $X = \begin{pmatrix} 1 & 0 \\ -1 & -1 \end{pmatrix}$

the Jacobian matrix is $\begin{pmatrix} -0.1875 & 0.15625 & 0 & 0.125 \\ 0 & -0.125 & 0 & 0.25 \\ 0.3125 & -0.09375 & 0 & 0.125 \\ 0 & 0.125 & 0 & 0.25 \end{pmatrix}$

and using a matrix norm, for example, $\| A \| = \left\{ \sum_{i=1}^{n} \sum_{i=1}^{n} |a_{ij}|^2 \right\}^{1/2}$

The norm of this matrix is $\| J \| = 0.658$ to 3 S.F.

At the solution $X = \begin{pmatrix} -4.875 & 2.5 \\ -3.15625 & -4.125 \end{pmatrix}$

the Jacobian matrix is $\begin{pmatrix} 1.918 & 0.493 & 0.734 & 0.395 \\ -0.391 & 1.801 & -0.3125 & 0.641 \\ -0.5195 & -0.296 & 1.359 & 0.395 \\ 0.234 & -0.449 & -0.3125 & 1.266 \end{pmatrix}$

and the norm of this matrix is $\|J\| = 3.547$.

Hence at the solution to which the iteration converges $\|f'(X)\| = 0.658$ and at the other solution $\|f'(X)\| = 3.547$ where $f'(X)$ is the Jacobian matrix of the function $f(X) = -A_1^{-1}[A_2 + X^2]$ written in vector form.

The condition for convergence is also satisfied in Example 5.2.2 where the equation to be solved is

$$X^2 + A_1X + A_2 = 0 \quad \text{and} \quad A_1 = \begin{pmatrix} 8 & -6 \\ 6 & 8 \end{pmatrix} \quad A_2 = \begin{pmatrix} -14 & -19 \\ 1 & 7 \end{pmatrix}.$$

With the rearrangement $X_{i+1} = f(X_i)$ where $f(X) = -A_1^{-1}[A_2 + X^2]$

then setting $f(X) = \begin{pmatrix} f_1(x_1,x_2,x_3,x_4) & f_2(x_1,x_2,x_3,x_4) \\ f_3(x_1,x_2,x_3,x_4) & f_4(x_1,x_2,x_3,x_4) \end{pmatrix}$

and writing the iterative process in vector form as $\underline{x}^{i+1} = \underline{f}(\underline{x}^i)$

the Jacobian matrix for $\underline{f}(\underline{x})$ is

$$J = \frac{1}{100}\begin{pmatrix} -16x_1-6x_3 & -8x_3 & -8x_2-6x_4-6x_1 & -6x_3 \\ -8x_2 & -8x_1-8x_4-6x_3 & -6x_2 & -8x_2-12x_4 \\ 12x_1-8x_3 & 6x_3 & 6x_2-8x_4-8x_1 & -8x_3 \\ 6x_2 & -8x_3+6x_1+6x_4 & -8x_2 & 6x_2-16x_4 \end{pmatrix}$$

At the solution $X = \begin{pmatrix} 1 & 1 \\ -1 & -2 \end{pmatrix}$

the Jacobian matrix is $\begin{pmatrix} -0.1 & 0.08 & -0.02 & 0.06 \\ -0.08 & 0.14 & -0.06 & 0.16 \\ 0.2 & 0.06 & 0.14 & 0.08 \\ 0.06 & 0.02 & -0.08 & 0.38 \end{pmatrix}$

and the norm of this matrix is

$$\|J\| = 0.544 \quad \text{to} \quad 3 \text{ S.F.}$$

The iteration converged to this solution whereas it did not converge to the solution

$$X = \begin{pmatrix} -7.729 & 3.518 \\ -6.482 & -7.271 \end{pmatrix} .$$

The norm of the Jacobian matrix at this solution is

$$\|J\| = 3.488 \quad .$$

Another rearrangement of the equation $X^2 + A_1X + A_2 = 0$ is $X = - A_2X^{-1} - A_1$ and the iterative process $X_{i+1} = - A_2X_i^{-1} - A_1$ may be defined.

This process is applied in the following example.

Example 5.2.3.

Consider the equation $X^2 + A_1X + A_2 = 0$

where $A_1 = \begin{pmatrix} -5 & 4 \\ -3 & -6 \end{pmatrix}$ $A_2 = \begin{pmatrix} -4 & -6 \\ -5 & 3 \end{pmatrix}$

Then using the iteration $X_{i+1} = - A_2X_i^{-1} - A_1$ with the initial

matrix $X_o = \begin{pmatrix} 2 & -1 \\ 2 & 3 \end{pmatrix}$ the sequence of matrices obtained is

$$X_1 = \begin{pmatrix} 5 & -2 \\ 5.625 & 5.875 \end{pmatrix}$$

$$X_2 = \begin{pmatrix} 4.7476923 & -3.0646154 \\ 4.1384615 & 5.8769231 \end{pmatrix}$$

$$X_3 = \begin{pmatrix} 4.9673996 & -2.9960576 \\ 4.0299469 & 6.0266111 \end{pmatrix}$$

$$X_4 = \begin{pmatrix} 4.9982567 & -3.005823 \\ 4.0050549 & 6.0018588 \end{pmatrix}$$

$$X_5 = \begin{pmatrix} 4.9994554 & -3.0005824 \\ 3.9996941 & 6.00081701 \end{pmatrix} .$$

The sequence is converging quite rapidly to the matrix $X = \begin{pmatrix} 5 & -3 \\ 4 & 6 \end{pmatrix}$ which is a solution.

It can be shown that $X = \begin{pmatrix} -0.905 & -0.476 \\ -0.381 & 0.905 \end{pmatrix}$ is also a solution where the figures are rounded to 3 decimal places.

Taking an initial matrix $X_o = \begin{pmatrix} -0.9 & -0.5 \\ -0.4 & 0.9 \end{pmatrix}$ which is close to this solution, the sequence of matrices obtained from

the proces $X_{i+1} = - A_2 X_i^{-1} - A_1$ is

$$X_1 = \begin{pmatrix} -0.9405941 & -0.6336634 \\ -0.2673267 & 0.8514851 \end{pmatrix}$$

$$X_2 = \begin{pmatrix} -0.1632651 & -0.7959182 \\ -0.5612245 & -0.1734699 \end{pmatrix}$$

$$X_3 = \begin{pmatrix} -1.3902412 & -9.2682957 \\ 9.0975691 & -4.682927 \end{pmatrix}$$

$$X_4 = \begin{pmatrix} 4.1928037 & -3.6836735 \\ 3.0426962 & 6.556122 \end{pmatrix}$$

$$X_5 = \begin{pmatrix} 5.2059164 & -2.9691271 \\ 4.0830008 & 6.1509156 \end{pmatrix}$$

$$X_6 = \begin{pmatrix} 5.0023925 & -3.0233801 \\ 3.9741637 & 6.0174908 \end{pmatrix}$$

$$X_7 = \begin{pmatrix} 5.0053418 & -3.0002228 \\ 3.997452 & 6.0026052 \end{pmatrix}$$

$$X_8 = \begin{pmatrix} 5.00061156 & -3.0001283 \\ 3.9992161 & 6.00035471 \end{pmatrix} .$$

The reason for convergence to the particular solution

$X = \begin{pmatrix} 5 & -3 \\ 4 & 6 \end{pmatrix}$ can again be seen by writing the iterative process $X_{i+1} = - A_2 X_i^{-1} - A_1$ in the vector form

$$\begin{pmatrix} x_1^{i+1} \\ x_2^{i+1} \\ x_3^{i+1} \\ x_4^{i+1} \end{pmatrix} = \begin{pmatrix} f_1(x_1^i, x_2^i, x_3^i, x_4^i) \\ f_2(x_1^i, x_2^i, x_3^i, x_4^i) \\ f_3(x_1^i, x_2^i, x_3^i, x_4^i) \\ f_4(x_1^i, x_2^i, x_3^i, x_4^i) \end{pmatrix}$$

where $$f_1(x_1,x_2,x_3,x_4) = \frac{5 - 6x_3 + 4x_4}{x_1x_4 - x_2x_3}$$

$$f_2(x_1,x_2,x_3,x_4) = \frac{-4 + 3x_3 + 5x_4}{x_1x_4 - x_2x_3}$$

$$f_3(x_1,x_2,x_3,x_4) = \frac{3 + 3x_3 + 5x_4}{x_1x_4 - x_2x_3}$$

$$f_4(x_1,x_2,x_3,x_4) = \frac{6 - 5x_2 - 3x_1}{x_1x_4 - x_2x_3}$$

and the Jacobian matrix for $\underline{f}(\underline{x})$ is

$$J = \frac{1}{(x_1x_4 - x_2x_3)^2} \begin{pmatrix} -4x_4^2+6x_3x_4 & 6x_3^2+4x_3x_4 & -6x_1x_4+4x_2x_4 & -4x_2x_3+6x_1x_3 \\ 6x_2x_3+4x_2x_4 & -4x_1x_4+6x_1x_3 & -4x_2^2+6x_1x_2 & -6x_1^2+4x_1x_2 \\ -5x_4^2-3x_3x_4 & 3x_3^2+5x_3x_4 & 3x_1x_4+5x_2x_3 & -5x_2x_3-3x_1x_3 \\ 3x_2x_3+5x_2x_4 & -5x_1x_4-3x_1x_3 & -5x_2^2-3x_1x_2 & 3x_1^2+5x_1x_2 \end{pmatrix}$$

At the solution $X = \begin{pmatrix} 5 & -3 \\ 4 & 6 \end{pmatrix}$

the Jacobian matrix is $J = \begin{pmatrix} 0 & 0 & -0.143 & 0.095 \\ 0 & 0 & -0.071 & -0.119 \\ -0.143 & 0.095 & 0 & 0 \\ -0.071 & -0.119 & 0 & 0 \end{pmatrix}$

and the norm of this matrix is $\| J \| = 0.312$ to 3 S.F.

At the solution $X = \begin{pmatrix} -0.905 & -0.476 \\ -0.381 & 0.905 \end{pmatrix}$ with elements

rounded to 3 d.p.,

the Jacobian matrix is $J = \begin{pmatrix} 5.343 & 0.508 & 3.191 & -1.343 \\ 1.343 & -1.207 & 1.678 & -3.191 \\ -3.061 & -1.289 & 1.550 & -1.941 \\ 1.609 & -3.060 & 2.424 & -4.609 \end{pmatrix}$

and the norm of this matrix is $\| J \| = 10.65$.

Again, convergence has taken place to the solution at which $\| J \| < 1$.

This arrangement $X_{i+1} = - A_1 - A_2 X_i^{-1}$ converges to the solution $X = \begin{pmatrix} 5 & -3 \\ 4 & 6 \end{pmatrix}$ even for an initial matrix with very large elements. e.g. if $X_o = \begin{pmatrix} 100 & -40 \\ 200 & 50 \end{pmatrix}$ the sequence of matrices obtained is

$$X_1 = \begin{pmatrix} 4.9230769 & -3.9415385 \\ 3.0653846 & 5.9923077 \end{pmatrix}$$

$$X_2 = \begin{pmatrix} 5.1341157 & -2.9104995 \\ 3.9416772 & 6.1187618 \end{pmatrix}$$

$$X_3 = \begin{pmatrix} 5.0192364 & -3.0102593 \\ 3.9890913 & 5.9801838 \end{pmatrix}$$

$$X_4 = \begin{pmatrix} 4.9996714 & -2.9968518 \\ 3.9962887 & 5.9998474 \end{pmatrix}$$

and is clearly converging to $\begin{pmatrix} 5 & -3 \\ 4 & 6 \end{pmatrix}$.

In this example the rearrangement $X_{i+1} = -A_1 - A_2X_i^{-1}$ only leads to the solution $X = \begin{pmatrix} 5 & -3 \\ 4 & 6 \end{pmatrix}$.

However for this equation the rearrangement

$X_{i+1} = -A_1^{-1}[A_2 + X_i^2]$ gives $\|J\| > 1$ at $X = \begin{pmatrix} 5 & -3 \\ 4 & 6 \end{pmatrix}$

and $\|J\| < 1$ at $X = \begin{pmatrix} -0.905 & -0.476 \\ -0.381 & 0.905 \end{pmatrix}$.

Taking $X_o = \begin{pmatrix} 2 & -1 \\ 2 & 3 \end{pmatrix}$ the iterative function $X_{i+1} = -A_1^{-1}[A_2 + X_i^2]$ gives

$$X_1 = \begin{pmatrix} 0.1904762 & -0.6190476 \\ 0.7380952 & 1.9761905 \end{pmatrix}$$

$$X_2 = \begin{pmatrix} -0.9554044 & -0.4346183 \\ -0.0890968 & 1.2920446 \end{pmatrix}$$

$$X_3 = \begin{pmatrix} -0.9145441 & -0.4296536 \\ -0.3810602 & 0.9995105 \end{pmatrix}$$

$$X_4 = \begin{pmatrix} -0.9078254 & -0.4659061 \\ -0.3848105 & 0.9267439 \end{pmatrix}$$

$$X_5 = \begin{pmatrix} -0.9049649 & -0.4738173 \\ -0.3820642 & 0.9099319 \end{pmatrix}$$

$$X_6 = \begin{pmatrix} -0.9049441 & -0.4756691 \\ -0.3811776 & 0.906002 \end{pmatrix}.$$

This is clearly converging to the other solution. Hence just as in the scalar case two different rearrangements of the equation $X^2 + A_1X + A_2 = 0$ can lead to two different solutions.

The conclusion which may be drawn from these examples is that the method of simple iteration applied to the matrix quadratic equation can lead to solutions under certain conditions.

In the scalar case the iterative process $x^{i+1} = f(x^i)$ will lead to a solution $x = a$ if $|f'(a)| < 1$. In the matrix case it has been shown that a similar condition exists in that the iteration $X_{i+1} = f(X_i)$ can lead to a solution matrix when $\| f'(X) \| < 1$ where the derivative $f'(X)$ is the Jacobian matrix of the set of functions $\underline{f}(\underline{x})$.

In practical terms it would be difficult to ascertain whether the conditions for convergence are satisfied in a particular problem, since the constituent polynomials of $f(X)$ would have to be formed as well as the Jacobian matrix. There is also the problem of

finding a matrix X_o sufficiently close to a solution X to ensure that $\| J_f(\underline{x}) \| < 1$.

The iterative process itself, however, defined on the matrices can be quite straightforward with few computational difficulties.

We have considered two arrangements leading to the iterative processes

$$(1) \quad X_{i+1} = - A_1^{-1}[A_2 + X_i^2]$$

$$(2) \quad X_{i+1} = - A_2 X_i^{-1} - A_1 \quad .$$

Clearly Method (1) cannot be applied if the matrix A_1 is singular. However, provided det $A_1 \neq 0$ then the iterative process will not break down, and if convergence occurs it will be to a root of $X^2 + A_1 X + A_2 = 0$. At each stage of the iterative process two matrix multiplications and one addition are required.

Since Method (2) involves the calculation of X_i^{-1} at each stage the process would break down if one of the iterates was singular.

In Method (1), $\| f'(X) \|$ is small when the determinant of the matrix A_1 is large in comparison to the size of the elements in a solution X, and in Method (2) $\| f'(X) \|$ is small when there is a solution which has a large determinant. However, if the size of the elements in a solution matrix are unknown, the likelihood of the arrangements leading to a solution would be difficult to estimate. In the scalar polynomial equation, the approximate location of a root can be estimated by using the result that there is a root of $f(x) = 0$ in the interval $[a, b]$ if $f(a).f(b) < 0$.

5.3 BERNOULLI'S ALGORITHM.

This algorithm is an iterative method for finding the dominant root of a polynomial equation. The dominant root is defined as the root which has the largest numerical value.

Consider the polynomial equation $x^n + a_1x^{n-1} + a_2x^{n-2} + \ldots + a_n = 0$ and define the values x_i by means of the recurrence relation

$$x_{i+1} + a_1x_i + a_2x_{i-1} + \ldots + a_nx_{i-n+1} = 0 \qquad i = 0,1,2\ldots.$$

then, for suitable starting values, the sequence y_i where $y_i = \frac{x_i}{x_{i-1}}$ converges to the dominant root of the polynomial equation. It can be shown that the conditions for convergence are satisfied by choosing $x_o = 1 \quad x_{-1} = x_{-2} = \ldots x_{-n+1} = 0$.

If the equation to be solved is $x^2 - 5x - 6 = 0$ then defining the recurrence relation

$$x_{i+1} - 5x_i - 6x_{i-1} = 0 \quad \text{with} \quad x_{-1} = 0 \quad x_o = 1$$

then $x_1 = 5, \quad x_2 = 31, \quad x_3 = 185, \quad x_4 = 1111, \quad x_5 = 6665$

$$x_6 = 39991$$

and the sequence y_i becomes

$$y_1 = \frac{x_1}{x_o} = 5$$

$$y_2 = 6.2, \qquad y_3 = 5.9677419, \qquad y_4 = 6.0054054$$

$$y_5 = 5.9990999, \qquad y_6 = 6.00015$$

and the sequence is clearly converging quite rapidly to the root $x = 6$.

The method does not lead to such rapid convergence when the

roots are closer in numerical value, for example, the equation $x^2 - 5x + 6 = 0$ has roots $x = 2$ and $x = 3$,

and defining the relation

$$x_{i+1} - 5x_i + 6x_{i-1} = 0 \quad \text{with} \quad x_{-1} = 0 \quad x_o = 1$$

then $x_1 = 5$, $x_2 = 19$, $x_3 = 65$, $x_4 = 211$, $x_5 = 665$,

$x_6 = 2059$, $x_7 = 6305$, $x_8 = 19171$, $x_9 = 58025$

and the sequence y_i defined by $y_i = \frac{x_i}{x_{i-1}}$ becomes

$y_1 = 5$

$y_2 = 3.8$

$y_3 = 3.4210526$

$y_4 = 3.2461539$

$y_5 = 3.1516588$

$y_6 = 3.0962406$

$y_7 = 3.0621661$

$y_8 = 3.0406027$

$y_9 = 3.026707$.

Though the sequence is obviously converging to the root $x = 3$ the rate of convergence is much slower than in the first example.

Dennis, Traub & Weber [1978] have shown that this algorithm may be applied to the unilateral matrix equation and the sequence will converge to the dominant solvent, provided one exists. A dominant solvent is defined as a solvent with eigenvalues greater in modulus than those of any other solvent.

Consider the matrix equation

$$X^n + A_1X^{n-1} + A_2X^{n-2} + \ldots + A_n = 0$$

and define the matrices X_i by the recurrence relation

$$X_{i+1} + A_1X_i + \ldots + A_nX_{i-n+1} = 0 \quad .$$

Then for suitable starting values the sequence $X_iX_{i-1}^{-1}$ converges to the dominant solvent if one exists. Conditions for convergence are satisfied by choosing

$$X_o = I \qquad X_{-1} = X_{-2} = \ldots \quad X_{-n+1} = 0 \quad .$$

The method is illustrated in the following examples.

Example 5.3.1.

Consider the equation $X^2 + A_1X + A_2 = 0$

where $$A_1 = \begin{pmatrix} -8 & -6 \\ -5 & 5 \end{pmatrix} \qquad A_2 = \begin{pmatrix} 12 & 12 \\ 10 & -14 \end{pmatrix}$$

and define the recurrence relation

$$X_{i+1} + A_1X_i + A_2X_{i-1} = 0$$

with $$X_{-1} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \qquad X_o = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

the sequence of matrices obtained is

$$X_1 = \begin{pmatrix} 8 & 6 \\ 5 & -5 \end{pmatrix} \qquad X_2 = \begin{pmatrix} 82 & 6 \\ 5 & 69 \end{pmatrix}$$

$$X_3 = \begin{pmatrix} 530 & 450 \\ 375 & -445 \end{pmatrix} \qquad X_4 = \begin{pmatrix} 5446 & 30 \\ 25 & 5381 \end{pmatrix} .$$

$$X_5 = \begin{pmatrix} 32858 & 32466 \\ 27055 & -37485 \end{pmatrix} \qquad X_6 = \begin{pmatrix} 359542 & -30114 \\ -25095 & 424789 \end{pmatrix}$$

and $$X_1X_o^{-1} = \begin{pmatrix} 8 & 6 \\ 5 & -5 \end{pmatrix}$$

$$X_2X_1^{-1} = \begin{pmatrix} 6.2857143 & 6.3428571 \\ 5.2857143 & -7.4571429 \end{pmatrix}$$

$$X_3X_2^{-1} = \begin{pmatrix} 6.098081 & 5.9914712 \\ 4.9928927 & -6.8834399 \end{pmatrix}$$

$$X_4X_3^{-1} = \begin{pmatrix} 6.0175976 & 6.0177954 \\ 5.0148295 & -7.020959 \end{pmatrix}$$

$$X_5X_4^{-1} = \begin{pmatrix} 6.005876 & 5.9999672 \\ 4.9999727 & -6.994053 \end{pmatrix}$$

$$X_6X_5^{-1} = \begin{pmatrix} 6.001137 & 6.0009847 \\ 5.0008206 & -7.0009966 \end{pmatrix}$$

The sequence $X_iX_{i-1}^{-1}$ is converging quite rapidly to the solution $X = \begin{pmatrix} 6 & 6 \\ 5 & -7 \end{pmatrix}$.

As in the scalar case, the convergence is not always so rapid as can be seen in the following example.

Example 5.3.2.

Consider the equation $X^3 + A_1X^2 + A_2X + A_3 = 0$

where $A_1 = \begin{pmatrix} -6 & 6 \\ -3 & -15 \end{pmatrix}$ $A_2 = \begin{pmatrix} 2 & -42 \\ 21 & 65 \end{pmatrix}$ $A_3 = \begin{pmatrix} 18 & 66 \\ -33 & -81 \end{pmatrix}$

and define the recurrence relation

$$X_{i+1} + A_1X_i + A_2X_{i-1} + A_3X_{i-2} = 0$$

with $X_{-2} = X_{-1} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$ and $X_o = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

then $X_1 = \begin{pmatrix} 6 & -6 \\ 3 & 15 \end{pmatrix}$

$$X_2 = \begin{pmatrix} 16 & -84 \\ 42 & 142 \end{pmatrix}$$

$$X_3 = \begin{pmatrix} -60 & -780 \\ 390 & 1110 \end{pmatrix}$$

$$X_4 = \begin{pmatrix} -1274 & -6090 \\ 3045 & 7861 \end{pmatrix}$$

$$X_5 = \begin{pmatrix} -12474 & -43386 \\ 21693 & 52605 \end{pmatrix}$$

$$X_6 = \begin{pmatrix} -99224 & -292824 \\ 146412 & 340012 \end{pmatrix}$$

$$X_7 = \begin{bmatrix} -715800 & -1910040 \\ 955020 & 2149260 \end{bmatrix}$$

$$X_8 = \begin{bmatrix} -4884374 & -12180630 \\ 6090315 & 13386571 \end{bmatrix}$$

and the sequence $X_i X_{i-1}^{-1}$ is

$$X_1 X_o^{-1} = \begin{bmatrix} 6 & -6 \\ 3 & 15 \end{bmatrix}$$

$$X_2 X_1^{-1} = \begin{bmatrix} 4.5556 & -3.7778 \\ 1.8889 & 10.2222 \end{bmatrix}$$

$$X_3 X_2^{-1} = \begin{bmatrix} 4.1793 & -3.0207 \\ 1.5103 & 8.7103 \end{bmatrix}$$

$$X_4 X_3^{-1} = \begin{bmatrix} 4.0444 & -2.6444 \\ 1.3222 & 8.0111 \end{bmatrix}$$

$$X_5 X_4^{-1} = \begin{bmatrix} 3.9925 & -2.4261 \\ 1.2131 & 7.6317 \end{bmatrix}$$

$$X_6 X_5^{-1} = \begin{bmatrix} 3.9742 & -2.2888 \\ 1.1444 & 7.4073 \end{bmatrix}$$

$$X_7 X_6^{-1} = \begin{bmatrix} 3.9704 & -2.1982 \\ 1.0991 & 7.2677 \end{bmatrix}$$

$$X_8 X_7^{-1} = \begin{pmatrix} 3.9727 & -2.1368 \\ 1.0684 & 7.1778 \end{pmatrix} .$$

The dominant solvent of this equation is $\begin{pmatrix} 4 & -2 \\ 1 & 7 \end{pmatrix}$ and the sequence is converging to this solution but very slowly.

Dennis, Traub & Weber [1978] describe an algorithm which may be applied when the Bernoulli iteration converges slowly. It is based on an algorithm for scalar polynomials described by Traub [1966].

In the scalar case, the iteration involves the ratio of polynomials of the same degree. The derivative of the polynomial is not required. The algorithm for a scalar polynomial is illustrated first.

Given the polynomial equation $P(x) = 0$ where

$$P(x) = x^n + a_1 x^{n-1} + a_2 x^{n-2} + \ldots + a_n = 0 \quad .$$

A set of polynomials of degree (n-1) are defined recursively as follows

$$G_{i+1}(x) = x\, G_i(x) - b_i\, P(x)$$

where $G_o(x) = x$ and b_i is the leading coefficient of the polynomial $G_i(x)$.

The sequence of leading coefficients of the G polynomials is in fact the sequence x_i obtained from the Bernoulli algorithm and hence the ratio $\frac{b_i}{b_{i-1}}$ converges to the root of $P(x) = 0$ with the largest numerical value.

If however the rate of convergence is slow, this first stage of iteration may be stopped at a certain point say $G_L(x)$ and a second stage iteration defined by

$$x_{i+1} = \frac{G_L(x_i)}{G_{L-1}(x_i)} \qquad \text{with} \qquad x_o = \frac{b_L}{b_{L-1}} \quad .$$

The method is illustrated for the equation $x^2 - 5x + 6 = 0$ in which the Bernoulli iteration converges slowly as shown at the beginning of this section.

Let $P(x) = x^2 - 5x + 6$

Stage one.

Let $G_{i+1}(x) = x\,G_i(x) - b_i[x^2 - 5x + 6]$ and $G_o(x) = x$.

Hence $G_1(x) = x[x] - 1.[x^2 - 5x + 6]$

$\therefore\ G_1(x) = 5x - 6$

$\therefore\ G_2(x) = x[5x - 6] - 5[x^2 - 5x + 6]$

$\therefore\ G_2(x) = 19x - 30$

$G_3(x) = x[19x - 30] - 19[x^2 - 5x + 6]$

$\therefore\ G_3(x) = 65x - 114$

$G_4(x) = x[65x - 114] - 65[x^2 - 5x + 6]$

$\therefore\ G_4(x) = 211x - 390$

$$G_5(x) = x[211x - 390] - 211[x^2 - 5x + 6]$$

$$\therefore \quad G_5(x) = 665x - 1266$$

The sequence of ratios of the leading coefficients of the G polynomials obtained so far is

$$\frac{b_1}{b_o} = \frac{5}{1} = 5$$

$$\frac{b_2}{b_1} = \frac{19}{5} = 3.8$$

$$\frac{b_3}{b_2} = \frac{65}{19} = 3.4210526$$

$$\frac{b_4}{b_3} = \frac{211}{65} = 3.2461539$$

$$\frac{b_5}{b_4} = \frac{665}{211} = 3.1516588$$

The sequence is clearly converging and the second stage iteration is now defined as

$$x_{i+1} = \frac{G_5(x_i)}{G_4(x_i)} \quad \text{with} \quad x_o = \frac{b_5}{b_4} = 3.1516588 \quad .$$

The sequence obtained is

$$x_1 = 3.0176476$$

$$x_2 = 3.0022889$$

$$x_3 = 3.0003008$$

$$x_4 = 3.0000396$$

The sequence is clearly converging much more rapidly than the Bernoulli iteration did in the same example.

A complete description of several iterative processes involving the use of the G polynomials is given by Traub [1966].

The method is applied to the unilateral matrix equation by Dennis, Traub & Weber [1978].

Again the algorithm is in two stages.

Given the unilateral matrix equation $P(X) = 0$

where $P(X) = X^n + A_1X^{n-1} + A_2X^{n-2} + \ldots + A_n$.

Stage one.

A set of matrix polynomials of degree (n-1) are defined recursively as follows

$$G_{i+1}(X) = G_i(X).X - B_i.P(X)$$

where $G_o(X) = X^{n-1}$ and B_i is the leading coefficient matrix in the matrix polynomial $G_i(X)$.

The sequence of leading coefficient matrices B_i is in fact the sequence X_i obtained by application of the Bernoulli algorithm and if there is a dominant solvent the matrix product $B_i\ B_{i-1}^{-1}$ will converge to this solvent.

If the rate of convergence is slow then the second stage of the algorithm may be implemented.

Stage two.

If the first stage has stopped at

$G_L(X)$ then the second stage iteration is defined as

$$X_{i+1} = G_L(X_i).\ G_{L-1}^{-1}(X_i) \quad \text{with} \quad X_o = B_L\ B_{L-1}^{-1} \ .$$

The algorithm is illustrated by application to the matrix equation used in Example 5.3.2 and the faster rate of convergence can be seen in Stage two.

Example 5.3.3.

Consider the equation $X^3 + A_1X^2 + A_2X + A_3 = 0$

where $A_1 = \begin{pmatrix} -6 & 6 \\ -3 & -15 \end{pmatrix}$ $A_2 = \begin{pmatrix} 2 & -42 \\ 21 & 65 \end{pmatrix}$ $A_3 = \begin{pmatrix} 18 & 66 \\ -33 & -81 \end{pmatrix}$

Stage one.

Define $G_{i+1}(X) = G_i(X).X - B_i\ P(X)$

with $G_o(X) = X^2$ and hence $B_o = I$.

Then $G_1(X) = [X^2]X - I[X^3 + A_1X^2 + A_2X + A_3]$

$\therefore\ G_1(X) = - A_1X^2 - A_2X - A_3$

Hence $B_1 = - A_1 = \begin{pmatrix} 6 & -6 \\ 3 & 15 \end{pmatrix}$

$G_2(X) = [- A_1X^2 - A_2X - A_3]X + A_1[X^3 + A_1X^2 + A_2X + A_3]$

$\therefore\ G_2(X) = [A_1^2 - A_2]X^2 + [A_1A_2 - A_3]X + A_1A_3$

$\therefore\ G_2(X) = \begin{pmatrix} 16 & -84 \\ 42 & 142 \end{pmatrix}X^2 + \begin{pmatrix} 96 & 576 \\ -288 & -768 \end{pmatrix}X + \begin{pmatrix} -306 & -882 \\ 441 & 1017 \end{pmatrix}$

Hence $B_2 = \begin{pmatrix} 16 & -84 \\ 42 & 142 \end{pmatrix}$

$$G_3(X) = \left[\begin{bmatrix} 16 & -84 \\ 42 & 142 \end{bmatrix}X^2 + \begin{bmatrix} 96 & 576 \\ -288 & -768 \end{bmatrix}X + \begin{bmatrix} -306 & -882 \\ 441 & 1017 \end{bmatrix}\right]X$$

$$- \begin{bmatrix} 16 & -84 \\ 42 & 142 \end{bmatrix} [X^3 + A_1X^2 + A_2X + A_3]$$

Hence $$G_3(X) = \begin{bmatrix} -60 & -780 \\ 390 & 1110 \end{bmatrix}X^2 + \begin{bmatrix} 1426 & 5250 \\ -2625 & -6449 \end{bmatrix}X + \begin{bmatrix} -3060 & -7860 \\ 3930 & 8730 \end{bmatrix}$$

Hence $$B_3 = \begin{bmatrix} -60 & -780 \\ 390 & 1110 \end{bmatrix}.$$

Two more iterations give

$$G_4(X) = \begin{bmatrix} -1274 & -6090 \\ 3045 & 7861 \end{bmatrix}X^2 + \begin{bmatrix} 13440 & 40320 \\ -20160 & -47040 \end{bmatrix}X + \begin{bmatrix} -24660 & -59220 \\ 29610 & 64170 \end{bmatrix}$$

Hence $$B_4 = \begin{bmatrix} -1274 & -6090 \\ 3045 & 7861 \end{bmatrix}$$

$$G_5(X) = \begin{bmatrix} -12474 & -43386 \\ 21693 & 52605 \end{bmatrix}X^2 + \begin{bmatrix} 105778 & 283122 \\ -141561 & -318905 \end{bmatrix}X + \begin{bmatrix} -178038 & -409206 \\ 204603 & 435771 \end{bmatrix}$$

Hence $$B_5 = \begin{bmatrix} -12474 & -43386 \\ 21693 & 52605 \end{bmatrix}.$$

The sequence $B_i\ B_{i-1}^{-1}$ obtained so far is

$$B_1B_o^{-1} = \begin{bmatrix} 6 & -6 \\ 3 & 15 \end{bmatrix}$$

$$B_2B_1^{-1} = \begin{pmatrix} 4.5556 & -3.7778 \\ 1.8889 & 10.2222 \end{pmatrix}$$

$$B_3B_2^{-1} = \begin{pmatrix} 4.1793 & -3.0207 \\ 1.5103 & 8.7103 \end{pmatrix}$$

$$B_4B_3^{-1} = \begin{pmatrix} 4.0444 & -2.6444 \\ 1.3222 & 8.0111 \end{pmatrix}$$

$$B_5B_4^{-1} = \begin{pmatrix} 3.9925 & -2.4261 \\ 1.2131 & 7.6317 \end{pmatrix} .$$

This is precisely the sequence of matrices obtained by the Bernoulli iteration and is clearly converging but slowly. The second stage is therefore implemented.

Stage two.

Define the iteration by

$$X_{i+1} = G_5(X_i)\, G_4^{-1}(X_i) \quad \text{with} \quad X_o = B_5B_4^{-1} .$$

The sequence of matrices obtained now is

$$X_1 = \begin{pmatrix} 3.973 & -2.089 \\ 1.045 & 7.107 \end{pmatrix}$$

$$X_2 = \begin{pmatrix} 3.993 & -2.018 \\ 1.009 & 7.019 \end{pmatrix}$$

$$X_3 = \begin{pmatrix} 3.999 & -2.003 \\ 1.002 & 7.004 \end{pmatrix}$$

$$X_4 = \begin{pmatrix} 3.999 & -2.001 \\ 1.000 & 7.001 \end{pmatrix}$$

$$X_5 = \begin{pmatrix} 4.000 & -2.000 \\ 1.000 & 7.001 \end{pmatrix}$$

and $S_1 = \begin{pmatrix} 4 & -2 \\ 1 & 7 \end{pmatrix}$ is the dominant solvent of the matrix equation $X^3 + A_1X^2 + A_2X + A_3 = 0$.

The faster rate of convergence can clearly be seen in the Stage two iteration.

The stage one algorithm suffers from the same disadvantage as the Bernoulli iteration in that the elements of the coefficient matrices in the G polynomials become very large and problem of inaccuracy may arise when inverses have to be determined.

5.4 THE NEWTON METHOD APPLIED TO THE MATRIX EQUATION.

Given a polynomial in a single variable

$$f(x) = a_o x^n + a_1 x^{n-1} + ... + a_n$$

then the Newton-Raphson iteration applied to the equation $f(x) = 0$ is

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)} \quad .$$

This can also be applied to a set of n equations in n variables

$$f_1(x_1,x_2,\ldots, x_n) = 0$$

$$f_2(x_1,x_2,\ldots, x_n) = 0$$

$$\ldots\ldots\ldots\ldots\ldots\ldots$$

$$f_n(x_1,x_2,\ldots, x_n) = 0$$

The set of equations is written in vector form as $\underline{f}(\underline{x}) = 0$

where $\underline{x} = (x_1,x_2,\ldots, x_n)$

and $\underline{f}(\underline{x}) = (f_1(\underline{x}),f_2(\underline{x}),\ldots,f_n(\underline{x}))$

and the derivative $\underline{f}'(\underline{x})$ becomes the Jacobian matrix

$$J_{\underline{x}} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \cdots & \frac{\partial f_n}{\partial x_n} \end{pmatrix}$$

Then the Newton method applied to the set of equations is

$$\underline{x}^{(i+1)} = \underline{x}^{(i)} - J^{-1}_{\underline{x}^{(i)}} \underline{f}(\underline{x}^{(i)}) \quad .$$

Newton's method can be shown under certain conditions to have quadratic convergence, provided a good initial approximation is available.

Since any matrix equation involving m×m matrices may be expressed as a set of m^2 polynomial equations in m^2 unknowns,

the Newton method may be applied to the system of equations to find the elements of the solution matrix. This can be illustrated by consideration of the unilateral quadratic matrix equation involving 2×2 matrices.

Consider the equation $X^2 + A_1X + A_2 = 0$

where $X = \begin{pmatrix} x_1 & x_2 \\ x_3 & x_4 \end{pmatrix}$ $A_1 = \begin{pmatrix} a_1 & a_2 \\ a_3 & a_4 \end{pmatrix}$ $A_2 = \begin{pmatrix} b_1 & b_2 \\ b_3 & b_4 \end{pmatrix}$

then the matrix equation becomes

$$\begin{pmatrix} x_1^2+x_2x_3+a_1x_1+a_2x_3+b_1 & x_1x_2+x_2x_4+a_1x_2+a_2x_4+b_2 \\ x_1x_3+x_3x_4+a_3x_1+a_4x_3+b_3 & x_2x_3+x_4^2+a_3x_2+a_4x_4+b_4 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

or $\begin{pmatrix} f_1(\underline{x}) & f_2(\underline{x}) \\ f_3(\underline{x}) & f_4(\underline{x}) \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$ where $\underline{x} = (x_1,x_2,x_3,x_4)$.

The Newton method may then be applied to the system $\underline{f}(\underline{x}) = 0$

where $\underline{f}(\underline{x}) = [f_1(\underline{x}), f_2(\underline{x}), f_3(\underline{x}), f_4(\underline{x})]$.

The Jacobian matrix for $\underline{f}(\underline{x})$ is

$$\begin{pmatrix} 2x_1+a_1 & x_3 & x_2+a_2 & 0 \\ x_2 & x_1+x_4+a_1 & 0 & x_2+a_2 \\ x_3+a_3 & 0 & x_1+x_4+a_4 & x_3 \\ 0 & x_3+a_3 & x_2 & 2x_4+a_4 \end{pmatrix}$$

The method is illustrated in the following example.

Example 5.4.1.

Consider the matrix equation $X^2 + A_1X + A_2 = 0$

where $X = \begin{pmatrix} x_1 & x_2 \\ x_3 & x_4 \end{pmatrix}$ $A_1 = \begin{pmatrix} 2 & 1 \\ -1 & 1 \end{pmatrix}$ $A_2 = \begin{pmatrix} -3 & 3 \\ -9 & -11 \end{pmatrix}$

then $J_{\underline{x}} = \begin{pmatrix} 2x_1+2 & x_3 & x_2+1 & 0 \\ x_2 & x_1+x_4+2 & 0 & x_2+1 \\ x_3-1 & 0 & x_1+x_4+1 & x_3 \\ 0 & x_3-1 & x_2 & 2x_4+1 \end{pmatrix}$.

Choosing the initial value $\underline{x}^o = \begin{pmatrix} 0 \\ -2 \\ 1 \\ 2 \end{pmatrix}$

then $J_{\underline{x}^{(o)}} = \begin{pmatrix} 2 & 1 & -1 & 0 \\ -2 & 4 & 0 & -1 \\ 0 & 0 & 3 & 1 \\ 0 & 0 & -2 & 5 \end{pmatrix}$

$\underline{f}(\underline{x}^{(o)} = \begin{pmatrix} -4 \\ -3 \\ -6 \\ -5 \end{pmatrix}$

and $\underline{x}^{(1)} = \begin{pmatrix} 0 \\ -2 \\ 1 \\ 2 \end{pmatrix} - \begin{pmatrix} 2 & 1 & -1 & 0 \\ -2 & 4 & 0 & -1 \\ 0 & 0 & 3 & 1 \\ 0 & 0 & -2 & 5 \end{pmatrix}^{-1} \begin{pmatrix} -4 \\ -3 \\ -6 \\ -5 \end{pmatrix}$

158

It is usually more convenient to write $\underline{x}^{(1)} = \begin{pmatrix} x_1^1 \\ x_2^1 \\ x_3^1 \\ x_4^1 \end{pmatrix}$

and to solve the linear equations obtained from

$$\begin{pmatrix} 2 & 1 & -1 & 0 \\ -2 & 4 & 0 & -1 \\ 0 & 0 & 3 & 1 \\ 0 & 0 & -2 & 5 \end{pmatrix} \begin{pmatrix} -x_1^1 \\ -2-x_2^1 \\ 1-x_3^1 \\ 2-x_4^1 \end{pmatrix} = \begin{pmatrix} -4 \\ -3 \\ -6 \\ -5 \end{pmatrix}$$

rather than calculating the inverse of the Jacobian matrix.

Solving the linear equations gives $\underline{x}^{(1)} = \begin{pmatrix} 1.729 \\ 0.012 \\ 2.471 \\ 3.588 \end{pmatrix}$

where the values are given to 3 decimal places.

Similarly, further iterations give

$$\underline{x}^{(2)} = \begin{pmatrix} 1.098 \\ -0.828 \\ 2.048 \\ 3.072 \end{pmatrix} \quad \underline{x}^{(3)} = \begin{pmatrix} 1.002 \\ -0.995 \\ 1.996 \\ 3.001 \end{pmatrix} \quad \underline{x}^{(4)} = \begin{pmatrix} 1.000 \\ -0.998 \\ 2.001 \\ 3.000 \end{pmatrix}$$

$$\underline{x}^{(5)} = \begin{pmatrix} 1.000 \\ -1.000 \\ 2.000 \\ 3.000 \end{pmatrix}.$$

Hence the iteration has converged to the vector solution $\begin{pmatrix} 1 \\ -1 \\ 2 \\ 3 \end{pmatrix}$ and the elements of the matrix solution have been formed.

The solution of the matrix equation is $X = \begin{pmatrix} 1 & -1 \\ 2 & 3 \end{pmatrix}$.

In the following example the method is applied to a cubic matrix equation.

Example 5.4.2.

Consider the equation $X^3 - A = 0$

where $X = \begin{pmatrix} x_1 & x_2 \\ x_3 & x_4 \end{pmatrix}$ $A = \begin{pmatrix} -1 & 3 \\ 0 & 8 \end{pmatrix}$.

The matrix equation is equivalent to $\begin{pmatrix} f_1(\underline{x}) & f_2(\underline{x}) \\ f_3(\underline{x}) & f_4(\underline{x}) \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$

where $\underline{x} = (x_1\ x_2\ x_3\ x_4)$

and

$$f_1(x_1x_2x_3x_4) = x_1^3 + 2x_1x_2x_3 + x_2x_3x_4 + 1$$

$$f_2(x_1x_2x_3x_4) = x_1^2x_2 + x_2^2x_3 + x_1x_2x_4 + x_2x_4^2 - 3$$

$$f_3(x_1x_2x_3x_4) = x_1^2x_3 + x_1x_3x_4 + x_2x_3^2 + x_3x_4^2$$

$$f_4(x_1x_2x_3x_4) = x_1x_2x_3 + 2x_2x_3x_4 + x_4^3 - 8 \quad .$$

The Jacobian matrix for $\underline{f}(\underline{x})$ is

$$J_{\underline{x}} = \begin{pmatrix} 3x_1^2+2x_2x_3 & 2x_1x_3+x_3x_4 & 2x_1x_2+x_2x_4 & x_2x_3 \\ 2x_1x_2+x_2x_4 & x_1^2+2x_2x_3+x_1x_4+x_4^2 & x_2^2 & x_1x_2+2x_2x_4 \\ 2x_1x_3+x_3x_4 & x_3^2 & x_1^2+x_1x_4+2x_2x_3+x_4^2 & x_1x_3+2x_3x_4 \\ x_2x_3 & x_1x_3+2x_3x_4 & x_1x_2+2x_2x_4 & 2x_2x_3+3x_4^2 \end{pmatrix}.$$

Choosing the initial value $\underline{x}^{(o)} = \begin{pmatrix} -1 \\ 1 \\ 0 \\ 1 \end{pmatrix}$ and setting $\underline{x}^{(1)} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}$

then $J_{\underline{x}^{(o)}} = \begin{pmatrix} 3 & 0 & -1 & 0 \\ -1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 3 \end{pmatrix}$ and $\underline{f}(\underline{x}^{(o)}) = \begin{pmatrix} 0 \\ -2 \\ 0 \\ 7 \end{pmatrix}$.

The Newton iteration $J_{\underline{x}^{(o)}}[\underline{x}^{o} - \underline{x}^{(1)}] = f(\underline{x}^{(o)})$

becomes $\begin{pmatrix} 3 & 0 & -1 & 0 \\ -1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 3 \end{pmatrix} \begin{pmatrix} -1-x_1 \\ 1-x_2 \\ -x_3 \\ 1-x_4 \end{pmatrix} = \begin{pmatrix} 0 \\ -2 \\ 0 \\ 7 \end{pmatrix}$.

Solving the linear equations gives $\underline{x}^{(1)} = \begin{pmatrix} -1 \\ \frac{2}{3} \\ 0 \\ 3\frac{1}{3} \end{pmatrix}$.

Further iterations, with the elements given to 2 decimal places are

$$\underline{x}^{(2)} = \begin{pmatrix} -1.00 \\ 0.71 \\ 0 \\ 2.48 \end{pmatrix} \qquad \underline{x}^{(3)} = \begin{pmatrix} -1.00 \\ 0.88 \\ 0 \\ 2.10 \end{pmatrix} \qquad \underline{x}^{(4)} = \begin{pmatrix} -1.00 \\ 1.48 \\ 0 \\ 2.00 \end{pmatrix}$$

$$\underline{x}^{(5)} = \begin{pmatrix} -1.00 \\ 1.00 \\ 0 \\ 2.00 \end{pmatrix} .$$

The iteration has converged to a solution vector $\underline{x} = \begin{pmatrix} -1 \\ 1 \\ 0 \\ 2 \end{pmatrix}$

and hence a solution of the matrix equation $X^3 - A = 0$ has the same elements and is

$$X = \begin{pmatrix} -1 & 1 \\ 0 & 2 \end{pmatrix}$$

The Newton Method may also be applied direct to the matrix equation by use of the derivative operator as described in 5.1.

Consider the equation $X^2 + A_1X + A_2 = 0$.

Let $f(X) = X^2 + A_1X + A_2$.

Then the derivative operator is

$$f'(X) = [X + A_1][\ ] + [\ ]X$$

where the square brackets are replaced by the matrix which the derivative operator is operating on.

The Newton iteration is

$$X_{i+1} = X_i - [f'(X_i)]^{-1} f(X_i), \quad f'(X_i) \text{ nonsingular.}$$

This may be rearranged as

$$f'(X_i)[X_i - X_{i+1}] = f(X_i) \quad .$$

The derivative operator is therefore operating on the matrix $[X_i - X_{i+1}]$ and the iteration becomes

$$[X_i+A_1][X_i-X_{i+1}] + [X_i-X_{i+1}][X_i] = X_i^2 + A_1X_i + A_2$$

and this simplifies to

$$[X_i+A_1][X_{i+1}] + [X_{i+1}][X_i] = X_i^2 - A_2 \quad .$$

The Newton method has therefore reduced the quadratic matrix equation to the problem of the iterative solution of a linear matrix equation of the Liapunov type.

A matrix equation of any degree can be converted to the iterative solution of a linear equation, e.g. for the cubic matrix equation

$$X^3 + A_1X^2 + A_2X + A_3 = 0 \quad .$$

The Newton iteration is

$$[X_i^2+A_1X_i+A_2]X_{i+1} + X_iX_{i+1}X_i + [X_{i+1}X_i+A_1X_{i+1}]X_i = 2X_i^3 + A_1X_i^2 - A_3$$

which is linear in X_{i+1}.

The Riccati equation XEX + DX + XF + G = 0 may also be solved by this method.

The Newton iteration for this equation is

$$[X_iE+D]X_{i+1} + X_{i+1}[EX_i+F] = X_iEX_i - G \quad .$$

Kleinman [1968] and Davis [1981] have derived iterative methods which are based on the Newton iteration.

One of the main difficulties in applying Newton's method to the matrix equation is in finding a suitable initial matrix for convergence to take place.

Kleinman's scheme is a method for the solution of the special form of the Riccati equation

$$PBR^{-1}B^TP - A^TP - PA - Q = 0$$

where P is the unknown matrix to be determined.

The iterative process is

$$A_i^TP_i + P_iA_i = - Q - P_{i-1}BR^{-1}B^TP_{i-1}$$

where $A_i = A - BR^{-1}B^TP_{i-1} \qquad i = 1,2,...$

and $A_o = A - BL_o$ and L_o is any matrix such that $A - BL_o$ is a stability matrix, that is $(A - BL_o)$ has eigenvalues with negative real parts.

Kleinman shows that $\lim_{i\to\infty} P_i = P$ and that the sequence $\{P_i\}$ is quadratically convergent.

Davis [1981] applied Newton's Method to the solution of the quadratic matrix equation $F(X) = A_oX^2 + A_1X + A_2 = 0$. He describes an algorithm using a subroutine SQUINT which stands for Solving the Quadratic by Iterating Newton Triangularizations. After an initial guess X_o is chosen, successive iterates are generated by the formula

$$X_{i+1} = X_i - T_i$$

where $T_i = [F'(X_i)]^{-1} F(X_i)$ $\quad i = 0,1,2,\ldots$

and T_i is given as the solution of the system

$$[A_o X_i + A_1]T_i + A_o T_i X_i = F(X_i) \qquad i = 0,1,2,\ldots.$$

The SQUINT subroutine obtains the T_i by simultaneously reducing $[A_o X_i + A_1]$ and A_o to upper triangular form then reducing X_i to lower triangular form. The transformed triangular system is solved and T_i is computed.

One of the problems of applying Newton's Method to the matrix equation is that of finding a suitable initial guess. Davis suggests

$$X_o = \left[\frac{\|A_1\| + \sqrt{\|A_1\|^2 + 4\|A_o\|\|A_2\|}}{2\|A_o\|} \right] I \quad .$$

This is designed to provide a rough estimate for the magnitude of a possible solution, but convergence is not guaranteed.

Several examples are now given showing a variety of matrix equations solved by the Newton method. Where successive iterates are given, the elements are rounded to 3 decimal places. In some examples, the solutions were obtained by use of programs written for a micro computer. In these cases, convergence was deemed to have occurred when each element in $f(X_i)$ had a numerical value less than 10^{-8}.

Example 5.4.3.

Consider the equation $X^2 + A_1X + A_2 = 0$

where $A_1 = \begin{pmatrix} 2 & 1 \\ -1 & 1 \end{pmatrix}$ $A_2 = \begin{pmatrix} -3 & 3 \\ -9 & -11 \end{pmatrix}$.

The Newton iteration is $[X_i + A_1]X_{i+1} + X_{i+1}X_i = X_i^2 - A_2$.

Choosing $X_o = \begin{pmatrix} 0 & -2 \\ 1 & 2 \end{pmatrix}$ the sequence of iterates obtained is

$$x_1 = \begin{pmatrix} 1.729 & 0.012 \\ 2.471 & 3.588 \end{pmatrix} \qquad X_2 = \begin{pmatrix} 1.098 & -0.828 \\ 2.048 & 3.072 \end{pmatrix}$$

$$X_3 = \begin{pmatrix} 1.002 & -0.995 \\ 1.996 & 3.001 \end{pmatrix} \qquad X_4 = \begin{pmatrix} 1.000 & -1.000 \\ 2.000 & 3.000 \end{pmatrix}$$

and the iteration has converged to $\begin{pmatrix} 1 & -1 \\ 2 & 3 \end{pmatrix}$ which is a

solution of $X^2 + A_1X + A_2 = 0$.

Example 5.4.4.

Consider the Riccati Equation $XEX + DX + XF + G = 0$

where $E = \begin{pmatrix} -1 & 0 \\ 2 & 1 \end{pmatrix}$ $D = \begin{pmatrix} 1 & 2 \\ -2 & 1 \end{pmatrix}$ $F = \begin{pmatrix} 1 & -2 \\ 2 & 1 \end{pmatrix}$

$$G = \begin{pmatrix} 1 & -3 \\ -11 & -10 \end{pmatrix}.$$

The Newton iteration is $[X_iE + D]X_{i+1} + X_{i+1}[EX_i + F] = X_iEX_i - G$.

166

Choosing $X_o = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

then

$$X_1 = \begin{pmatrix} 0.750 & -0.750 \\ 0.500 & 3.000 \end{pmatrix} \qquad X_2 = \begin{pmatrix} 0.038 & -1.070 \\ 1.003 & 3.057 \end{pmatrix}$$

$$X_3 = \begin{pmatrix} -0.010 & -1.005 \\ 1.006 & 3.003 \end{pmatrix} \qquad X_4 = \begin{pmatrix} 0.000 & -1.000 \\ 1.000 & 3.000 \end{pmatrix}.$$

The iteration has converged to $\begin{pmatrix} 0 & -1 \\ 1 & 3 \end{pmatrix}$ which is a solution of $XEX + DX + XF + G = 0$.

Example 5.4.5.

Consider the cubic equation $X^3 + A_1X^2 + A_2X + A_3 = 0$

where $A_1 = \begin{pmatrix} -6 & 6 \\ -3 & -15 \end{pmatrix}$ $A_2 = \begin{pmatrix} 2 & -42 \\ 21 & 65 \end{pmatrix}$ $A_3 = \begin{pmatrix} 18 & 66 \\ -33 & -81 \end{pmatrix}$.

The Newton iteration is

$$[X_i^2+A_1X_i+A_2]X_{i+1} + [X_i+A_1]X_{i+1}X_i + X_{i+1}X_i^2 = 2X_i^3 + A_1X_i^2 - A_3 .$$

Choosing $X_o = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}$

gives

$$X_1 = \begin{pmatrix} -0.102 & -1.801 \\ 0.861 & 2.053 \end{pmatrix} \qquad X_2 = \begin{pmatrix} -0.096 & -2.154 \\ 0.943 & 2.744 \end{pmatrix}$$

$$X_3 = \begin{pmatrix} -0.059 & -2.130 \\ 1.005 & 3.002 \end{pmatrix} \qquad X_4 = \begin{pmatrix} -0.006 & -2.013 \\ 1.002 & 3.004 \end{pmatrix}$$

$$X_5 = \begin{pmatrix} 0.000 & -2.000 \\ 1.000 & 3.000 \end{pmatrix} .$$

The iteration has converged to $\begin{pmatrix} 0 & -2 \\ 1 & 3 \end{pmatrix}$ which is a solution of the equation $X^3 + A_1X^2 + A_2X + A_3 = 0$.

Example 5.4.6.

Consider the equation $X^2 + A_1X + A_2 = 0$

where $A_1 = \begin{pmatrix} 2 & 1 & -3 \\ -1 & -1 & 0 \\ 2 & 1 & -2 \end{pmatrix}$ $A_2 = \begin{pmatrix} -4 & -2 & -5 \\ 2 & -2 & -2 \\ 2 & 0 & -17 \end{pmatrix}$.

The Newton iteration is $[X_i + A_1]X_{i+1} + X_{i+1}X_i = X_i^2 - A_2$.

Choosing $X_o = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$

the iteration converges in seven iterations to

$$X = \begin{pmatrix} 1.098 & -0.064 & 2.376 \\ -0.055 & 2.077 & 0.753 \\ -1.039 & -0.404 & 4.908 \end{pmatrix}$$

where the elements have been rounded to 3 decimal places.

In this example, slight changes in the choice of X_o produce iterative sequences which converge to different solutions. In each case convergence is deemed to have occurred when each element of $[X_i^2 + A_1X_i + A_2]$ has a numerical value less than 10^{-8}. The solution is given to 3 decimal places in each case.

Choosing $X_o = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ the iteration converges to

$$X = \begin{pmatrix} -1.097 & 0.903 & 3.859 \\ 1.565 & -0.534 & 0.369 \\ 2.344 & 2.344 & 0.948 \end{pmatrix}$$ in 16 iterations

$X_o = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$ converges to $\begin{pmatrix} -0.881 & 1.119 & 3.590 \\ 2.532 & 0.532 & -0.834 \\ -0.642 & -0.642 & 4.664 \end{pmatrix}$ in 30 iterations

$X_o = \begin{pmatrix} 4 & -1 & 0 \\ 2 & -1 & 3 \\ 0 & 0 & 2 \end{pmatrix}$ converges to $\begin{pmatrix} -1.066 & -1.305 & 8.201 \\ 2.449 & 3.513 & -5.987 \\ 1.104 & 0.824 & -0.862 \end{pmatrix}$ in 35 iterations

$X_o = \begin{pmatrix} -3 & 0 & 1 \\ 0 & -1 & 1 \\ -2 & -2 & -3 \end{pmatrix}$ converges to $\begin{pmatrix} -3.050 & -1.050 & 1.029 \\ 1.042 & -0.958 & -0.389 \\ -0.432 & -0.432 & -3.075 \end{pmatrix}$ in 5 iterations.

$X_o = \begin{pmatrix} -10 & -5 & -1 \\ 20 & -4 & 6 \\ -30 & 4 & 20 \end{pmatrix}$ converges to $\begin{pmatrix} -0.731 & 1.269 & 3.402 \\ 1.663 & -0.337 & 0.247 \\ 2.865 & 2.865 & 0.299 \end{pmatrix}$ in 14 iterations.

Example 5.4.7.

Consider the equation $X^2A_1X + A_2X + A_3 = 0$

where $A_1 = \begin{pmatrix} 0 & 1 \\ -1 & 2 \end{pmatrix}$ $A_2 = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$ and $A_3 = \begin{pmatrix} 0 & -1 \\ -2 & 0 \end{pmatrix}$.

The Newton iteration is

$$[X_i^2A_1+A_2]X_{i+1} + X_iX_{i+1}A_1X_i + X_{i+1}X_iA_1X_i = 2X_i^2A_1X_i - A_3$$

$X_o = \begin{pmatrix} 1 & -1 \\ 1 & 0 \end{pmatrix}$ converges to $\begin{pmatrix} -1.039 & -2.039 \\ -0.638 & 0.362 \end{pmatrix}$ in 14 iterations

$X_o = \begin{pmatrix} 1 & 2 \\ -1 & -1 \end{pmatrix}$ converges to $\begin{pmatrix} 1.786 & 0.183 \\ -0.330 & 0.262 \end{pmatrix}$ in 20 iterations

$X_o = \begin{pmatrix} 3 & 5 \\ -1 & -2 \end{pmatrix}$ converges to $\begin{pmatrix} -1.039 & 2.039 \\ -0.638 & 0.362 \end{pmatrix}$ in 20 iterations

$X_o = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$ converges to $\begin{pmatrix} 1.786 & 0.183 \\ -0.330 & 0.262 \end{pmatrix}$ in 24 iterations

$X_o = \begin{pmatrix} 2 & -3 \\ 1 & -1 \end{pmatrix}$ and $X_o = \begin{pmatrix} 2 & 5 \\ -1 & -2 \end{pmatrix}$ do not converge .

The number of iterations required for convergence, when it occurs, depends upon the "distance" of an initial approximation from a solution matrix where "distance" is measured in terms of a matrix norm.

This is illustrated in example 5.4.7 where the initial values $X_o^A = \begin{pmatrix} 1 & -1 \\ 1 & 0 \end{pmatrix}$ and $X_o^B = \begin{pmatrix} 3 & 5 \\ -1 & -2 \end{pmatrix}$ converge to the solution

$$X_1 = \begin{pmatrix} -1.039 & -2.039 \\ -0.638 & 0.362 \end{pmatrix}$$ in 14 iteration and 20 iterations

respectively.

$$X_1 - X_o^A = \begin{pmatrix} -2.039 & -1.039 \\ -1.638 & 0.362 \end{pmatrix} \text{ and } \| X_1 - X_o^A \| = 3.078$$

$$X_1 - X_o^B = \begin{pmatrix} -4.039 & -2.961 \\ 0.362 & 2.362 \end{pmatrix} \text{ and } \| X_1 - X_o^B \| = 7$$

As would be expected, the two versions of the Newton Method are in fact equivalent. Using the derivative operator the Newton iteration can be rearranged as

$$f'(X_i)[X_i - X_{i+1}] = f(X_i) \quad .$$

If $f(X) = X^2 + A_1X + A_2$ then the derivative operator

$f'(X_i) = [X_i + A_1][\ ] + [\ ]X_i$ operating on the matrix $[X_i - X_{i+1}]$ can be written as

$$[(X_i + A_1) \circledast I_m + I_m \circledast X_i^T](\underline{x}^{(i)} - \underline{x}^{(i+1)})$$

where $(\underline{x}^{(i)} - \underline{x}^{(i+1)})$ is the column vector formed from the transposed rows of the matrix $(X_i - X_{i+1})$.

Since $$f(X_i) = \begin{pmatrix} f_{11}(\underline{x}) & f_{12}(\underline{x}) & \cdots & f_{1m}(\underline{x}) \\ f_{21}(\underline{x}) & f_{22}(\underline{x}) & \cdots & f_{2m}(\underline{x}) \\ \cdots & \cdots & \cdots & \cdots \\ f_{m1}(\underline{x}) & f_{m2}(\underline{x}) & \cdots & f_{mm}(\underline{x}) \end{pmatrix}$$

where $f_{ij}(\underline{x}) = 0 \quad i = 1,\ldots,m \quad j = 1,\ldots,m$ are simply the constituent equations of the matrix equation, then the Newton iteration can be written as

$$[(X_i + A_1) \otimes I_m + I_m \otimes X_i^T](\underline{x}^{(i)} - \underline{x}^{(i+1)}) = \underline{f}(\underline{x}^i) \quad .$$

Since the Newton iteration using the Jacobian matrix can be written as

$$J_{\underline{x}^{(i)}}[\underline{x}^{(i)} - \underline{x}^{(i+1)}] = \underline{f}(\underline{x}^{(i)})$$

it only remains to show that $[(X + A_1) \otimes I_m + I_m \otimes X^T]$ and the Jacobian matrix are the same.

Let $X = (x_{ij}) \qquad A_1 = (a_{ij})$ and $A_2 = (b_{ij})$

then the equation $X^2 + A_1X + A_2 = 0$ can be written as

$$\begin{pmatrix} f_{11}(\underline{x}) & f_{12}(\underline{x}) & \ldots\ldots & f_{1m}(\underline{x}) \\ f_{21}(\underline{x}) & f_{22}(\underline{x}) & \ldots\ldots & f_{2m}(\underline{x}) \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ f_{m1}(\underline{x}) & f_{m2}(\underline{x}) & \ldots\ldots & f_{mm}(\underline{x}) \end{pmatrix} = 0$$

where $\underline{x} = (x_{11}, x_{12}, \ldots, x_{mm})$

and $$f_{rs}(\underline{x}) = \sum_{j=1}^{m} (x_{rj}\, x_{js} + a_{rj}\, x_{js} + b_{rs}) \quad .$$

The Jacobian matrix for the set of functions

$(f_{11}(\underline{x}), f_{12}(\underline{x}) \ldots\ldots f_{mm}(\underline{x}))$ is

$$J_{\underline{x}} = \begin{pmatrix} \frac{\partial f_{11}}{\partial x_{11}}(\underline{x}) & \frac{\partial f_{11}}{\partial x_{12}}(\underline{x}) & \cdots\cdots & \frac{\partial f_{11}}{\partial x_{mm}}(\underline{x}) \\ \frac{\partial f_{12}}{\partial x_{11}}(\underline{x}) & \frac{\partial f_{12}}{\partial x_{12}}(\underline{x}) & \cdots\cdots & \frac{\partial f_{12}}{\partial x_{mm}}(\underline{x}) \\ \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ \frac{\partial f_{mm}}{\partial x_{11}}(\underline{x}) & \frac{\partial f_{mm}}{\partial x_{12}}(\underline{x}) & \cdots\cdots & \frac{\partial f_{mm}}{\partial x_{mm}}(\underline{x}) \end{pmatrix}$$

$$\text{where } \frac{\partial f_{rs}}{\partial x_{jk}} = \begin{cases} x_{rj} + a_{rj} & \text{if } s = k \\ x_{ks} & \text{if } r = j \\ x_{jj} + x_{kk} + a_{jj} & \text{if } r = j \text{ and } s = k \\ 0 & \text{otherwise} \end{cases}$$

$$\therefore J_{\underline{x}} = \begin{pmatrix} X^T & 0 & \cdots & 0 \\ 0 & X^T & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & X^T \end{pmatrix} + \begin{pmatrix} x_{11}I_m & x_{12}I_m & \cdots & x_{1m}I_m \\ x_{21}I_m & x_{22}I_m & \cdots & x_{2m}I_m \\ \cdots & \cdots & \cdots & \cdots \\ x_{m1}I_m & x_{m2}I_m & \cdots & x_{mm}I_m \end{pmatrix} +$$

$$+ \begin{pmatrix} a_{11}I_m & a_{12}I_m & \cdots & a_{1m}I_m \\ a_{21}I_m & a_{22}I_m & \cdots & a_{2m}I_m \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1}I_m & a_{m2}I_m & \cdots & a_{mm}I_m \end{pmatrix}$$

$$= I_m \otimes X^T + X \otimes I_m + A_1 \otimes I_m$$

$$= [(X + A_1) \otimes I_m + I_m \otimes X^T]$$

Hence the Jacobian $J_{\underline{x}}$ is equivalent to the matrix $[(X + A_1) \textcircled{X} I_m + I_m \textcircled{X} X^T]$ and the two versions are equivalent.

Though the two versions are equivalent, the amount of computation involved in using the derivative operator is much less as can be seen in the following example.

Example 5.4.8.

Consider the equation $X^2 + A_1X + A_2 = 0$

where $A_1 = \begin{pmatrix} 1 & 2 \\ -1 & 0 \end{pmatrix}$ $A_2 = \begin{pmatrix} -9 & 0 \\ 1 & -1 \end{pmatrix}$.

In applying the Newton iteration using the Jacobian matrix the constituent equations must first be derived.

$$f_1(\underline{x}) = x_1^2 + x_2x_3 + x_1 + 2x_3 - 9$$

$$f_2(\underline{x}) = x_1x_2 + x_2x_4 + x_2 + 2x_4$$

$$f_3(\underline{x}) = x_3x_1 + x_3x_4 - x_1 + 1$$

$$f_4(\underline{x}) = x_4^2 + x_2x_3 - x_2 - 1 \quad .$$

The Jacobian matrix is

$$J_{\underline{x}} = \begin{pmatrix} 2x_1+1 & x_3 & x_2+2 & 0 \\ x_2 & x_1+x_4+1 & 0 & x_2+2 \\ x_3-1 & 0 & x_1+x_4 & x_3 \\ 0 & x_3-1 & x_2 & 2x_4 \end{pmatrix}$$

174

Let $\underline{x}_o = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \end{pmatrix}$ then $J_{\underline{x}_o} = \begin{pmatrix} 3 & 1 & 3 & 0 \\ 1 & 2 & 0 & 3 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$

and $\underline{f}(\underline{x}_o) = \begin{pmatrix} -4 \\ 2 \\ 1 \\ -1 \end{pmatrix}$.

Let $\underline{x}_1 = \begin{pmatrix} x_{11} \\ x_{12} \\ x_{21} \\ x_{22} \end{pmatrix}$

then substituting in the Newton iteration

$$J_{\underline{x}_o}[\underline{x}_o - \underline{x}_1] = \underline{f}(\underline{x}_o)$$

$$\begin{pmatrix} 3 & 1 & 3 & 0 \\ 1 & 2 & 0 & 3 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 1 - x_{11} \\ 1 - x_{12} \\ 1 - x_{21} \\ - x_{22} \end{pmatrix} = \begin{pmatrix} -4 \\ 2 \\ 1 \\ -1 \end{pmatrix}$$

and $\underline{x}_1'$ is found by solving the equations

$$3(1-x_{11}) + (1-x_{12}) + 3(1-x_{21}) = -4$$

$$(1-x_{11}) + 2(1-x_{12}) - 3x_{22} = 2$$

$$(1-x_{21}) - x_{22} = 1$$

$$(1-x_{21}) = -1 \quad .$$

Using the derivative operator it is not necessary to calculate the constituent equations and the iteration

$$[X_i + A_1][X_i - X_{i+1}] + [X_i - X_{i+1}]X_i = X^2 + A_1X_i + A_2$$

can be simplified to

$$[X_i + A_1]X_{i+1} + X_{i+1}X_i = X_i^2 - A_2$$

and setting $X_o = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$ and letting $X_1 = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{pmatrix}$

then $$\begin{pmatrix} 2 & 3 \\ 0 & 0 \end{pmatrix}X_1 + X_1 \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 11 & 1 \\ 0 & 2 \end{pmatrix}$$

and X_1 can be found from

$$\begin{pmatrix} 3 & 1 & 3 & 0 \\ 1 & 2 & 0 & 3 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} x_{11} \\ x_{12} \\ x_{21} \\ x_{22} \end{pmatrix} = \begin{pmatrix} 11 \\ 1 \\ 0 \\ 2 \end{pmatrix}.$$

The amount of computation is therefore greatly reduced.

5.5 AN ITERATIVE METHOD FOR THE SOLUTION OF MATRIX EQUATIONS USING THE CHARACTERISTIC POLYNOMIAL OF A SOLUTION.

The elimination method described in Chapter 4 may be used to form an iterative algorithm.

The elimination method is based on obtaining a possible characteristic equation of a solution and combining it with the matrix equation to obtain a linear equation in X and hence a solution.

If the characteristic equation is not known then the linear equation gives the solution X in terms of the matrix coefficients and the scalar coefficients of the characteristic polynomial.

This can be used to form an iterative process and the initial values required are approximations to the n characteristic coefficients of X_o. Hence only m initial values are required instead of the m^2 values required for the Newton method when X is an m×m matrix.

Consider the quadratic unilateral matrix equation

$$X^2 + A_1X + A_2 = 0.$$

Let the characteristic equation of X be

$$\lambda^2 + a_1\lambda + a_2 = 0 \quad .$$

Hence X satisfies the two equations

$$X^2 + A_1X + A_2 = 0 \quad \text{and} \quad X^2 + a_1X + a_2I = 0.$$

Eliminating X^2 gives

$$X = [a_1I - A_1]^{-1}[A_2 - a_2I] \quad .$$

If the characteristic coefficients are not known then an iterative process may be set up defined by

$$X_{i+1} = [a_1^{(i)}I - A_1]^{-1}[A_2 - a_2^{(i)}I]$$

where $a_1^{(i)}$ and $a_2^{(i)}$ are the characteristic coefficients of X_i.

The method is illustrated in the following example.

Example 5.5.1.

Consider the equation $X^2 + A_1X + A_2 = 0$

177

where $A_1 = \begin{pmatrix} 2 & 1 \\ -1 & 1 \end{pmatrix}$ $A_2 = \begin{pmatrix} -3 & 3 \\ -9 & -11 \end{pmatrix}$.

Choosing initial values $a_1^{(o)} = -2$ and $a_2^{(o)} = 2$

then $X_1 = [-2I - A_1]^{-1}[A_2 - 2I]$

$$\therefore X_1 = \begin{pmatrix} -4 & -1 \\ 1 & -3 \end{pmatrix}^{-1} \begin{pmatrix} -5 & 3 \\ -9 & -13 \end{pmatrix}$$

$$\therefore X_1 = \begin{pmatrix} 0.462 & -1.692 \\ 3.154 & 3.769 \end{pmatrix}.$$

Then $a_1^{(1)} = -4.231$ $a_2^{(1)} = 7.078$

and $X_2 = [-4.231I - A_1]^{-1}[A_2 - 7.078I]$

$$\therefore X_2 = \begin{pmatrix} -6.231 & -1 \\ 1 & -5.231 \end{pmatrix}^{-1} \begin{pmatrix} -10.078 & 3 \\ -9 & -18.078 \end{pmatrix}$$

$$\therefore X_2 = \begin{pmatrix} 1.301 & -1.005 \\ 1.969 & 3.264 \end{pmatrix}.$$

Then $a_1^{(2)} = -4.565$ $a_2^{(2)} = 6.225$

and $X_3 = [-4.565I - A_1]^{-1}[A_2 - 6.225I]$

$$X_3 = \begin{pmatrix} -6.565 & -1 \\ 1 & -5.565 \end{pmatrix}^{-1} \begin{pmatrix} -9.225 & 3 \\ -9 & -17.225 \end{pmatrix}$$

$$\therefore \quad X_3 = \begin{pmatrix} 1.128 & -0.904 \\ 1.820 & 2.933 \end{pmatrix}$$

Then $\quad a_1^{(3)} = -4.061 \qquad a_2^{(3)} = 4.954$

$$X_4 = [-4.061I - A_1]^{-1} [A_2 - 4.954I]$$

$$\therefore \quad X_4 = \begin{pmatrix} 0.987 & -0.983 \\ 1.973 & 2.958 \end{pmatrix}$$

$$\therefore \quad a_1^{(4)} = -3.945 \qquad a_2^{(4)} = 4.859$$

$$X_5 = [-3.945I - A_1]^{-1} [A_2 - 4.859I]$$

$$\therefore \quad X_5 = \begin{pmatrix} 0.982 & -1.010 \\ 2.019 & 3.003 \end{pmatrix}$$

$$\therefore \quad a_1^{(5)} = -3.985 \qquad a_2^{(5)} = 4.988$$

$$X_6 = [-3.985I - A_1]^{-1} [A_2 - 4.988I]$$

$$X_6 = \begin{pmatrix} 0.999(5) & -1.003 \\ 2.006 & 3.006 \end{pmatrix}$$

$$\therefore \quad a_1^{(6)} = -4.006 \qquad a_2^{(6)} = 5.016$$

$$\therefore \quad X_7 = [-4.006I - A_1]^{-1} [A_2 - 5.016I]$$

$$\therefore \quad X_7 = \begin{pmatrix} 1.002 & -1.000 \\ 1.998 & 3.000 \end{pmatrix}$$

$$\therefore \quad a_1^{(7)} = -4.002 \qquad a_2^{(7)} = 5.004$$

$$X_8 = [-4.002I - A_1]^{-1} [A_2 - 5.004I].$$

$$X_8 = \begin{pmatrix} 1.000 & -1.000 \\ 1.999 & 3.000 \end{pmatrix} .$$

The sequence converges to the solution $X = \begin{pmatrix} 1 & -1 \\ 2 & 3 \end{pmatrix}$.

In the next three examples a wide range of initial values are tested. Convergence is assumed when each element in $[X_i^2 + A_1X_i + A_2]$ has a numerical value less than 10^{-8}. The magnitude of the starting values does not appear to affect the number of iterations required for convergence.

Example 5.5.2.

Consider the equation $X^2 + A_1X + A_2 = 0$

where $A_1 = \begin{pmatrix} 1 & -1 \\ 2 & 1 \end{pmatrix}$ $A_2 = \begin{pmatrix} -6 & -3 \\ -4 & -7 \end{pmatrix}$

Starting values $a_1^{(o)}$	$a_2^{(o)}$	Number of iterations	
-1	1	10	All converge to the solution $X = \begin{pmatrix} 2 & 1 \\ 0 & 1.791 \end{pmatrix}$ which has characteristic coefficients -3.791 and 3.583.
0	1	10	
-5	-5	10	
-100	100	10	
-20	50	10	
-1000	1000	10	
500	-500	10	
-15	-15	12	All converge to the solution $X = \begin{pmatrix} -2.599 & 0.040 \\ -1.920 & -3.192 \end{pmatrix}$ which has characteristic coefficients 5.7913 and 8.3739 .
-20	-20	12	
50	50	12	
-500	-1000	12	

Example 5.5.3.

Consider the equation $X^2 + A_1X + A_2 = 0$

where $A_1 = \begin{pmatrix} -1 & 1 \\ 2 & 0 \end{pmatrix}$ $A_2 = \begin{pmatrix} 0 & 0 \\ -12 & -12 \end{pmatrix}$

Starting values $a_1^{(o)}$	$a_2^{(o)}$	Number of iterations	
-1	1	16	All converge to the solution $X = \begin{pmatrix} 1 & -1 \\ 2 & 4 \end{pmatrix}$ which has characteristic coefficients -5 and 6 .
-10	-10	22	
50	-50	23	
-1000	5000	12	
-50	1	20	

Example 5.5.4.

Consider the equation $X^2 + A_1X + A_2 = 0$

where $A_1 = \begin{pmatrix} -1 & 2 \\ 1 & 1 \end{pmatrix}$ $A_2 = \begin{pmatrix} 0 & 0 \\ -1 & -1 \end{pmatrix}$

Starting values $a_1^{(o)}$	$a_2^{(o)}$	Number of iterations	
-1	1	3	All converge to the solution $X = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}$ which has characteristic coefficients -1, 0 .
-1	0	1	
0	1	13	
0	0	2	
20	-20	25	
500	-50	10	
-1000	1000	16	
-1000	-1	10	
-10	-10		No convergence.
-50	-50		
-1000	-1000		

The iteration may be extended to equations involving matrices of higher order using the recurrence relation obtained in the elimination method described in Chapter 4.

In the 3×3 case, 3 starting values are required and

$$X_{i+1} = -J_2^{-1} K_2 \quad \text{where} \quad J_2 = K_1 - J_1A_1 \qquad J_1 = a_1^{(i)}I - A$$

$$K_2 = a_3^{(i)}I - J_1A_2 \qquad K_1 = a_2^{(i)}I - A_2$$

where $a_1^{(i)}$ $a_2^{(i)}$ $a_3^{(i)}$ are the characteristic coefficients of the matrix X_i.

The next two examples illustrate the iterative solution of the 3×3 quadratic matrix equation with a variety of starting values.

Example 5.5.5.

Consider the equation $X^2 + A_1X + A_2 = 0$

where $A_1 = \begin{pmatrix} 0 & 1 & 0 \\ 2 & -1 & -1 \\ 1 & 1 & 2 \end{pmatrix}$ $A_2 = \begin{pmatrix} -8 & 2 & 1 \\ 6 & -2 & -1 \\ 2 & -2 & -2 \end{pmatrix}$

Starting values			Number of iterations	
$a_1^{(o)}$	$a_2^{(o)}$	$a_3^{(o)}$		
-1	1	-1	8	All converge to the solution
-1	1	0	8	$X = \begin{pmatrix} 0.152 & 0.144 & 0.155 \\ 8.012 & -2.147 & -1.147 \\ -8.241 & 2.940 & 1.940 \end{pmatrix}$
-1	0	0	8	
0	0	0	8	
1	-1	1	8	which has characteristic
-100	2000	-1000	10	coefficients 0.055, -0.789, 0.156.
-10	10	-10	16	Converge to the solution
-10	20	-50	18	$X = \begin{pmatrix} -2.851 & 1.149 & 0.702 \\ 0.379 & 0.379 & 0.242 \\ -1.342 & 0.657 & 0.685 \end{pmatrix}$
				which has characteristic
				coefficients 1.787, -2.491, 0.426.

Starting values			Number of iterations	
$a_1^{(o)}$	$a_2^{(o)}$	$a_3^{(o)}$		
10	20	5	7	Converge to the solution $X = \begin{pmatrix} -2.816 & 0.315 & 0.315 \\ 0.111 & -0.415 & 0.585 \\ -0.251 & -1.797 & -2.797 \end{pmatrix}$ which has characteristic coefficients 6.028, 11.301, 6.273
50	-100	-20		Do not converge to a solution
20	-10	-2		
1000	100	500		

Example 5.5.6.

Consider the equation $X^2 + A_1X + A_2 = 0$

where $A_1 = \begin{pmatrix} 2 & 1 & -3 \\ -1 & -1 & 0 \\ 2 & -1 & -2 \end{pmatrix}$ $A_2 = \begin{pmatrix} -4 & -2 & -5 \\ 2 & -2 & -2 \\ 2 & 0 & -17 \end{pmatrix}$

Starting values			Number of iterations	
$a_1^{(o)}$	$a_2^{(o)}$	$a_3^{(o)}$		
-1	1	-1	6	All converge to the solution $X = \begin{pmatrix} 1.098 & -0.064 & 2.376 \\ -0.055 & 2.077 & 0.753 \\ -1.039 & -0.404 & 4.908 \end{pmatrix}$ which has characteristic coefficients -8.083, 20.633, -16.740
-1	1	0	6	
-1	0	0	6	
-5	10	-50	6	
-500	500	-500	6	
-100	20	200	6	

Starting values			Number of
$a_1^{(o)}$	$a_2^{(o)}$	$a_3^{(o)}$	iterations
0	0	0	7
1	-1	1	7
0	0	1	7
-20	-20	-30	7
-100	-100	-500	7
-10000	-5000	-1	7

All converge to the solution

$$X = \begin{pmatrix} -3.0496 & -1.0496 & 1.029 \\ 1.042 & -0.958 & -0.389 \\ -0.432 & -0.432 & -3.075 \end{pmatrix}$$

which has characteristic coefficients 7.083, 16.615, 12.900

The method may be used for unilateral matrix equations of any degree and any order.

The method is applied to the cubic unilateral matrix equation in the following example.

Example 5.5.7.

Consider the matrix equation $X^3 + A_1X^2 + A_2X + A_3 = 0$

where $A_1 = \begin{pmatrix} 4 & -2 \\ -2 & 1 \end{pmatrix}$ $A_2 = \begin{pmatrix} 1 & 3 \\ 2 & 1 \end{pmatrix}$ $A_3 = \begin{pmatrix} -4 & -1 \\ -12 & -3 \end{pmatrix}$.

The iteration is

$$X_{i+1} = [A_2 - a_2^{(i)}I - a_1^{(i)}A_1 + (a_1^i)^2I]^{-1}[a_2^{(i)}A_1 - a_1^{(i)}a_2^{(i)}I - A_3]$$

where $a_1^{(i)}$, $a_2^{(i)}$ are the characteristic coefficients of X_i.

Starting values $a_1^{(o)}$	Starting values $a_2^{(o)}$	Number of iterations	
-1	1	18	All converge to the solution $X = \begin{pmatrix} 0.825 & -2.204 \\ -0.538 & -0.013 \end{pmatrix}$ which has characteristic coefficients -0.812, -1.196
-1	0	18	
0	1	19	
0	0	20	
-1	-1	20	
-2	-2	20	
-2	-2	20	
1000	-500	20	
50	-50	20	
-10	-10		Do not converge to a solution
-1	-10		
-5	-5		

Example 5.5.8.

Consider the cubic matrix equation $X^3 + A_1X^2 + A_2X + A_3 = 0$

where $A_1 = \begin{pmatrix} 2 & -1 & 0 \\ 1 & 0 & 3 \\ -1 & 0 & 0 \end{pmatrix}$ $A_2 = \begin{pmatrix} 0 & -2 & 1 \\ 8 & -1 & 0 \\ 0 & 1 & 0 \end{pmatrix}$

$A_3 = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 1 & 0 & 1 \end{pmatrix}$.

The iteration is

$$X_{i+1} = [A_2 - T_i - A_1S_i + S_i^2]^{-1}[A_1T_i - S_iT_i - A_3]$$

where $S_i = [A_1 - a_1^{(i)}I]^{-1}[A_2 - a_2^{(i)}I]$

$T_i = [A_1 - a_1^{(i)}I]^{-1}[A_3 - a_3^{(i)}I]$

and $a_1^{(i)}$, $a_2^{(i)}$, $a_3^{(i)}$ are the characteristic coefficients of X_i.

Starting values			Number of iterations
$a_1^{(o)}$	$a_2^{(o)}$	$a_3^{(o)}$	
-3	3	-1	14
3	3	1	11
-1	-1	1	13
-2	-24	55	15

All converge to the solution

$$X = \begin{pmatrix} -0.188 & 0.121 & -0.222 \\ -0.799 & -0.013 & -0.480 \\ -0.278 & -0.063 & -0.644 \end{pmatrix}$$

which has characteristic coefficients 0.845, 0.137, 0.052

Unlike the Newton method, the number of iterations required for convergence is not always affected by the "distance" of the initial values from the true characteristic coefficients. This is illustrated particularly in Example 5.5.6 where convergence takes place in the same number of iterations for widely differing initial values.

In the 2×2 case the iterative process may be defined explicitly in terms of the characteristic coefficients alone and the solution matrix evaluated at the final stage when convergence has occurred for the characteristic coefficients.

Consider the equation $X^2 + A_1X + A_2 = 0$

where $A_1 = \begin{pmatrix} p_1 & p_2 \\ p_3 & p_4 \end{pmatrix}$ $A_2 = \begin{pmatrix} q_1 & q_2 \\ q_3 & q_4 \end{pmatrix}$.

The iteration is $X_{i+1} = [a_1^{(i)}I - A_1]^{-1}[A_2 - a_2^{(i)}I]$.

However, since $a_1^{(i+1)} = -$ Trace X_{i+1}

then

$$a_1^{(i+1)} = \frac{2a_1^{(i)} a_2^{(i)} + c\, a_2^{(i)} + e\, a_1^{(i)} + T}{D}$$

where c and d are the characteristic coefficients of A_1

e and f are the characteristic coefficients of A_2

$T = p_4q_1 + p_1q_4 - p_2q_3 - p_3q_2$

and $D = (a_1^{(i)})^2 + c\, a_1^{(i)} + d$

and since $a_2^{(i+1)} =$ det. X_{i+1}

then $a_2^{(i+1)} = \dfrac{(a_2^{(i)})^2 + e\, a_2^{(i)} + f}{D}$.

Since c, d, e, f and T may all be evaluated from the coefficient matrices, then the iterative process can be written in vector form as

$\underline{a}^{(i+1)} = \underline{f}(\underline{a}^{(i)})$

where $\underline{a} = (a_1, a_2)$ and $\underline{f}(\underline{a}) = [f_1(\underline{a}),\ f_2(\underline{a})]$

where $f_1(\underline{a}) = \dfrac{2a_1a_2 + c\, a_2 + e\, a_1 + T}{D}$

$f_2(\underline{a}) = \dfrac{a_2^2 + e\, a_2 + f}{D}$.

In considering convergent sequences, the norm of the Jacobian matrix $J_{f(\underline{a})}$ may be evaluated, as in section 5.2.

Example 5.5.9.

Consider the equation $X^2 + A_1X + A_2 = 0$

where $A_1 = \begin{pmatrix} 5 & -2 \\ 6 & 3 \end{pmatrix}$ $A_2 = \begin{pmatrix} -4 & -6 \\ -2 & -3 \end{pmatrix}$

Starting values $a_1^{(o)}$	$a_2^{(o)}$	Number of iterations
1	1	10
0	1	10
1	0	10
-1	-1	10
20	20	12
30	-30	12
500	-500	12
-10000	5000	12

All converge to

$a_1 = 0.2803$ $a_2 = 0$

which give the solution

$$X = \begin{pmatrix} 0.5991 & 0.8986 \\ -0.5863 & -0.8794 \end{pmatrix}.$$

The Jacobian matrix is

$$J_{f(\underline{a})} = \begin{pmatrix} \frac{\partial f_1(\underline{a})}{\partial a_1} & \frac{\partial f_1(\underline{a})}{\partial a_2} \\ \frac{\partial f_2(\underline{a})}{\partial a_1} & \frac{\partial f_2(\underline{a})}{\partial a_2} \end{pmatrix}$$

where $$\frac{\partial f_1(\underline{a})}{\partial a_1} = \frac{D[2a_2+e] - [2a_1a_2 + ca_2+ea_1 + T][2a_1+c]}{D^2}$$

:89

$$\frac{\partial f_1(\underline{a})}{\partial a_2} = \frac{2a_1 + c}{D}$$

$$\frac{\partial f_2}{\partial a_1} = -\frac{[2a_1 + c][a_2^2 + ea_2 + f]}{D^2}$$

$$\frac{\partial f_2}{\partial a_2} = \frac{2a_2 + e}{D} \quad .$$

At $a_1 = 0.2803, \quad a_2 = 0$

$$J_{f(\underline{a})} = \begin{pmatrix} -0.3318 & 0.2996 \\ 0 & 0.2818 \end{pmatrix}$$

and $\| J_{f(\underline{a})} \| = 0.528$

$\therefore \quad \| J_{f(\underline{a})} \| < 1$ and convergence occurs.

Consider the equation $X^2 + A_1X + A_2 = 0$

where $A_1 = \begin{pmatrix} -1 & 2 \\ 1 & 1 \end{pmatrix}$ $\qquad A_2 = \begin{pmatrix} 0 & 0 \\ -1 & -1 \end{pmatrix}$.

This is the equation considered in Example 5.5.4.

Many starting values converged to the solution $X = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}$

which has $a_1 = -1$ and $a_2 = 0$,

but no starting values converged to the solution $X = \begin{pmatrix} 1 & -2 \\ -0.5 & -2 \end{pmatrix}$

which has $a_1 = 1, \quad a_2 = -3$.

At the second solution, the Jacobian matrix is

$$J_{f(\underline{a})} = \begin{bmatrix} -3.5 & 1 \\ -3 & 2.5 \end{bmatrix}$$

and $\| J_{f(\underline{a})} \| = 5.34$ and $\| J_{f(\underline{a})} \| > 1$.

However at the solution $X = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}$ which has $a_1 = -1$

$a_2 = 0$ the Jacobian matrix is

$$J_{f(\underline{a})} = \begin{bmatrix} -0.5 & -1 \\ 0 & -0.5 \end{bmatrix}$$

and $\| J_{f(\underline{a})} \| = 1.22$.

Hence $\| J_{f(\underline{a})} \| > 1$ and yet convergence still occurred for many starting values.

This example illustrates the property that if $\| J_{f(\underline{x})} \| < 1$ and an initial choice $\underline{x}_o$ can be found sufficiently close to the solution then the sequence must converge to it.

However, even if $\| J_{f(\underline{x})} \| > 1$ at the solution then the sequence may still converge for suitable choices of $\underline{x}_o$.

5.6 CONCLUSION.

The advantage of iterative methods is that solutions for equations can often be found which are not obtainable by algebraic methods. Many of the methods described in Chapter 3 can fail under certain conditions.

In 3.2 Methods I and II fail when the transforming matrix T is singular and in 3.3 Method I fails when both U and M are singular. An iterative method can often find solvents which cannot be expressed in the form TDT^{-1}.

Most of the methods described in this chapter share the disadvantage that considerable computation is required at each iteration, though with computer methods this can be overcome.

The other problem encountered in applying iterative methods is that of finding a suitable initial matrix which will satisfy the conditions for convergence. In the method of simple iteration however, convergence can occur even when $\| X - X_o \|$ is very large. This is illustrated in Example 5.2.3. This is often the case when the method of simple iteration is applied to the scalar polynomial equation. As shown in 5.2 the iteration $x_{i+1} = \frac{5}{x_i} + 4$ converged to $x = 5$ for all initial values except $x = 0$.

Though it has been shown that in the matrix case, convergence will occur if $\| J_{\underline{x}} \| < 1$, it is unfortunately not possible to predict this in advance since it is difficult to obtain a close approximation to a solution. In the scalar case it is possible to obtain a value of x close to a solution by means of the intermediate value theorem.

It has been shown that a matrix solution may be obtained by applying the Bernoulli iteration to the matrix equation. The disadvantage of this method is that it can lead to only one solution and if a dominant solvent does not exist then convergence will not occur.

Any iterative method for the solution of a system of equations can be applied to the constituent equations which are equivalent to

a matrix equation. However, because of the computation involved in forming the constituent equations, it is preferable to use a method which can be applied directly to the matrix equation.

It has been shown in Section 5.4 that the two methods of applying the Newton iteration are equivalent, but the computation is reduced in applying it direct to the matrix equation. The main problem in applying Newton's method is that of finding a suitable initial matrix for convergence to occur. This means that m^2 initial values are required for equations involving m×m matrices.

The method described in Section 5.5 is new and is of interest since only m initial values are required. The method has been applied to many examples of matrix equations involving 2×2 and 3×3 matrices.

193

CHAPTER 6

The Square Root of a Matrix

6.1 INTRODUCTION.

In the field of complex numbers, any element has precisely two square roots. In the ring of matrices however the number of square roots of a matrix P depends on the nature of the matrix P and also on its size. In 6.5 it is shown that there may be an infinite number of square roots. If X and P are m×m matrices and P has distinct eigenvalues then it can be shown [6.2] that the number of solutions of $X^2 = P$ is 2^m. The number of square roots of a matrix clearly increases rapidly with the size of the matrix.

The equation $X^2 = P$ is simply a special form of the general unilateral matrix quadratic equation $X^2 + A_1X + A_2 = 0$ obtained by setting $A_1 = 0$ and $A_2 = -P$. Hence the methods described in previous chapters for the general quadratic equation may be applied. However, because of the particular form of the equation, other methods may be used which are not applicable to the general unilateral equation.

The method described in 6.2 uses the relationship between the eigenvalues of X and P while the method described in 6.3 derives a relationship between the characteristic coefficients of X and P and uses it to obtain a solution X. In 6.5 it is shown that the equation $X^2 = P$ has an infinite number of solutions if P is derogatory.

The finding of the square root of a matrix has relevance in other techniques. It is necessary for example in the application of Method V for the solution of the Matrix Riccati equation described in 3.3.

6.2 THE SQUARE ROOT OF A MATRIX OBTAINED BY CONSIDERATION OF ITS EIGENVALUES.

This method is described by Gantmacher [1959] and makes use of the fact that if $X^2 = P$, then the eigenvalues of X are the square roots of the eigenvalues of P.

Let the Jordan normal form of P be the block diagonal matrix

$$\overline{P} = \left\{\lambda_1 I + H_1,\ \lambda_2 I + H_2 \ldots\ldots\ldots \lambda_u I_u + H_u\right\}$$

where the λ_i are the eigenvalues of P and the matrix H_i is of the form

$$H_i = \begin{bmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & 1 \\ 0 & 0 & 0 & \ldots & 0 \end{bmatrix} .$$

The matrix H_i is nilpotent and $H_i^m = 0$ if the matrix is m×m.

The square root of a Jordan block can be expressed by means of the series

$$\sqrt{\lambda_i I + H_i} = \lambda_i^{\frac{1}{2}} I + \frac{1}{2}\lambda_i^{-\frac{1}{2}} H_i + \frac{\left(\frac{1}{2}\right)\left(-\frac{1}{2}\right)}{2!}\lambda_i^{-3/2} H_i^2 + \frac{\left(\frac{1}{2}\right)\left(-\frac{1}{2}\right)\left(-\frac{3}{2}\right)}{3!}\lambda_i^{-5/2} H_i^3 + \ldots\ldots .$$

Since $H_i^m = 0$ the series eventually terminates.

For example the square root of $\begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix}$ is given by

$$\sqrt{\lambda}\, I + \frac{1}{2\sqrt{\lambda}} H = \begin{pmatrix} \sqrt{\lambda} & \frac{1}{2\sqrt{\lambda}} \\ 0 & \sqrt{\lambda} \end{pmatrix} .$$

The square root of $\begin{pmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{pmatrix}$ is given by

$$\sqrt{\lambda}\, I + \frac{1}{2\sqrt{\lambda}} H + \frac{\left(\frac{1}{2}\right)\left(-\frac{1}{2}\right)}{2!}\lambda^{-3/2}H^2 = \begin{pmatrix} \sqrt{\lambda} & \frac{1}{2\sqrt{\lambda}} & -\frac{1}{8\sqrt{\lambda^3}} \\ 0 & \sqrt{\lambda} & \frac{1}{2\sqrt{\lambda}} \\ 0 & 0 & \sqrt{\lambda} \end{pmatrix}$$

The square root of any matrix can be found by expressing it in Jordan Normal Form.

Consider the equation $X^2 = P$.

Let $P = U\bar{P}U^{-1}$ where $\bar{P}$ is the Jordan Normal Form of P and let $X = T\bar{X}T^{-1}$ where $\bar{X}$ is the Jordan Normal Form of X.

Then $\quad T\bar{X}^2T^{-1} = U\bar{P}U^{-1}$

$\therefore \quad \bar{X}^2 = Q\bar{P}Q^{-1}$ where $Q = T^{-1}U$.

The matrices $\bar{X}^2$ and $\bar{P}$ are similar and therefore have equal eigenvalues. The eigenvalues of X are the square roots of the eigenvalues of P.

Let $\bar{P} = \text{diag}(J_1, J_2, J_3, \ldots, J_r)$ where $J_i = \lambda_i I + H_i$ and λ_i are the eigenvalues of P.

Then $\quad \bar{X} = \text{diag}(\sqrt{J_1}, \sqrt{J_2}, \ldots \sqrt{J_r})$.

Since the square roots of the Jordan Blocks can be obtained from the terminating series already described then the matrix $\bar{X}$ can be evaluated.

Hence if the transforming matrix T can be obtained then the solution X can be found from $X = T\bar{X}T^{-1}$.

$$X^2 = T\bar{X}^2T^{-1} = T\bar{P}T^{-1} \quad [\text{Since } \bar{X}^2 = \bar{P}] \ .$$

But since $X^2 = P = U\bar{P}U^{-1}$

then $T\overline{P}T^{-1} = U\overline{P}U^{-1}$.

We may therefore choose as the transforming matrix T the matrix UB where B is an arbitrary non singular matrix such that $B\overline{P} = \overline{P}B$.

Then $X = UB(\sqrt{\lambda_1 I+H_1},\ \sqrt{\lambda_2 I+H_2}\ \ldots\ldots \sqrt{\lambda_r I+H_r})B^{-1}U^{-1}$.

The method is illustrated in the following examples.

Example 6.2.1.

Consider the equation $X^2 = P$ where $P = \begin{pmatrix} -2 & 6 \\ -3 & 7 \end{pmatrix}$.

The Jordan Normal Form of P is $\overline{P} = \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix}$

where $P = U\overline{P}U^{-1}$ and $U = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$ $U^{-1} = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}$

then $\overline{X} = \begin{pmatrix} \sqrt{1} & 0 \\ 0 & \sqrt{4} \end{pmatrix}$ and there are therefore four possibilities for $\overline{X}$

$$\overline{X} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \quad \overline{X} = \begin{pmatrix} -1 & 0 \\ 0 & 2 \end{pmatrix} \quad \overline{X} = \begin{pmatrix} 1 & 0 \\ 0 & -2 \end{pmatrix} \quad \overline{X} = \begin{pmatrix} -1 & 0 \\ 0 & -2 \end{pmatrix}$$

and $X = T\overline{X}T^{-1}$ where T = UB and B is an arbitrary matrix which commutes with $\overline{P}$.

$\therefore$ $B = \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}$ and hence $T = \begin{pmatrix} 2a & b \\ a & b \end{pmatrix}$ $T^{-1} = \begin{pmatrix} \frac{1}{a} & -\frac{1}{a} \\ -\frac{1}{b} & \frac{2}{b} \end{pmatrix}$

If $\bar{X} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$ then $X = T\bar{X}T^{-1} = \begin{pmatrix} 0 & 2 \\ -1 & 3 \end{pmatrix}$

If $\bar{X} = \begin{pmatrix} -1 & 0 \\ 0 & 2 \end{pmatrix}$ then $X = T\bar{X}T^{-1} = \begin{pmatrix} -4 & 6 \\ -3 & 5 \end{pmatrix}$

If $\bar{X} = \begin{pmatrix} 1 & 0 \\ 0 & -2 \end{pmatrix}$ then $X = T\bar{X}T^{-1} = \begin{pmatrix} 4 & -6 \\ 3 & -5 \end{pmatrix}$

If $\bar{X} = \begin{pmatrix} -1 & 0 \\ 0 & -2 \end{pmatrix}$ then $X = T\bar{X}T^{-1} = \begin{pmatrix} 0 & -2 \\ 1 & -3 \end{pmatrix}$.

Example 6.2.2.

This example illustrates the case where the matrix P has an infinite number of square roots.

Consider the equation $X^2 = P$ where $P = \begin{pmatrix} 0 & 1 & 1 \\ 0 & 1 & 0 \\ -1 & 1 & 2 \end{pmatrix}$.

The Jordan Normal Form of P is $\bar{P} = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$

where $P = U\bar{P}U^{-1}$ and $U = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \end{pmatrix}$ $U^{-1} = \begin{pmatrix} 0 & 0 & 1 \\ -1 & 1 & 1 \\ 1 & 0 & -1 \end{pmatrix}$

and $\bar{X} = (\sqrt{1+H}, \sqrt{1})$.

198

There are four possibilities for $\overline{X}$

$$\overline{X} = \begin{pmatrix} 1 & \frac{1}{2} & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \overline{X} = \begin{pmatrix} 1 & \frac{1}{2} & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix} \quad \overline{X} = \begin{pmatrix} -1 & -\frac{1}{2} & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$\overline{X} = \begin{pmatrix} -1 & -\frac{1}{2} & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix}$$

and $X = T\overline{X}T^{-1}$ where $T = UB$ and B is an arbitrary matrix which commutes with $\overline{P}$.

Hence $$B = \begin{pmatrix} a & b & c \\ 0 & a & 0 \\ 0 & d & e \end{pmatrix}$$

and $$T = \begin{pmatrix} a & b+d & c+e \\ c & a+d & e \\ a & b & c \end{pmatrix} .$$

Choosing $\overline{X} = \begin{pmatrix} 1 & \frac{1}{2} & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ then $X = T\overline{X}T^{-1} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ 0 & 1 & 0 \\ -\frac{1}{2} & \frac{1}{2} & \frac{3}{2} \end{pmatrix}$.

However, choosing $\overline{X} = \begin{pmatrix} 1 & \frac{1}{2} & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix}$ leads to an infinite number of solutions

$$X = T\overline{X}T^{-1} = \begin{pmatrix} -2pq-2p-2q-\frac{3}{2} & 2pq+2q+\frac{1}{2} & 2pq+2p+2q+\frac{5}{2} \\ -2-2q & 1+2q & 2+2q \\ -2p-2pq-\frac{1}{2} & \frac{1}{2}+2pq & 2pq+2p+\frac{3}{2} \end{pmatrix}$$

where $p = \frac{c}{e}$ $q = \frac{d}{a}$.

Example 6.2.3.

This example illustrates the case in which both the eigenvalues of P are complex and hence the eigenvalues of a solution X are also complex.

Consider the equation $X^2 = P$ where $P = \begin{pmatrix} -1 & -2 \\ 4 & -1 \end{pmatrix}$.

The Jordan Normal Form of P is

$$\overline{P} = \begin{pmatrix} -1+2\sqrt{2}i & 0 \\ 0 & -1-2\sqrt{2}i \end{pmatrix}$$

where $P = U\overline{P}U^{-1}$ and $U = \begin{pmatrix} \frac{i}{\sqrt{2}} & -\frac{i}{\sqrt{2}} \\ 1 & 1 \end{pmatrix}$ $U^{-1} = \begin{pmatrix} -\frac{i}{\sqrt{2}} & \frac{1}{2} \\ \frac{i}{\sqrt{2}} & \frac{1}{2} \end{pmatrix}$.

then $\overline{X} = \begin{pmatrix} \sqrt{-1+2\sqrt{2}i} & 0 \\ 0 & \sqrt{-1-2\sqrt{2}i} \end{pmatrix}$.

There are four possibilities for $\overline{X}$

$$\overline{X} = \begin{pmatrix} 1+\sqrt{2}i & 0 \\ 0 & 1-\sqrt{2}i \end{pmatrix} \qquad \overline{X} = \begin{pmatrix} -1-\sqrt{2}i & 0 \\ 0 & -1+\sqrt{2}i \end{pmatrix}$$

$$\overline{X} = \begin{pmatrix} -1-\sqrt{2}i & 0 \\ 0 & 1-\sqrt{2}i \end{pmatrix} \qquad \overline{X} = \begin{pmatrix} 1+\sqrt{2}i & 0 \\ 0 & -1+\sqrt{2}i \end{pmatrix}$$

and $X = T\overline{X}T^{-1}$ where $T = UB$ and B is an arbitrary matrix which commutes with U.

$$\therefore \quad B = \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix} \quad \text{and} \quad T = \begin{pmatrix} \frac{a}{\sqrt{2}}\,i & -\frac{b}{\sqrt{2}}\,i \\ a & b \end{pmatrix} \qquad T^{-1} = \begin{pmatrix} -\frac{i}{\sqrt{2}a} & \frac{1}{2a} \\ \frac{i}{\sqrt{2}b} & \frac{1}{2b} \end{pmatrix}$$

$$\text{If} \quad \bar{X} = \begin{pmatrix} 1+\sqrt{2}i & 0 \\ 0 & 1-\sqrt{2}i \end{pmatrix} \qquad X = T\bar{X}T^{-1} = \begin{pmatrix} 1 & -1 \\ 2 & 1 \end{pmatrix}$$

$$\text{If} \quad \bar{X} = \begin{pmatrix} -1-\sqrt{2}i & 0 \\ 0 & -1+\sqrt{2}i \end{pmatrix} \qquad X = T\bar{X}T^{-1} = \begin{pmatrix} -1 & 1 \\ -2 & -1 \end{pmatrix}$$

$$\text{If} \quad \bar{X} = \begin{pmatrix} -1-\sqrt{2}i & 0 \\ 0 & 1-\sqrt{2}i \end{pmatrix} \qquad X = T\bar{X}T^{-1} = \begin{pmatrix} -\sqrt{2}i & -\frac{i}{\sqrt{2}} \\ \sqrt{2}i & -\sqrt{2}i \end{pmatrix}$$

$$\text{If} \quad \bar{X} = \begin{pmatrix} 1+\sqrt{2}i & 0 \\ 0 & -1+\sqrt{2}i \end{pmatrix} \qquad X = T\bar{X}T^{-1} = \begin{pmatrix} \sqrt{2}i & \frac{i}{\sqrt{2}} \\ -\sqrt{2}i & \sqrt{2}i \end{pmatrix}.$$

This example shows that there can be real matrix solutions for the equation $X^2 - P = 0$ and yet no real λ for which $\det[\lambda^2 I - P] = 0$.

6.3 EXTENSION OF ELIMINATION METHODS TO FIND THE SQUARE ROOT OF A MATRIX.

An extension of the elimination method described in Chapter 4 may be applied to the problem of finding the square root of a matrix.

In applying the method to the unilateral quadratic matrix equation $X^2 + A_1X + A_2 = 0$, possible characteristic polynomials for X must be found by solving $\det[\lambda^2 I + A_1\lambda + A_2] = 0$. Alternatively an iteration may be set up by choosing starting values for the

characteristic coefficients of X. Either of these methods could be applied to the equation $X^2 = P$.

However in the case of the equation $X^2 - P = 0$, the elimination method may be extended by using the fact that the characteristic equation of X is $\lambda^m + a_1\lambda^{m-1} + \ldots + a_m = 0$ for some a_i to be determined. X can then be eliminated completely and if X and P are m×m matrices then a polynomial in P of degree m is obtained which has coefficients which are scalar polynomials in the a_i. Since P is annihilated by a unique polynomial of degree m if P is non derogatory then by using coefficients of P^i a set of m equations in the unknowns $a_1, a_2, \ldots, a_m$ is obtained.

The solution of these equations gives the characteristic polynomial of a solution X and by elimination a solution X may be found.

The 2×2 case is illustrated as follows.

Consider the equation $X^2 - P = 0$.

The characteristic equation of X is of the form $\lambda^2 + a_1\lambda + a_2 = 0$ then X satisfies the two equations $X^2 - P = 0$

and $X^2 + a_1X + a_2I = 0$.

Eliminating X^2 between the two equations gives a linear equation in X

$$a_1X + (a_2I + P) = 0 \qquad (1)$$

A second linear equation may be obtained by multiplying (1) on the right by X and multiplying $X^2 - P = 0$ by a_1. Subtracting the resulting equations gives

$$[a_2I + P]X + a_1P = 0 \quad . \qquad (2)$$

Multiplying (1) on the left by $[a_2I + P]$ and (2) by a_1 and subtracting eliminates X completely and gives a polynomial in P

$$P^2 + (2a_2 - a_1^2)P + a_2^2 I = 0 \ .$$

But P also satisfies its own characteristic equation

$$P^2 + \alpha_1 P + \alpha_2 I = 0 \ .$$

Comparing coefficients in the two polynomials

$$2a_2 - a_1^2 = \alpha_1$$

$$a_2^2 = \alpha_2 \ .$$

Since the α_1 α_2 can be evaluated from the known matrix P, the values a_1, a_2 can be evaluated and hence X may be found from

$$X = -\frac{1}{a_1}(a_2 I + P) \quad , \quad a_1 \neq 0 \ .$$

Example 6.3.1.

Consider the equation $X^2 = P$ where $P = \begin{pmatrix} -1 & -2 \\ 4 & -1 \end{pmatrix}$

the characteristic equation of P is

$$P^2 + 2P + 9I = 0 \ .$$

If the characteristic equation of X is

$$X^2 + a_1 X + a_2 I = 0$$

then a_1, a_2 satisfy the equations

$$2a_2 - a_1^2 = 2$$

$$a_2^2 = 9 \ .$$

If $a_2 = 3$ then $a_1 = \pm 2$

If $a_2 = -3$ then $a_1 = \pm 2\sqrt{2}i$.

Substituting these values in $X = \frac{1}{a_1}(- a_2 I - P)$

gives the four solutions

$$X = \begin{pmatrix} -1 & 1 \\ -2 & -1 \end{pmatrix} \qquad X = \begin{pmatrix} 1 & -1 \\ 2 & 1 \end{pmatrix} \qquad X = \begin{pmatrix} -\sqrt{2}i & -\frac{i}{\sqrt{2}} \\ \sqrt{2}i & -\sqrt{2}i \end{pmatrix}$$

$$X = \begin{pmatrix} \sqrt{2}i & \frac{i}{\sqrt{2}} \\ -\sqrt{2}i & \sqrt{2}i \end{pmatrix}$$

The method may be extended to matrices of higher order. The 3×3 case is illustrated as follows.

Consider the equation $X^2 = P$ where X and P are 3×3 matrices.

Let the characteristic equation of X be

$$X^3 + a_1X^2 + a_2X + a_3I = 0 .$$

Eliminating X^3 gives a second quadratic equation in X

$$a_1X^2 + [a_2I+P]X + a_3I = 0 .$$

Proceeding as in the 2×2 case, two linear equations in X are obtained

$$[a_2I+P]X + [a_3I+a_1P] = 0$$

$$[a_3I+a_1P]X + a_2P + P^2 = 0 .$$

Eliminating X completely gives a polynomial of degree 3 in P

$$P^3 + (2a_2 - a_1^2)P^2 + (a_2^2 - 2a_1a_3)P - a_3^2 I = 0 \quad .$$

Comparison with the characteristic equation of P

$$P^3 + \alpha_1 P^2 + \alpha_2 P + \alpha_3 I = 0$$

gives three equations

$$2a_2 - a_1^2 = \alpha_1$$

$$a_2^2 - 2a_1a_3 = \alpha_2$$

$$- a_3^2 = \alpha_3 \quad .$$

Since α_1 α_2 α_3 are known, the values of a_1, a_2, a_3 may be evaluated and solutions X found from

$$X = [a_2 I + P]^{-1} [-a_3 I - a_1 P] \quad , \quad [a_2 I + P] \text{ nonsingular.}$$

Example 6.3.2.

Consider the equation $X^2 - P = 0$ where $P = \begin{pmatrix} 0 & 1 & 4 \\ 3 & 3 & 2 \\ 2 & 6 & 1 \end{pmatrix}$.

The characteristic equation of P is

$$P^3 - 4P^2 - 20P - 49I = 0$$

and if the characteristic equation of X is

$$X^3 + a_1 X^2 + a_2 X + a_3 I = 0 \quad .$$

Eliminating X completely and comparing the coefficients of the resulting polynomial in P with the characteristic coefficients of P, the 3 equations are obtained

$$2a_2 - a_1^2 = -4 \qquad (1)$$

$$a_2^2 - 2a_1a_3 = -20 \qquad (2)$$

$$a_3^2 = 49 \quad . \qquad (3)$$

From equation (3) $a_3 = \pm 7$.

Taking $a_3 = 7$ and $a_2 = \dfrac{a_1^2 - 4}{2}$

leads to $a_1^4 - 8a_1^2 - 56a_1 + 96 = 0$

or $(a_1-4)(a_1^3+4a_1^2+8a_1-24) = 0$

$\therefore$ $a_1 = 4$, $a_2 = 6$, $a_3 = 7$ satisfy these equations.

Substituting these values in $X = [a_2I+P]^{-1}[-a_3I-a_1P]$

gives $X = [6I+P]^{-1}[-7I-4P]$

$$\therefore \quad X = \begin{pmatrix} -1 & 1 & -2 \\ -1 & -2 & 0 \\ 0 & -2 & -1 \end{pmatrix} .$$

Since there are 4 values of a_1 for each of the two values of a_3 this indicates that there are 8 square roots of P.

As the orders of X and P increase, the coefficients of the polynomial obtained in P are seen to follow a pattern.

4×4 Case.

The polynomial in P is

$$P^4 + (2a_2-a_1^2)P^3 + (2a_4+a_2^2-2a_1a_3)P^2 + (2a_2a_4-a_3^2)P + a_4^2 = 0$$

5×5 Case.

$$P^5 + (2a_2-a_1^2)P^4 + (2a_4+a_2^2-2a_1a_3)P^3 + (2a_2a_4-2a_1a_5-a_3^2)P^2$$

$$+ (a_4^2-2a_3a_5)P - a_5^2I = 0 \ .$$

6×6 Case.

$$P^6 + (2a_2-a_1^2)P^5 + (2a_4+a_2^2-2a_1a_3)P^4 + (2a_6+2a_2a_4-2a_1a_5-a_3^2)P^3$$

$$+ (2a_2a_6+a_4^2-2a_3a_5)P^2 + (2a_4a_6-a_5^2)P + a_6^2I = 0 \ .$$

m×m Case.

If the characteristic equation of X is

$$X^m + a_1X^{m-1} + a_2X^{m-2} + \ldots\ldots + a_mI = 0$$

the polynomial obtained in P is

$$c_oP^m + c_1P^{m-1} + c_2P^{m-2} + \ldots\ldots + c_m$$

where $c_o = a_o = 1$ and $c_i = \sum_{j+k=2i} a_ja_k(-1)^j \quad i = 1,2,\ldots m, \quad j,k = 0,1,\ldots,m$

and if the characteristic polynomial of P is

$$P^m + \alpha_1P^{m-1} + \alpha_2P^{m-2} + \ldots\ldots + \alpha_mI \ .$$

Then by equating the coefficients a set of m equations is obtained

$$\sum_{j+k=2i} a_ja_k(-1)^j = \alpha_i \ . \quad [i = 1 \text{ to } m] \ .$$

The problem of finding the m^2 unknown elements in the matrix X has therefore been reduced to that of determining m unknown

values in the set of m equations. Iterative methods may be applied to the set of equations to obtain values $a_1, a_2, \ldots, a_m$.

The solution X may then be obtained from

$$X = - J_{m-1}^{-1} J_m \quad \text{where} \quad J_{-1} = 0 \qquad J_o = I$$

$$\text{and} \quad J_i = a_i I + J_{i-2} P \ .$$

The final stage in evaluating $J_{m-1}^{-1} J_m$ involves (m+1) matrix multiplications and the inversion of an m×m matrix.

As in 4.3 it is possible that the matrix J_{m-1} may be singular and that a particular set of solutions $a_1, a_2, \ldots, a_m$ for the set of polynomial equations may not lead to a matrix solution for the equation $X^2 - P = 0$.

Example 6.3.3.

Consider the equation $X^2 - P = 0$ where $P = \begin{pmatrix} 0 & 1 & 1 \\ 0 & 1 & 0 \\ -1 & 1 & 2 \end{pmatrix}$.

The characteristic equation of P is $P^3 - 3P^2 + 3P - I = 0$.

The three equations in a_1, a_2, a_3 are therefore

$$2a_2 - a_1^2 = -3 \qquad (1)$$

$$a_2^2 - 2a_1 a_3 = 3 \qquad (2)$$

$$- a_3^2 = -1 \qquad (3)$$

From (3) $a_3 = \pm 1$.

If $a_3 = 1$ then $a_1 = \dfrac{a_2^2 - 3}{2}$

and hence $a_2^4 - 6a_2^2 - 8a_2 - 3 = 0$

or $(a_2+1)^3(a_2-3) = 0$.

This leads to two possibilities

$$a_1 = -1, \quad a_2 = -1, \quad a_3 = 1$$

or $a_1 = 3$, $a_2 = 3$, $a_3 = 1$.

If $a_3 = -1$ then $a_1 = \frac{3-a_2^2}{2}$

and hence $a_2^4 - 6a_2^2 - 8a_2 - 3 = 0$

or $(a_2+1)^3(a_2-3) = 0$.

This leads to two possibilities

$$a_1 = 1, \quad a_2 = -1, \quad a_3 = -1$$

$$a_1 = -3, \quad a_2 = 3, \quad a_3 = -1 .$$

Choosing $a_1 = -3, \quad a_2 = 3, \quad a_3 = -1$

then $J_1 = -3I, \quad J_2 = 3I+P, \quad J_3 = -I-3P$

and $X = - J_2^{-1} J_3$

$$\therefore \quad X = - \begin{pmatrix} 3 & 1 & 1 \\ 0 & 4 & 0 \\ -1 & 1 & 5 \end{pmatrix}^{-1} \begin{pmatrix} -1 & -3 & -3 \\ 0 & -4 & 0 \\ 3 & -3 & -7 \end{pmatrix}$$

$$\therefore \quad X = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ 0 & 1 & 0 \\ -\frac{1}{2} & \frac{1}{2} & \frac{3}{2} \end{pmatrix} .$$

However choosing $a_1 = -1, \quad a_2 = -1, \quad a_3 = 1$

then $J_1 = -I, \quad J_2 = -I+P, \quad J_3 = I-P$

then $$J_2 = \begin{pmatrix} -1 & 1 & 1 \\ 0 & 0 & 0 \\ -1 & 1 & 1 \end{pmatrix}$$

and hence $X = -J_2^{-1} J_3$ cannot be evaluated since J_2 is singular.

As shown in 6.2.2 there are in fact an infinite number of solutions for this equation which all have $\lambda^3 - \lambda^2 - \lambda + 1$ as their characteristic polynomial.

6.4 THE SOLUTION OF $X^2 = P$ BY USE OF THE COMPANION FORM.

The equations relating the characteristic coefficients of X with the characteristic coefficients of P, which were obtained in 6.3 can also be obtained by using the Companion form of a matrix.

Let C_x be the companion form of X

then $\det[X - \mu I] \equiv \det[C_x - \mu I]$ for all scalar μ

$\therefore$ $\det[X - \mu I].\det[C_x + \mu I] \equiv \det[C_x - \mu I]\det[C_x + \mu I]$

But $\det[C_x + \mu I] \equiv \det[X + \mu I]$.

Hence $\det[X - \mu I]\det[X + \mu I] \equiv \det[C_x - \mu I].\det[C_x + \mu I]$

$\therefore$ $\det\{[X - \mu I][X + \mu I]\} \equiv \det\{[C_x - \mu I][C_x + \mu I]\}$

$\therefore$ $\det[X^2 - \mu^2 I] \equiv \det[C_x^2 - \mu^2 I]$

or $\det[X^2 - \lambda I] \equiv \det[C_x^2 - \lambda I]$ since μ is an arbitrary scalar.

Consider the equation $X^2 = P$

then $X^2 - \lambda I = P - \lambda I$

$\therefore$ $\det[X^2 - \lambda I] = \det[P - \lambda I]$

$\therefore$ $\det[C_x^2 - \lambda I] \equiv \det[P - \lambda I]$.

Hence by comparing coefficients, the equations relating the characteristic coefficients of X and P may be obtained.

2×2 Case.

Let the characteristic equations of X and P be

$$X^2 + a_1 X + a_2 I = 0 \qquad P^2 + \alpha_1 P + \alpha_2 I = 0$$

then $$C_x = \begin{pmatrix} 0 & 1 \\ -a_2 & -a_1 \end{pmatrix}$$

$\therefore$ $$C_x^2 = \begin{pmatrix} -a_2 & -a_1 \\ a_1 a_2 & -a_2 + a_1^2 \end{pmatrix}$$

and $$\det[C_x^2 - \lambda I] = \det \begin{bmatrix} -a_2 - \lambda & -a_1 \\ a_1 a_2 & -a_2 + a_1^2 - \lambda \end{bmatrix} .$$

By row operations

$$\det[C_x^2 - \lambda I] = \det \begin{bmatrix} -a_2 - \lambda & -a_1 \\ -a_1 \lambda & -a_2 - \lambda \end{bmatrix}$$

$$= [-a_2 - \lambda]^2 - \lambda a_1^2$$

$$= \lambda^2 + (2a_2 - a_1^2)\lambda + a_2^2 \quad .$$

But $\det[P - \lambda I] = \lambda^2 + \alpha_1\lambda + \alpha_2$.

Equating coefficients gives the two equations

$$2a_2 - a_1^2 = \alpha_1$$

$$a_2^2 = \alpha_2 \quad .$$

For matrices of higher order, the evaluation of $\det[C_x^2 - \lambda I]$ can always be reduced to the evaluation of a 2×2 determinant by row and column operations.

3x3 Case.

Let the characteristic equation of X be

$$X^3 + a_1X^2 + a_2X + a_3I = 0$$

then $C_x = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -a_3 & -a_2 & -a_1 \end{bmatrix}$ and $C_x^2 = \begin{bmatrix} 0 & 0 & 1 \\ -a_3 & -a_2 & -a_1 \\ a_1a_3 & (a_1a_2-a_3) & (a_1^2-a_2) \end{bmatrix}$

$$\therefore \quad \det[C_x^2-\lambda I] = \det \begin{bmatrix} -\lambda & 0 & 1 \\ -a_3 & -\lambda-a_2 & -a_1 \\ a_1a_3 & a_1a_2-a_3 & a_1^2-a_2-\lambda \end{bmatrix}.$$

$$= \det \begin{bmatrix} -\lambda & 0 & 1 \\ -a_3 & -a_2-\lambda & -a_1 \\ 0 & -a_3-a_1\lambda & -a_2-\lambda \end{bmatrix}$$

$$= - \det \begin{bmatrix} 1 & 0 & 0 \\ 0 & -a_2-\lambda & -a_3-a_1\lambda \\ 0 & -a_3-a_1\lambda & -a_2\lambda-\lambda^2 \end{bmatrix}$$

$$= [-a_3-a_1\lambda]^2 - \lambda[-a_2-\lambda]^2$$

$$= - \lambda^3 + (a_1^2-2a_2)\lambda^2 + (2a_1a_3-a_2^2)\lambda + a_3^2$$

and since $\det[P-\lambda I] = - \lambda^3 - \alpha_1\lambda^2 - \alpha_2\lambda - \alpha_3$

then comparing coefficients $2a_2 - a_1^2 = \alpha_1$

$$a_2^2 - 2a_1a_3 = \alpha_2$$

$$- a_3^2 = \alpha_3 \ .$$

4×4 Case.

Let the characteristic equation of X be

$$X^4 + a_1X^3 + a_2X^2 + a_3X + a_4I = 0$$

then $$C_x = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -a_4 & -a_3 & -a_2 & -a_1 \end{pmatrix}$$

and $$C_x^2 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -a_4 & -a_3 & -a_2 & -a_1 \\ a_1a_4 & a_1a_3-a_4 & a_1a_2-a_3 & a_1^2-a_2 \end{pmatrix}$$

$$\therefore \quad \det[C_x^2-\lambda I] = \det \begin{bmatrix} -\lambda & 0 & 1 & 0 \\ 0 & -\lambda & 0 & 1 \\ -a_4 & -a_3 & -a_2-\lambda & -a_1 \\ a_1a_4 & a_1a_3-a_4 & a_1a_2-a_3 & a_1^2-a_2-\lambda \end{bmatrix}$$

$$= \det \begin{bmatrix} -\lambda & 0 & 1 & 0 \\ 0 & -\lambda & 0 & 1 \\ -a_4 & -a_3 & -a_2-\lambda & -a_1 \\ 0 & -a_4 & -a_3-a_1\lambda & -a_2-\lambda \end{bmatrix}$$

$$= \det \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -a_4-a_2\lambda-\lambda^2 & -a_3-a_1\lambda \\ 0 & 0 & -a_3\lambda-a_1\lambda^2 & -a_4-a_2\lambda-\lambda^2 \end{bmatrix}$$

$$= [-a_4-a_2\lambda-\lambda^2]^2 - \lambda[-a_3-a_1\lambda]^2$$

$$= \lambda^4 + (2a_2-a_1^2)\lambda^3 + (2a_4+a_2^2-2a_1a_3)\lambda^2 + (2a_2a_4-a_3^2)\lambda + a_4^2$$

and since $\det[P-\lambda I] = \lambda^4 + \alpha_1\lambda^3 + \alpha_2\lambda^2 + \alpha_3\lambda + \alpha_4$

comparing coefficients $2a_2 - a_1^2 = \alpha_1$

$$2a_4 + a_2^2 - 2a_1a_3 = \alpha_2$$

$$2a_2a_4 - a_3^2 = \alpha_3$$

$$a_4^2 = \alpha_4$$

---

m×m Case.

The evaluation of $\det[C_x^2 - \lambda I]$ can always be reduced to the evaluation of a 2×2 determinant and the elements of the determinant form a pattern as can be seen from the following table.

Size of X	Value of $\det[C_x^2 - \lambda I]$
2×2	$[-a_2-\lambda]^2 - \lambda a_1^2$
3×3	$[-a_3-a_1\lambda]^2 - \lambda[-a_2-\lambda]^2$
4×4	$[-a_4-a_2\lambda-\lambda^2]^2 - \lambda[-a_3-a_1\lambda]^2$
5×5	$[-a_5-a_3\lambda-a_1\lambda^2]^2 - \lambda[-a_4-a_2\lambda-\lambda^2]^2$
6×6	$[-a_6-a_4\lambda-a_2\lambda^2-\lambda^3]^2 - \lambda[-a_5-a_3\lambda-a_1\lambda^2]^2$

If X is an m×m matrix, then

$$\det[C_x^2-\lambda I] = [-a_m-a_{m-2}\lambda-a_{m-4}\lambda^2....]^2 - \lambda[-a_{m-1}-a_{m-3}\lambda-a_{m-5}\lambda^2....]$$

where $a_o = 1$ and a_{-1}, a_{-2} etc. are all zero.

This therefore gives an alternative method for deriving the m equations relating the characteristic coefficients of X and P. Having obtained $a_1, a_2, \ldots, a_m$ the solution X is obtained from

$$X = - J_{m-1}^{-1} J_m \quad \text{where} \quad J_{-1} = 0 \qquad J_o = I$$

$$\text{and} \quad J_i = a_i I + J_{i-2} P \ .$$

6.5 THE EQUATION $X^2=P$ WHERE P IS A DEROGATORY MATRIX.

Given a derogatory matrix P, the minimum polynomial is of lower degree than the characteristic polynomial and there is more than one Jordan Block associated with a particular eigenvalue. In this section it is shown that the equation $X^2 = P$ has an infinite number of solutions if P is derogatory.

Consider the equation $X^2 = P$.

Let $P = T\bar{P}T^{-1}$ where $\bar{P}$ is the Jordan Normal form of P

then $X^2 = T\bar{P}T^{-1}$

or $Y^2 = \bar{P}$ where $Y = T^{-1}XT$.

Any equation of the form $X^2 = P$ can therefore be replaced by an equivalent equation $Y^2 = \bar{P}$ where $\bar{P}$ is in Jordan Normal Form and the solution X then obtained from $X = TYT^{-1}$.

It is sufficient therefore to consider only the case where P is in Jordan Normal Form.

2×2 Case.

If P is derogatory then it is of the form $\begin{pmatrix} \alpha & 0 \\ 0 & \alpha \end{pmatrix}$

∴ In the 2×2 case, if P is derogatory then $X^2 - P = 0$ has an infinite number of solutions of the form

$$X = \begin{pmatrix} a & \frac{\alpha - a^2}{b} \\ b & -a \end{pmatrix}$$

e.g. if $X^2 = P$ where $P = \begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix}$

then X is any matrix of the form $\begin{pmatrix} a & \frac{3-a^2}{b} \\ b & -a \end{pmatrix}$.

3×3 Case.

If P is derogatory $\overline{P}$ is of the form

$$P_1 = \begin{pmatrix} \alpha & 0 & 0 \\ 0 & \alpha & 0 \\ 0 & 0 & \alpha \end{pmatrix} \quad P_2 = \begin{pmatrix} \alpha & 1 & 0 \\ 0 & \alpha & 0 \\ 0 & 0 & \alpha \end{pmatrix} \text{ or } P_3 = \begin{pmatrix} \alpha & 0 & 0 \\ 0 & \alpha & 0 \\ 0 & 0 & \beta \end{pmatrix}$$

If $X^2 = P_1$ there are an infinite number of solutions of the form

$$X = \begin{pmatrix} a & \frac{\alpha-a^2}{b} & 0 \\ b & -a & 0 \\ 0 & 0 & \sqrt{\alpha} \end{pmatrix}$$

e.g. if $P_1 = \begin{pmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{pmatrix}$ $X = \begin{pmatrix} a & \frac{4-a^2}{b} & 0 \\ b & -a & 0 \\ 0 & 0 & 2 \end{pmatrix}$.

If $X^2 = P_2$ there are an infinite number of solutions of the form

$$X = \begin{pmatrix} \sqrt{\alpha} & 2ab\sqrt{\alpha} + \frac{1}{2\sqrt{\alpha}} & -2b\sqrt{\alpha} \\ 0 & \sqrt{\alpha} & 0 \\ 0 & 2a\sqrt{\alpha} & -\sqrt{\alpha} \end{pmatrix}$$

e.g. if $P_2 = \begin{pmatrix} 4 & 1 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{pmatrix}$ then $X = \begin{pmatrix} 2 & 4ab+\frac{1}{4} & -4b \\ 0 & 2 & 0 \\ 0 & 4a & -2 \end{pmatrix}$.

If $X^2 = P_3$ there are an infinite number of solutions of the form

$$X = \begin{pmatrix} a & \frac{\alpha-a^2}{b} & 0 \\ b & -a & 0 \\ 0 & 0 & \sqrt{\beta} \end{pmatrix}.$$

4×4 Case.

If P is derogatory then $\overline{P}$ has one of the following forms

$$P_1 = \begin{pmatrix} \alpha & 0 & 0 & 0 \\ 0 & \alpha & 0 & 0 \\ 0 & 0 & \alpha & 0 \\ 0 & 0 & 0 & \alpha \end{pmatrix} \qquad P_2 = \begin{pmatrix} \alpha & 0 & 0 & 0 \\ 0 & \alpha & 0 & 0 \\ 0 & 0 & \alpha & 0 \\ 0 & 0 & 0 & \beta \end{pmatrix}$$

$$P_3 = \begin{pmatrix} \alpha & 0 & 0 & 0 \\ 0 & \alpha & 0 & 0 \\ 0 & 0 & \beta & 0 \\ 0 & 0 & 0 & \beta \end{pmatrix} \qquad P_4 = \begin{pmatrix} \alpha & 1 & 0 & 0 \\ 0 & \alpha & 0 & 0 \\ 0 & 0 & \alpha & 0 \\ 0 & 0 & 0 & \beta \end{pmatrix}$$

$$P_5 = \begin{pmatrix} \alpha & 1 & 0 & 0 \\ 0 & \alpha & 1 & 0 \\ 0 & 0 & \alpha & 0 \\ 0 & 0 & 0 & \alpha \end{pmatrix} \qquad P_6 = \begin{pmatrix} \alpha & 1 & 0 & 0 \\ 0 & \alpha & 0 & 0 \\ 0 & 0 & \alpha & 1 \\ 0 & 0 & 0 & \alpha \end{pmatrix}$$

If P has one of the first four forms then solutions may be found by partitioning X according to the Jordan Normal form of P and then using the results obtained for the 2×2 and 3×3 cases.

e.g. $X^2 = \begin{pmatrix} \alpha & 0 & 0 & 0 \\ 0 & \alpha & 0 & 0 \\ 0 & 0 & \alpha & 0 \\ 0 & 0 & 0 & \alpha \end{pmatrix}$ can be written as

$$\left(\begin{array}{c|c} X_1^2 & 0 \\ \hline 0 & X_2^2 \end{array}\right) = \left(\begin{array}{cc|cc} \alpha & 0 & 0 & 0 \\ 0 & \alpha & 0 & 0 \\ \hline 0 & 0 & \alpha & 0 \\ 0 & 0 & 0 & \alpha \end{array}\right)$$

and the equations $X_1^2 = \begin{pmatrix} \alpha & 0 \\ 0 & \alpha \end{pmatrix}$ $X_2^2 = \begin{pmatrix} \alpha & 0 \\ 0 & \alpha \end{pmatrix}$ can then be solved.

Hence $X = \left(\begin{array}{c|c} X_1 & 0 \\ \hline 0 & X_2 \end{array}\right) = \begin{pmatrix} a & \frac{\alpha-a^2}{b} & 0 & 0 \\ b & -a & 0 & 0 \\ 0 & 0 & c & \frac{\alpha-c^2}{d} \\ 0 & 0 & d & -c \end{pmatrix}$.

Similarly a solution for $X^2 = \left(\begin{array}{ccc|c} \alpha & 0 & 0 & 0 \\ 0 & \alpha & 0 & 0 \\ 0 & 0 & \alpha & 0 \\ \hline 0 & 0 & 0 & \beta \end{array}\right)$ is $X = \begin{pmatrix} a & \frac{\alpha-a^2}{b} & 0 & 0 \\ b & -a & 0 & 0 \\ 0 & 0 & \sqrt{\alpha} & 0 \\ 0 & 0 & 0 & \sqrt{\beta} \end{pmatrix}$

A solution for $X = \left(\begin{array}{cc|cc} \alpha & 0 & 0 & 0 \\ 0 & \alpha & 0 & 0 \\ \hline 0 & 0 & \beta & 0 \\ 0 & 0 & 0 & \beta \end{array}\right)$

is $X = \begin{pmatrix} a & \frac{\alpha-a^2}{b} & 0 & 0 \\ b & -a & 0 & 0 \\ 0 & 0 & c & \frac{\beta-c^2}{d} \\ 0 & 0 & d & -c \end{pmatrix}$

and a solution for $X^2 = \left(\begin{array}{ccc|c} \alpha & 1 & 0 & 0 \\ 0 & \alpha & 0 & 0 \\ 0 & 0 & \alpha & 0 \\ \hline 0 & 0 & 0 & \beta \end{array}\right)$

is $X = \begin{pmatrix} \sqrt{\alpha} & 2ab\sqrt{\alpha}+\frac{1}{2\sqrt{\alpha}} & -2b\sqrt{\alpha} & 0 \\ 0 & \sqrt{\alpha} & 0 & 0 \\ 0 & 2a\sqrt{\alpha} & -\sqrt{\alpha} & 0 \\ 0 & 0 & 0 & \sqrt{\beta} \end{pmatrix}$.

Consider the equation $X^2 = P_5 = \begin{pmatrix} \alpha & 1 & 0 & 0 \\ 0 & \alpha & 1 & 0 \\ 0 & 0 & \alpha & 0 \\ 0 & 0 & 0 & \alpha \end{pmatrix}$

Following the method described in 6.2 then a solution X can be written as $X = \bar{T}\bar{X}T^{-1}$

where $$X = \begin{pmatrix} \sqrt{\alpha} & \frac{1}{2\sqrt{\alpha}} & -\frac{1}{8\sqrt{\alpha^3}} & 0 \\ 0 & \sqrt{\alpha} & \frac{1}{2\sqrt{\alpha}} & 0 \\ 0 & 0 & \sqrt{\alpha} & 0 \\ 0 & 0 & 0 & \sqrt{\alpha} \end{pmatrix}$$

or $$\left\{\left(\sqrt{\alpha}\, I + \frac{1}{2\sqrt{\alpha}} H - \frac{1}{8\sqrt{\alpha^3}} H^2\right), \quad \sqrt{\alpha}\right\}$$

and T is a general matrix which commutes with P.

Hence T is of the form $$\left(\begin{array}{ccc:c} & T_1 & & b \\ & & & 0 \\ & & & 0 \\ \hdashline 0 & 0 & a & c \end{array}\right)$$

where $$T_1 = \begin{pmatrix} d & e & f \\ 0 & d & e \\ 0 & 0 & d \end{pmatrix}$$

and the inverse matrix $T^{-1} = \left(\begin{array}{ccc:c} & S_1 & & -\frac{b}{cd} \\ & & & 0 \\ & & & 0 \\ \hdashline 0 & 0 & \frac{-a}{cd} & \frac{1}{c} \end{array}\right)$

and S_1 is of the same form as T_1 .

Then choosing $\overline{X}$ so that the values of $\sqrt{\alpha}$ are of opposite signs in the two Jordan Blocks

$$X = \left(\begin{array}{ccc|c} & T_1 & & \begin{matrix}b\\0\\0\end{matrix} \\ \hline 0 & 0 & a & c \end{array}\right)\left(\begin{array}{ccc|c} & X_1 & & \begin{matrix}0\\0\\0\end{matrix} \\ \hline 0 & 0 & 0 & -\sqrt{\alpha} \end{array}\right)\left(\begin{array}{ccc|c} & S_1 & & \begin{matrix}-\frac{b}{cd}\\0\\0\end{matrix} \\ \hline 0 & 0 & -\frac{a}{cd} & \frac{1}{c} \end{array}\right)$$

where $X_1 = \sqrt{\alpha}\, I + \frac{1}{2\sqrt{\alpha}} H - \frac{1}{8\sqrt{\alpha^3}} H^2$.

and since T_1 commutes with H and H^2 then $T_1X_1 = X_1T_1$.

Hence $$X = \left(\begin{array}{ccc|c} & & & -\frac{2b}{c}\sqrt{\alpha} \\ & & & 0 \\ & & & 0 \\ \hline 0 & 0 & \frac{2a}{d}\sqrt{\alpha} & -\sqrt{\alpha} \end{array}\right)$$

and since $TT^{-1} =$ $$\left(\begin{array}{ccc|c} & T_1 & & \begin{matrix}b\\0\\0\end{matrix} \\ \hline 0 & 0 & a & c \end{array}\right)\left(\begin{array}{ccc|c} & S_1 & & \begin{matrix}-\frac{b}{cd}\\0\\0\end{matrix} \\ \hline 0 & 0 & -\frac{a}{cd} & \frac{1}{c} \end{array}\right) = I$$

then $$\left(\begin{array}{ccc|c} & & & 0 \\ & T_1S_1 - \frac{ba}{cd}H^2 & & 0 \\ & & & 0 \\ \hline 0 & 0 & 0 & 1 \end{array}\right) = \left(\begin{array}{ccc|c} & & & 0 \\ & I & & 0 \\ & & & 0 \\ \hline 0 & 0 & 0 & 1 \end{array}\right).$$

Hence $$T_1S_1 = I + \frac{ba}{cd} H^2 \quad .$$

Hence $X_1 T_1 S_1 = X_1 + \frac{ba}{cd} X_1 H^2$.

The solution is therefore

$$X = \begin{pmatrix} \sqrt{\alpha} & \frac{1}{2\sqrt{\alpha}} & \frac{2ab}{cd}\sqrt{\alpha} - \frac{1}{8\sqrt{\alpha^3}} & -\frac{2b}{c}\sqrt{\alpha} \\ 0 & \sqrt{\alpha} & \frac{1}{2\sqrt{\alpha}} & 0 \\ 0 & 0 & \sqrt{\alpha} & 0 \\ 0 & 0 & \frac{2a}{d}\sqrt{\alpha} & -\sqrt{\alpha} \end{pmatrix}$$

and hence the equation $X^2 = P_5$ has an infinite number of solutions.

Consider the equation $X^2 = P_6 = \begin{pmatrix} \alpha & 1 & 0 & 0 \\ 0 & \alpha & 0 & 0 \\ 0 & 0 & \alpha & 1 \\ 0 & 0 & 0 & \alpha \end{pmatrix}$

then a solution exists $X = T\bar{X}T^{-1}$

where $\bar{X} = \begin{pmatrix} \sqrt{\alpha} & \frac{1}{2\sqrt{\alpha}} & 0 & 0 \\ 0 & \sqrt{\alpha} & 0 & 0 \\ 0 & 0 & \sqrt{\alpha} & \frac{1}{2\sqrt{\alpha}} \\ 0 & 0 & 0 & \sqrt{\alpha} \end{pmatrix}$

and T is a matrix which commutes with P_6 .

Let $T = \begin{pmatrix} T_1 & T_2 \\ T_3 & T_4 \end{pmatrix}$ and the inverse matrix $T^{-1} = \begin{pmatrix} S_1 & S_2 \\ S_3 & S_4 \end{pmatrix}$

where the matrices T_i and S_i are of the form $\begin{pmatrix} a_i & b_i \\ 0 & a_i \end{pmatrix}$.

Then choosing $\overline{X}$ so that the values of $\sqrt{\alpha}$ are of opposite signs in the two Jordan Blocks

$$X = T\overline{X}T^{-1} = \begin{pmatrix} T_1 & T_2 \\ T_3 & T_4 \end{pmatrix} \left(\begin{array}{cc|cc} \sqrt{\alpha} & \frac{1}{2\sqrt{\alpha}} & 0 & 0 \\ 0 & \sqrt{\alpha} & 0 & 0 \\ \hline 0 & 0 & -\sqrt{\alpha} & -\frac{1}{2\sqrt{\alpha}} \\ 0 & 0 & 0 & -\sqrt{\alpha} \end{array}\right) \begin{pmatrix} S_1 & S_2 \\ S_3 & S_4 \end{pmatrix}$$

$$\therefore \quad X = \begin{pmatrix} \sqrt{\alpha}\, T_1 + \frac{1}{2\sqrt{\alpha}} T_1 H & -\sqrt{\alpha}\, T_2 - \frac{1}{2\sqrt{\alpha}} T_2 H \\ \sqrt{\alpha}\, T_3 + \frac{1}{2\sqrt{\alpha}} T_3 H & -\sqrt{\alpha}\, T_4 - \frac{1}{2\sqrt{\alpha}} T_4 H \end{pmatrix} \begin{pmatrix} S_1 & S_2 \\ S_3 & S_4 \end{pmatrix}$$

and since $T_i H = H T_i$ for $i = 1$ to 4

$$\text{then } X = \begin{pmatrix} (\sqrt{\alpha}\, I + \frac{1}{2\sqrt{\alpha}} H)(T_1 S_1 - T_2 S_3) & (\sqrt{\alpha}\, I + \frac{1}{2\sqrt{\alpha}} H)(T_1 S_1 - T_2 S_4) \\ (\sqrt{\alpha}\, I + \frac{1}{2\sqrt{\alpha}} H)(T_3 S_1 - T_4 S_3) & (\sqrt{\alpha}\, I + \frac{1}{2\sqrt{\alpha}} H)(T_3 S_2 - T_4 S_4) \end{pmatrix}$$

$$\therefore \quad X = \left(\begin{array}{c|c} \sqrt{\alpha}\, I + \frac{1}{2\sqrt{\alpha}} H & 0 \\ \hline 0 & \sqrt{\alpha}\, I + \frac{1}{2\sqrt{\alpha}} H \end{array}\right) \left(\begin{array}{c|c} T_1 S_1 - T_2 S_3 & T_1 S_2 - T_2 S_4 \\ \hline T_3 S_1 - T_4 S_3 & T_3 S_2 - T_4 S_4 \end{array}\right)$$

It can be shown that the matrix X contains arbitrary elements as follows.

Let $$T = \begin{pmatrix} a & b & c & d \\ 0 & a & 0 & c \\ e & f & g & h \\ 0 & e & 0 & g \end{pmatrix}$$

then the inverse matrix $T^{-1} = \left(\begin{array}{c|c} S_1 & S_2 \\ \hline S_3 & S_4 \end{array}\right)$

has blocks $$S_1 = \begin{pmatrix} \frac{-g}{ce-ag} & \frac{deg-g^2b-hce+fgc}{(ce-ag)^2} \\ 0 & \frac{-g}{ce-ag} \end{pmatrix}$$

$$S_2 = \begin{pmatrix} \frac{-c}{ag-ce} & \frac{cha-fc^2-dag+bcg}{(ag-ec)^2} \\ 0 & \frac{-c}{ag-ec} \end{pmatrix}$$

$$S_3 = \begin{pmatrix} \frac{e}{ce-ag} & \frac{ahe+gbe-afg-de^2}{(ce-ag)^2} \\ 0 & \frac{e}{ce-ag} \end{pmatrix}$$

$$S_4 = \begin{pmatrix} \frac{a}{ag-ec} & \frac{cda-ha^2-bce+fca}{(ag-ec)^2} \\ 0 & \frac{a}{ag-ec} \end{pmatrix}$$

Hence $$T_1S_1-T_2S_3 = \begin{pmatrix} \frac{-ag-ce}{ce-ag} & \frac{2[adeg-ahce+afgc-bgce]}{(ce-ag)^2} \\ 0 & \frac{-ag-ce}{ce-ag} \end{pmatrix}$$

$$T_1S_2-T_2S_4 = \begin{pmatrix} \frac{-2ac}{ag-ec} & \frac{2[a^2ch-afc^2-da^2g+bec^2]}{(ce-ag)^2} \\ 0 & \frac{-2ac}{ag-ec} \end{pmatrix}$$

$$T_3S_1-T_4S_3 = \begin{pmatrix} \frac{-2ge}{ce-ag} & \frac{2[de^2g-g^2eb-hce^2+fag^2]}{(ce-ag)^2} \\ 0 & \frac{-2ge}{(ce-ag)} \end{pmatrix}$$

$$T_3S_2-T_4S_4 = \begin{pmatrix} \frac{-ag-ce}{ag-ec} & \frac{2[echa-edag+ebcg-gfac]}{(ag-ec)^2} \\ 0 & \frac{-ag-ec}{ag-ec} \end{pmatrix}.$$

Hence the matrix X has arbitrary elements.

In the 4×4 case therefore, if the matrix P is derogatory then the equation $X^2 = P$ has an infinite number of solutions.

m×m Case.

If the matrix P is derogatory then there is more than one Jordan Block corresponding to a single eigenvalue.

If $X^2 = P$ where P is in Jordan Normal Form then X may be partitioned to correspond with the blocks in P associated with a particular eigenvalue.

e.g. if $P = \left(\begin{array}{ccc|cc} \alpha & 1 & 0 & 0 & 0 \\ 0 & \alpha & 0 & 0 & 0 \\ 0 & 0 & \alpha & 0 & 0 \\ \hline 0 & 0 & 0 & \beta & 1 \\ 0 & 0 & 0 & 0 & \beta \end{array}\right)$

then solutions X may be looked for of the form $\begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix}$

where $X_1^2 = \begin{pmatrix} \alpha & 1 & 0 \\ 0 & \alpha & 0 \\ 0 & 0 & \alpha \end{pmatrix}$ and $X_2^2 = \begin{pmatrix} \beta & 1 \\ 0 & \beta \end{pmatrix}$

and the equation $X^2 = P$ becomes $\begin{pmatrix} X_1^2 & 0 \\ 0 & X_2^2 \end{pmatrix} = \begin{pmatrix} P_1 & 0 \\ 0 & P_2 \end{pmatrix}$.

In general the equation $X^2 = P$ may be partitioned as

$$\begin{pmatrix} X_1^2 & 0 & 0 & \dots & 0 \\ 0 & X_2^2 & 0 & \dots & 0 \\ 0 & 0 & X_3^2 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & & X_r^2 \end{pmatrix} = \begin{pmatrix} P_1 & 0 & 0 & \dots & 0 \\ 0 & P_2 & 0 & \dots & 0 \\ 0 & 0 & P_3 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & P_r \end{pmatrix}.$$

If P is derogatory then at least one of the blocks P_i can be further partitioned into more than one Jordan block. In this case it can be shown that the equation $X_i^2 = P_i$ then has an infinite number of solutions.

Consider $X_i^2 = P_i$

where $P_i = \left(\begin{array}{cccc|cccc} \alpha & 1 & 0 & \ldots & 0 & & & \\ 0 & \alpha & 1 & \ldots & 0 & & & \\ \ldots & \ldots & \ldots & \ldots & \ldots & & \bigcirc & \\ 0 & 0 & 0 & \ldots & 1 & & & \\ 0 & 0 & 0 & \ldots & \alpha & & & \\ \hline & & & & & \alpha & 1 & 0 \ldots 0 \\ & & \bigcirc & & & 0 & \alpha & 1 \ldots 0 \\ & & & & & \ldots & \ldots & \ldots \\ & & & & & 0 & 0 & 0 \ldots 1 \\ & & & & & 0 & 0 & 0 \ldots \alpha \end{array}\right)$

then $\bar{X}_i = (\sqrt{\alpha I + H}) \ , \ \sqrt{\alpha I + H})$ {the two diagonal blocks may be of different order}

$$\therefore \quad \bar{X}_i = \left[\left(\sqrt{\alpha}\, I \frac{1}{2\sqrt{\alpha}} H - \frac{1}{8\sqrt{\alpha^3}} H^2 + \ldots\right), \left(\sqrt{\alpha}\, I + \frac{1}{2\sqrt{\alpha}} H - \frac{1}{8\sqrt{\alpha^3}} H^2 + \ldots\right)\right]$$

and $X_i = T\bar{X}_i T^{-1}$ where $T = \begin{pmatrix} T_1 & T_2 \\ T_3 & T_4 \end{pmatrix}$ $\quad T^{-1} = \begin{pmatrix} S_1 & S_2 \\ S_3 & S_4 \end{pmatrix}$

and the matrices T_i, S_i are of the form $\begin{pmatrix} a_1 & a_2 & a_3 & \ldots & a_r \\ 0 & a_1 & a_2 & \ldots & a_{r-1} \\ 0 & 0 & a_1 & \ldots & a_{r-2} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & a_1 \end{pmatrix}$.

Choosing opposite signs of $\sqrt{\alpha}$ for the different blocks of $\bar{X}$ then

$$X_i = \begin{pmatrix} T_1 & T_2 \\ T_3 & T_4 \end{pmatrix} \begin{pmatrix} (\sqrt{\alpha} I + \frac{1}{2\sqrt{\alpha}} H - \frac{1}{8\sqrt{\alpha^3}} H^2 + \ldots) & 0 \\ 0 & (-\sqrt{\alpha} I - \frac{1}{2\sqrt{\alpha}} H + \frac{1}{8\sqrt{\alpha^3}} H^2 + \ldots) \end{pmatrix} \begin{pmatrix} S_1 & S_2 \\ S_3 & S_4 \end{pmatrix}$$

$$X_i=\begin{pmatrix} (\sqrt{\alpha}I+\frac{1}{2\sqrt{\alpha}}H-\frac{1}{8\sqrt{\alpha^3}}H^2+\ldots) & 0 \\ 0 & (\sqrt{\alpha}I+\frac{1}{2\sqrt{\alpha}}H-\frac{1}{8\sqrt{\alpha^3}}H^2+\ldots) \end{pmatrix}\begin{pmatrix} T_1S_1-T_2S_3 & T_1S_2-T_2S_4 \\ T_3S_1-T_4S_3 & T_3S_2-T_4S_4 \end{pmatrix}$$

Since the matrix $\begin{pmatrix} T_1S_1-T_2S_3 & T_1S_2-T_2S_4 \\ T_3S_1-T_4S_3 & T_3S_2-T_4S_4 \end{pmatrix}$ contains

arbitrary elements, then the equation $X_i^2 = P_i$ has an infinite number of solutions when the matrix P_i contains more than one Jordan Block corresponding to a single eigenvalue.

It is straightforward to show that this matrix X_i is indeed a solution of $X_i^2 = P_i$ since X_i can be written as the product

$$X_i = \begin{pmatrix} \sqrt{\alpha I+H} & 0 \\ 0 & \sqrt{\alpha I+H} \end{pmatrix}\begin{pmatrix} T_1 & T_2 \\ T_3 & T_4 \end{pmatrix}\begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix}\begin{pmatrix} S_1 & S_2 \\ S_3 & S_4 \end{pmatrix}.$$

Since the matrix $\begin{pmatrix} \sqrt{\alpha I+H} & 0 \\ 0 & \sqrt{\alpha I+H} \end{pmatrix}$ commutes with T, T^{-1} and

$\begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix}$

then $X_i^2 = \begin{pmatrix} \sqrt{\alpha I+H} & 0 \\ 0 & \sqrt{\alpha I+H} \end{pmatrix}^2\begin{pmatrix} T_1 & T_2 \\ T_3 & T_4 \end{pmatrix}\begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix}\begin{pmatrix} S_1 & S_2 \\ S_3 & S_4 \end{pmatrix}\begin{pmatrix} T_1 & T_2 \\ T_3 & T_4 \end{pmatrix}\begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix}\begin{pmatrix} S_1 & S_2 \\ S_3 & S_4 \end{pmatrix}$

and since $\begin{pmatrix} S_1 & S_2 \\ S_3 & S_4 \end{pmatrix}\begin{pmatrix} T_1 & T_2 \\ T_3 & T_4 \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix}$

then $X_i^2 = \begin{pmatrix} \alpha I+H & 0 \\ 0 & \alpha I+H \end{pmatrix}\begin{pmatrix} T_1 & T_2 \\ T_3 & T_4 \end{pmatrix}\begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix}\begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix}\begin{pmatrix} S_1 & S_2 \\ S_3 & S_4 \end{pmatrix}$

$$\therefore \quad X_i^2 = \begin{pmatrix} \alpha I+H & 0 \\ 0 & \alpha I+H \end{pmatrix} \begin{pmatrix} T_1 & T_2 \\ T_3 & T_4 \end{pmatrix} \begin{pmatrix} S_1 & S_2 \\ S_3 & S_4 \end{pmatrix}$$

$$\therefore \quad X_i^2 = \begin{pmatrix} \alpha I+H & 0 \\ 0 & \alpha I+H \end{pmatrix}$$

$$\therefore \quad X_i^2 = P_i \quad .$$

It can therefore be concluded that if P is derogatory the equation $X^2 = P$ has an infinite number of solutions since one of the blocks $X_i^2 = P_i$ has an infinite number of solutions.

The equations derived in 6.3 and 6.4 assumed that P was non derogatory since the elimination method depends on comparing the coefficients of a scalar equation in P which has the same degree. However when P is derogatory there is not a unique scalar equation of degree m which is satisfied by P.

However, even if P is derogatory, a solution obtained by this method is still valid even though it may be a particular example of a general solution.

The following example illustrates the case.

Example 6.5.1.

Consider the equation $X^2 = P$ where $P = \begin{pmatrix} 4 & 1 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{pmatrix}$.

The characteristic equation of P is

$$P^3 - 12P^2 + 48P - 64I = 0 \quad .$$

If the characteristic equation of X is

$$X^3 + a_1X^2 + a_2X + a_3I = 0$$

then the elimination method gives the three equations

$$2a_2 - a_1^2 = -12$$

$$a_2^2 - 2a_1a_3 = 48$$

$$a_3^2 = 64 \quad .$$

There are four possible solutions to these equations

(1) $a_1 = -2 \qquad a_2 = -4 \qquad a_3 = 8$

(2) $a_1 = 2 \qquad a_2 = -4 \qquad a_3 = -8$

(3) $a_1 = 6 \qquad a_2 = 12 \qquad a_3 = 8$

(4) $a_1 = -6 \qquad a_2 = 12 \qquad a_3 = -8$.

Choosing solution (1) and substituting in

$$[a_2I + P]X = [-a_3I - a_1P]$$

leads to $\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 & x_2 & x_3 \\ x_4 & x_5 & x_6 \\ x_7 & x_8 & x_9 \end{pmatrix} = \begin{pmatrix} 0 & 2 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$

$\therefore \begin{pmatrix} x_4 & x_5 & x_6 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 2 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$.

Hence X is any matrix with $x_4 = 0 \quad x_5 = 2 \quad x_6 = 0$ which

also has characteristic equation $X^3 - 2X^2 - 4X + 8I = 0$.

The elimination method indicates that there are an infinite number of solutions but obtaining the general solution would not be straight forward.

However other choices of characteristic coefficients may lead to a particular solution. In this example solutions (3) and (4) lead to such solutions.

Solution (3) $a_1 = 6 \quad a_2 = 12 \quad a_3 = 8$

on substitution in $X = [a_2 I + P]^{-1}[-a_3 I - a_1 P]$

gives the solution

$$X = \begin{pmatrix} -2 & -\frac{1}{4} & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -2 \end{pmatrix}.$$

Solution (4) $a_1 = -6 \quad a_2 = 12 \quad a_3 = -8$

gives the solution

$$X = \begin{pmatrix} 2 & \frac{1}{4} & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix}.$$

These solutions are particular examples of a general solution which is

$$X = \begin{pmatrix} p & (p-q)ab + \frac{1}{2p} & b(q-p) \\ 0 & p & 0 \\ 0 & a(p-q) & q \end{pmatrix}.$$

where $p^2 = q^2 = 4$

and a, b are arbitrary elements.

6.6 CONCLUSION.

In this chapter two methods of determining the square root of a matrix have been considered. The method described in 6.2, in which the square roots of the eigenvalues of P are used to determine the solution X of $X^2 = P$, involves difficulties in practical application. Not only the eigenvalues of P must be determined but also its Jordan Normal Form. The square root of each Jordan Block must be found by use of the terminating series. Finally a matrix which commutes with $\overline{P}$ must be obtained in order to find the transforming matrix T to obtain the solution $X = T\overline{X}T^{-1}$.

The elimination method described in 6.3 does not require the eigenvalues or Jordan Form of P to be determined but does require the computation of the characteristic equation of P. A system of m equations in m unknowns is obtained by comparing coefficients. Since the number of solutions, if finite of $X^2 = P$ is 2^m where X and P are m×m matrices, it soon becomes difficult to solve the m equations algebraically but methods such as Newton may be applied.

The advantage of this method is that there are only m unknowns rather than m^2 values to be determined. Eliminating X completely has advantages over the elimination method described in 4.3 where it is necessary to find the roots of $\det[\lambda^2 I - P] = 0$. As shown in example 6.2.3, there can be real matrix solutions of the equation $X^2 - P = 0$ and yet no real roots of $\det[\lambda^2 I - P] = 0$. This problem does not arise in this method as the coefficients of the characteristic polynomial of a real matrix are also real.

In conclusion it may be said that this elimination method seems to compare favourably with known algebraic and iterative methods of evaluating the square root of a matrix.

CHAPTER 7

Conclusion

Various methods of solution of matrix equations have been considered in this thesis. An attempt has been made to find matrix solutions by consideration of the equivalent system of constituent equations. Decision methods, multivariable resultants and iterative methods have been applied. Though particular solutions may be obtained by these methods, computational difficulties exist for large order matrices. Since the number of variables to be obtained is m^2 when the unknown matrix is m×m, it becomes difficult to draw any general conclusions about the matrical properties of the coefficient matrices. This seems to indicate that methods which are applied direct to the matrix equation offer many advantages.

The elimination method described in Chapter 4 has been shown to be successful for matrix equations of order 2×2 and 3×3. The adaptation of the method for the solution of the matrix quadratic equation involving m×m matrices has also been described. The computational difficulties involved do not seem to be any greater than many of the methods described in Chapter 3. The viability of the process for large order matrices using computer methods is a possible area for further investigation.

One of the iterative processes described in Chapter 5 is an adaptation of this elimination method, using the characteristic equation of a solution X. The method has been shown to lead to solutions for a variety of equations involving 2×2 and 3×3 matrices.

The method offers scope for further investigation such as whether the method is stable for large order matrices, conditions for convergence and choice of initial matrix.

In Chapter 6 a method for finding the square root of a given matrix was described. A set of equations connecting the characteristic coefficients of X and P was obtained. An area for further work is the possibility of extending the method to the solution of the unilateral matrix equation $X^2 + A_1X + A_2 = 0$.

As stated in earlier chapters, work on matrix equations in recent years has concentrated on the numerical solution of the Lyapunov and Riccati equations. Many approaches are variations of the eigenvector or Schur vector methods described in Chapter 3. It is possible that some early methods such as that described by Ingraham [1941] could now be adapted for modern computer techniques by use of computer algebra. This is another area for further work.

235

References

BAKER, H.F. [1925]. 'The reciprocation of one quadratic into another'. Proc. Cambridge Philos. Soc. Vol. 23, pp.22-27.

BELL, J. [1950]. 'Families of solutions of the unilateral matrix equation'. Proc. Amer. Math. Soc. Vol. I, pp.151-159.

BUCCHEIM, A. [1884]. 'On the theory of matrices'. Proc. London Math. Soc. Vol. 16, pp.63-81.

CAYLEY, A. [1858]. 'A memoir on the theory of matrices'. Philos. Trans. Roy. Soc. London. Vol. 148, pp.17-37.

CHARLIER, J.P., & P. VAN DOOREN. [1987]. 'A systolic algorithm for Riccati and Lyapunov equations'. Philips Research Laboratory, Brussels.

DAVIS, G.J. [1981]. 'Numerical solution of a quadratic matrix equation'. S.I.A.M. J. Sci. Stat. Comput. Vol. 2, pp.164-175.

DENNIS, J.E., J.F. TRAUB & R.P. WEBER. [1976]. 'The algebraic theory of matrix polynomials'. S.I.A.M. J. Numer. Anal. Vol. 13, pp.831-845.

DENNIS, J.E., J.F. TRAUB & R.P. WEBER. [1978]. 'Algorithms for solvents of matrix polynomials'. S.I.A.M. J. Numer. Anal. Vol. 15, pp.523-533.

DICKSON, L.E. [1926]. Modern Algebraic Theories. Chicago.

FRANKLIN, P. [1932]. 'Algebraic Matrix equations'. J. Math. Physics. Mass. Inst. Technol. Vol. 10. pp.289-310.

FREESTED, W.C., R.F. WEBBER & R.W. BASS. [1968]. 'The "GASP" computer program - an integrated tool for optimal control and filter design'. Preprints Joint Automatic Control Conference. Ann. Arbor, Michigan, pp.198-202.

FROBENIUS, F. [1896]. 'Über die cogredienten transformation der bilinearen Formen'. S.B. Preu. Akad. Wiss. pp.7-16.

GANTMACHER, F.R. [1959]. The Theory of Matrices, Vol. I. Chelsea Pub. Co.

HODGE, W.V.D. & D. PEDOE. [1947]. Methods of Algebraic Geometry. Vol. I, Cambridge Univ. Press.

INCERTIS, F. [1983]. 'An extension on a new formulation of the algebraic Riccati equation problem'. I.E.E.E. Trans. Aut. Control. Vol. AC 28, pp.235-238.

INGRAHAM, M.H. [1934]. 'On the rational solution of the matrix equation P(X) = A'. Journal of Mathematics and Physics, Vol. 13, pp.46-50.

INGRAHAM, M.H. [1941]. 'Rational methods in matrix equations'. Bull. Amer. Math. Soc. Vol. 47, pp.61-70.

KLEINMAN, D.L. [1968]. 'On an iterative technique for Riccati equation computations'. I.E.E.E. Trans. Aut. Control. Vol. AC 13, pp.114-115.

KREIS, H. [1906]. 'Contribution à la theorie des systemes lineaires'. Thesis, Zurich.

LANCASTER, P. [1966]. Lambda matrices and vibrating systems. Pergamon Press.

LAUB, A.J. [1979]. 'A Schur method for solving algebraic Riccati equations'. I.E.E.E. Trans. Aut. Control. Vol. 24, pp.913-921.

MORRIS, J.L. [1983]. Computational Methods in elementary numerical analysis. John Wiley, p.304.

MOSES, J. [1966]. 'Solutions of systems of polynomial equations by elimination'.
Comm. Ass. Comput. Mach. Vol. 9, No.8, pp.634-637.

POPEEA, C. & L. LUPAS. [1976]. 'Matrix equation solutions by the Matrix sign function'.
Rev. Roum. Sci. Techn. - Électrotechn. et Énerg., Vol. 21, No.2, pp.281-290.

POTTER, J. [1966]. 'Matrix quadratic solutions'.
S.I.A.M. J. Appl. Math. Vol. 14, pp.496-501.

ROTH, W.E. [1928]. 'A solution of the matric equation P(X) = A.'
Trans. Amer. Math. Soc. Vol. 30, pp.579-596.

ROTH, W.E. [1930]. 'On the unilateral equation in Matrices'.
Trans. Amer. Math. Soc. Vol. 32, pp.61-80.

ROTH, W.E. [1950]. 'On the matric equation $X^2 + AX + XB + C = 0$'.
Proc. Amer. Math. Soc. Vol. I, pp.586-589.

SEIDENBERG, A. [1954]. 'A new decision method for elementary algebra'.
Annals of Mathematics. Vol. 60, pp.365-374.

SYLVESTER, J. [1884][A] 'Sur les quantities formant un groupe de nonions analogues aux quaternions de Hamilton'.
C.R. Acad. Sci. Paris. Vol. 98, pp.471-475.

SYLVESTER, J. [1884][B] 'Sur ℓ'equation en matrices PX = XQ'.
C.R. Acad. Sci. Paris. Vol. 99, pp.67-71.

SYLVESTER, J. [1884][C] 'Sur ℓa resolution générale de ℓ'equation lineaire en matrices d'un order quelconque'.
C.R. Acad. Sci. Paris. Vol. 99, pp.409-412
pp.432-436.

SYLVESTER, J. [1884][D] 'Sur la solution explicite de l'equation quadratique de Hamilton en quaternions ou en matrices du second ordre'.
C.R. Acad. Sci. Paris. Vol. 99, pp.555-557.

SYLVESTER, J. [1884][E] 'Sur les conditions de l'existence de recines égales dans l'equation du second degré de Hamilton et sur une méthode générale pour resoudre une équation unilaterale de n'importe quel degré en matrices d'un ordre quelconque'.
C.R. Acad. Sci. Paris. Vol. 99, pp.621-631.

TARSKI, A. [1951]. A Decision Method for Elementary Algebra and Geometry.
Univ. Calif. Press.

TRAUB, J.F. [1966]. 'A class of globally convergent iteration functions for the solution of polynomial equations'.
Math. Comput. Vol. 20, pp.113-138.

WILLIAMS, L.H. [1962]. 'Algebra of polynomials in several variables for a digital computer'.
J.A.C.M. Vol. 9, pp.29-40.

Appendix A

Formal Proofs

The result given on page 206 can be proved formally as follows.

Theorem I.

Let X be an m×m matrix which satisfies the equation $X^2 = P$, and let the characteristic polynomials of X and P be $\sum_{i=0}^{m} a_i \lambda^{m-i}$ and $\sum_{i=0}^{m} \alpha_i \lambda^{m-1}$ respectively where $a_o = \alpha_o = 1$. Then the characteristic coefficients of X and P are connected by the relation

$$\sum_{j+k=2i} a_j a_k (-1)^j = \alpha_i, \qquad i = 1,2,\ldots,m, \quad j,k = 0,1,\ldots,m \ .$$

Proof.

Since X satisfies its own characteristic equation

$$a_o X^m + a_1 X^{m-1} + a_2 X^{m-2} + \ldots + a_r X^{m-r} + \ldots + a_{m-1} X + a_m I = 0 \ .$$

Substituting P for X^2 in this equation, then X satisfies the linear equation

$$A_o X + A_1 = 0 \qquad (1)$$

where A_o, A_1 are scalar polynomials in the matrix P.

Multiplying equation (1) on the right by X and substituting P for X^2 again, then X also satisfies the linear equation

$$A_1 X + A_o P = 0 \ . \qquad (2)$$

Combining equations (1) and (2) by multiplying (1) on the left by A_1 and (2) on the left by A_o and subtracting, then X is

eliminated completely since $A_1 A_o = A_o A_1$.

The resulting equation is

$$A_1^2 - A_o^2 P = 0 \quad . \tag{3}$$

It is now shown that this is an equation in P of degree m, whether or not m is odd or even.

If m is even, then m = 2k for some positive integer k.

Then

$$A_o = [a_1 P^{k-1} + a_3 P^{k-2} + \ldots \; a_{2r-1} P^{k-r} + \ldots \; a_{m-1} I]$$

$$A_1 = [a_o P^k + a_2 P^{k-1} + \ldots \; a_{2r} P^{k-r} + \ldots \; a_m I]$$

$$A_1^2 - A_o^2 P = A_1^2 - (A_o P^{\frac{1}{2}})^2$$

$$= [A_1 - A_o P^{\frac{1}{2}}][A_1 + A_o P^{\frac{1}{2}}] \quad \text{since } A_1,\ A_o,\ P^{\frac{1}{2}} \text{ all commute}$$

$$A_1 - A_o P^{\frac{1}{2}} = [a_o P^k - a_1 P^{k-\frac{1}{2}} + a_2 P^{k-1} + \ldots + (-1)^r a_r P^{k-r/2} + \ldots + a_m I]$$

$$A_1 + A_o P^{\frac{1}{2}} = [a_o P^k + a_1 P^{k-\frac{1}{2}} + a_2 P^{k-1} + \ldots + a_r P^{k-r/2} + \ldots + a_m I] \; .$$

Hence $[A_1 - A_o P^{\frac{1}{2}}][A_1 + A_o P^{\frac{1}{2}}]$

$$= a_o a_o P^{2k} + [a_o a_2 - a_1 a_1 + a_2 a_o] P^{2k-1}$$

$$+ [a_o a_4 - a_1 a_3 + a_2 a_2 - a_3 a_1 + a_4 a_o] P^{2k-2}$$

$$+ \ldots + [a_o a_{2r} - a_1 a_{2r-1} + a_2 a_{2r-2} - a_3 a_{2r-3} + \ldots + a_{2r} a_o] P^{2k-r}$$

$$+ \ldots + a_m^2 I$$

and since $a_o = 1$ and $m = 2k$ ..

3a

$$A_1^2 - A_o^2 P = P^m + [a_o a_2 - a_1 a_1 + a_2 a_o]P^{m-1}$$

$$+ [a_o a_4 - a_1 a_3 + a_2 a_2 - a_3 a_1 + a_4 a_o]P^{m-2}$$

$$+ \ldots + [\sum_{j+k=2r} a_j a_k (-1)^j]P^{m-r} + \ldots + a_m^2 I = 0. \qquad (4)$$

Considering now the case where m is an odd integer.

Let m = 2k-1 for some positive integer k.

Then

$$A_o = [a_o P^{k-1} + a_2 P^{k-2} + a_4 P^{k-3} + \ldots + a_{2r-2} P^{k-r} + \ldots + a_{m-1} I]$$

$$A_1 = [a_1 P^{k-1} + a_3 P^{k-2} + a_5 P^{k-3} + \ldots + a_{2r-1} P^{k-r} + \ldots \quad a_m I] .$$

Writing equation (3) as $A_o^2 P - A_1^2 = 0$, this may be factorized as

$$[A_o P^{\frac{1}{2}} - A_1][A_o P^{\frac{1}{2}} + A_1] = 0$$

$$A_o P^{\frac{1}{2}} - A_1 = [a_o P^{k-\frac{1}{2}} - a_1 P^{k-1} + a_2 P^{k-3/2} + \ldots + (-1)^r a_r P^{k-\frac{(r+1)}{2}}$$

$$+ \ldots + a_m I]$$

$$A_o P^{\frac{1}{2}} + A_1 = [a_o P^{k-\frac{1}{2}} + a_1 P^{k-1} + a_2 P^{k-3/2} + \ldots + a_r P^{k-\frac{(r+1)}{2}}$$

$$+ \ldots + a_m I]$$

Hence $[A_o P^{\frac{1}{2}} - A_1][A_o P^{\frac{1}{2}} + A_1]$

$$= a_o a_o P^{2k-1} + [a_o a_2 - a_1 a_1 + a_2 a_o]P^{2k-2}$$

$$+ [a_o a_4 - a_1 a_3 + a_2 a_2 - a_3 a_1 + a_4 a_o]P^{2k-3}$$

$$+ \ldots + [a_o a_{2r} - a_1 a_{2r-1} + a_2 a_{2r-2}$$

$$+ \ldots + a_{2r-1} a_1 + a_{2r} a_o]P^{(2k-1)-r} + \ldots + a_m^2 I$$

and since $a_o = 1$ and $m = 2k-1$

$$A_o^2 P - A_1^2 = P^m + [a_o a_2 - a_1 a_1 + a_2 a_o]P^{m-1}$$

$$+ [a_o a_4 - a_1 a_3 + a_2 a_2 - a_3 a_1 + a_4 a_o]P^{m-2}$$

$$+ \ldots + \Big[\sum_{j+k=2r} a_j a_k (-1)^j\Big]P^{m-r} + \ldots + a_m^2 I = 0$$

and this is identical to equation (4).

Comparing the coefficients of this polynomial in P with the characteristic polynomial of P gives

$$\sum_{j+k=2r} a_j a_k (-1)^j = \alpha_i \qquad i = 1,2,\ldots,m; \quad j,k = 0,1,\ldots,m$$

which proves the theorem.

The result given on page 214 can be proved formally as follows.

Theorem II.

If X is an m×m matrix with characteristic polynomial $\sum_{i=0}^{m} a_i \lambda^{m-i}$ where $a_o = 1$, and C_x is the companion form of X

then $$\det[C_x^2 - \lambda I] = \Big\{a_m + a_{m-2}\lambda + a_{m-4}\lambda^2 + \ldots a_{m-2r}\lambda^r + \ldots a_{m-2[\frac{m}{2}]}\lambda^{[\frac{m}{2}]}\Big\}^2$$

$$- \lambda\Big\{a_{m-1} + a_{m-3}\lambda + \ldots a_{m-2r-1}\lambda^r + \ldots a_{m-2[\frac{m-1}{2}]-1}\lambda^{[\frac{m-1}{2}]}\Big\}^2.$$

Proof.

Let $\lambda = \mu^2$.

Then

$$\det[C_x^2 - \lambda I] = \det[C_x^2 - \mu^2 I]$$

$$= \det[C_x - \mu I].\ \det[C_x + \mu I] \quad .$$

Two cases are now considered a) m is an even integer,

b) m is an odd integer .

Case a.

Let $m = 2k$ for some positive integer k.

Then

$$\det[C_x - \mu I] = a_o\mu^{2k} + a_1\mu^{2k-1} + a_2\mu^{2k-2} + \ldots + a_{2k}$$

$$\det[C_x + \mu I] = a_o\mu^{2k} - a_1\mu^{2k-1} + a_2\mu^{2k-2} - a_3\mu^{2k-3} + \ldots + a_{2k}$$

$$\det[C_x - \mu I].\ \det[C_x + \mu I] = \left[\sum_{r=0}^{2k} a_r\mu^{2k-r}\right]\left[\sum_{r=0}^{2k} (-1)^r a_r\mu^{2k-r}\right]$$

$$= \left[\left(\sum_{r=0}^{k} a_{2r}\mu^{2k-2r}\right) + \left(\sum_{r=1}^{k} a_{2r-1}\mu^{2k-(2r-1)}\right)\right]\left[\left(\sum_{r=0}^{k} a_{2r}\mu^{2k-2r}\right)\right.$$

$$\left. - \left(\sum_{r=1}^{k} a_{2r-1}\mu^{2k-(2r-1)}\right)\right]$$

$$= \left[\sum_{r=0}^{k} a_{2r}(\mu^2)^{k-r}\right]^2 - \mu^2\left[\sum_{r=1}^{k} a_{2r-1}(\mu^2)^{k-r}\right]^2 \quad .$$

Now substituting $\lambda = \mu^2$, the expression becomes

$$\left[\sum_{r=0}^{k} a_{2r}\lambda^{k-r}\right]^2 - \lambda\left[\sum_{r=1}^{k} a_{2r-1}\lambda^{k-r}\right]^2$$

and since $k = \frac{m}{2}$

$$\det[C_x^2 - \lambda I] = \left[\sum_{r=0}^{\frac{m}{2}} a_{2r}\lambda^{\frac{m}{2}-r}\right]^2 - \lambda\left[\sum_{r=1}^{\frac{m}{2}} a_{2r-1}\lambda^{\frac{m}{2}-r}\right]^2 .$$

Case b.

Let $m = 2k-1$ for some positive integer k.

Then

$$\det[C_x - \mu I] = -a_o\mu^{2k-1} - a_1\mu^{2k-2} - a_2\mu^{2k-3} + \ldots - a_{2k-2}\mu - a_{2k-1}$$

$$\det[C_x + \mu I] = a_o\mu^{2k-1} - a_1\mu^{2k-2} + a_2\mu^{2k-3} + \ldots + a_{2k-2}\mu - a_{2k-1}$$

$$\therefore \det[C_x - \mu I].\det[C_x + \mu I] = \left[-\sum_{r=0}^{2k-1} a_r\mu^{2k-(r+1)}\right]\left[\sum_{r=0}^{2k-1}(-1)^r a_r\mu^{2k-(r+1)}\right]$$

$$= \left[\left(-\sum_{r=1}^{k} a_{2r-1}\mu^{2k-2r}\right) - \left(\sum_{r=1}^{k} a_{2r-2}\mu^{2k-(2r-1)}\right)\right]\left[\left(-\sum_{r=1}^{k} a_{2r-1}\mu^{2k-2r}\right) + \left(\sum_{r=1}^{k} a_{2r-2}\mu^{2k-(2r-1)}\right)\right]$$

$$= \left[\sum_{r=1}^{k} a_{2r-1}(\mu^2)^{k-r}\right]^2 - \mu^2\left[\sum_{r=1}^{k} a_{2r-2}(\mu^2)^{k-r}\right]^2 .$$

Now substituting $\lambda = \mu^2$ the expression becomes

$$\left[\sum_{r=1}^{k} a_{2r-1}\lambda^{k-r}\right]^2 - \lambda\left[\sum_{r=1}^{k} a_{2r-2}\lambda^{k-r}\right]^2$$

and replacing k by $\frac{m+1}{2}$

$$\det[C_x^2 - \lambda I] = \left[\sum_{r=1}^{\frac{m+1}{2}} a_{2r-1}\lambda^{\frac{m+1}{2}-r}\right]^2 - \lambda\left[\sum_{r=1}^{\frac{m+1}{2}} a_{2r-2}\lambda^{\frac{m+1}{2}-r}\right]^2 .$$

In both cases a and b

$$\det[C_x^2 - \lambda I] = \left\{a_m + a_{m-2}\lambda + a_{m-4}\lambda^2 + \ldots a_{m-2r}\lambda^r + \ldots \right.$$

$$\left. a_{m-2[\frac{m}{2}]}\lambda^{[\frac{m}{2}]}\right\}^2$$

$$- \lambda\left\{a_{m-1} + a_{m-3}\lambda + \ldots + a_{m-2r-1}\lambda^r + \ldots a_{m-2[\frac{m-1}{2}]}\lambda^{[\frac{m-1}{2}]}\right\}^2 ,$$

$$a_o = 1 .$$

The result given on page 100 may be proved formally as follows.

Theorem III.

Let X be an m×m matrix which satisfies the equation $X^2 + A_1X + A_2 = 0$, and let the characteristic polynomial of X be $\sum_{i=0}^{m} a_i\lambda^{m-i}$ where $a_o = 1$.

Then X satisfies the linear equation

$$J_{m-1}X + K_{m-1} = 0 \quad \text{where} \quad J_1 = a_1I - A_1, \quad K_1 = a_2I - A_2$$

$$\text{and} \quad J_{i+1} = K_i - J_iA_1, \quad K_{i+1} = a_{i+2}I - J_iA_2 .$$

Proof.

Since X satisfies its own characteristic equation, then X simultaneously satisfies the two equations

$$X^m + a_1X^{m-1} + a_2X^{m-2} + \ldots + a_{m-2}X^2 + a_{m-1}X + a_mI = 0 \qquad (1)$$

$$X^2 + A_1X + A_2 = 0 \quad . \qquad (2)$$

Multiplying equation (2) on the right by X^{m-2} and subtracting from equation (1) gives

$$[a_1I - A_1]X^{m-1} + [a_2I - A_2]X^{m-2} + a_3X^{m-3} + \ldots a_{m-2}X^2 + a_{m-1}X + a_mI = 0$$

and defining $J_1 = a_1I - A_1$ and $K_1 = a_2I - A_2$, then X satisfies the equation

$$J_1X^{m-1} + K_1X^{m-2} + a_3X^{m-3} + \ldots + a_{m-2}X^2 + a_{m-1}X + a_mI = 0 \quad . \quad (3)$$

Multiplying equation (2) on the left by J_1 and on the right by X^{m-3} and subtracting from equation (3) gives

$$[K_1 - J_1A_1]X^{m-2} + [a_3I - J_1A_2]X^{m-3} + a_4X^{m-4} + \ldots + a_{m-1}X + a_mI = 0$$

and defining $J_2 = K_1 - J_1A_1$ and $K_2 = a_3I - J_1A_2$, then X satisfies the equation

$$J_2X^{m-2} + K_2X^{m-3} + a_4X^{m-4} + a_5X^{m-5} + \ldots + a_{m-1}X + a_mI = 0 \quad . \quad (4)$$

After applying the process i times an equation of degree (m-i) is obtained

$$J_iX^{m-i} + K_iX^{m-(i+1)} + a_{i+2}X^{m-(i+2)} + \ldots + a_{m-1}X + a_mI = 0 \ldots\ldots (i+2) \quad .$$

Multiplying equation (2) on the right by $X^{m-(i+2)}$ and on the left

by J_i and subtracting from equation (i+2) gives

$$[K_i - J_iA_1]X^{m-(i+1)} + [a_{i+2}I - J_iA_2]X^{m-(i+2)} + a_{i+3}X^{m-(i+3)}$$

$$+ \ldots + a_{m-1}X + a_mI = 0$$

and hence X satisfies the equation

$$J_{i+1}X^{m-(i+1)} + K_{i+1}X^{m-(i+2)} + a_{i+3}X^{m-(i+3)} + \ldots + a_{m-1}X + a_mI = 0 \ldots . (i+3)$$

where $J_{i+1} = K_i - J_iA_1$ and $K_{i+1} = a_{i+2}I - J_iA_2$.

The process is continued until a linear equation in X is obtained.

Then $J_{m-1}X + K_{m-1} = 0$, and the proof is complete.

Appendix B

Section α.

Extension of the method of Chapter 6 to the case of $X^2 + A_1X + A_2 = 0$.

Difficulties are encountered in attempting to apply the method to the general unilateral equation. Some special cases are considered in this appendix.

I The equation $X^2 + A_1X + A_2 = 0$ where $A_1 = kI$ for some scalar k.

If $A_1 = kI$, the equation may be written in the form

$$(X + \frac{k}{2} I)^2 + A_2 - \frac{k^2}{4} I = 0$$

and setting $Z = X + \frac{k}{2} I$ and $R = \frac{k^2}{4} I - A_2$

the equation becomes $Z^2 = R$.

The characteristic coefficients of the matrices Z and R are connected by the relations defined in Chapter 6, i.e. if the characteristic polynomials of Z and R are $\sum_{i=0}^{m} z_i \lambda^{m-i}$ and $\sum_{i=0}^{m} r_i \lambda^{m-i}$ respectively where $z_o = r_o = 1$

then $$\sum_{j+k=2i} z_j z_k (-1)^j = r_i \ , \qquad i = 1,2,\ldots,m, \quad j,k = 0,1,\ldots,m.$$

A relationship also exists between the characteristic coefficients of Z and X and also between R and A_2.

Let the characteristic polynomials of X and A_2 be $\sum_{i=0}^{m} a_i \lambda^{m-i}$ and $\sum_{i=0}^{m} \beta_i \lambda^{m-i}$ where $a_o = \beta_o = 1$

$$\det[Z - \lambda I] = \lambda^m + z_1\lambda^{m-1} + z_2\lambda^{m-2} + \ldots + z_m \ .$$

But $\det[Z - \lambda I] = \det[X - (\lambda - \frac{k}{2})I]$

$$= (\lambda - \tfrac{k}{2})^m + a_1(\lambda - \tfrac{k}{2})^{m-1} + \ldots + a_m \ .$$

Comparing coefficients of λ^i

$$z_n = \sum_{i=0}^{n} {}^{m-n+i}C_i \ a_{n-i} \left(\frac{k}{2}\right)^i (-1)^i \qquad n = 1,2,\ldots,m.$$

Also $\det[R - \lambda I] = \lambda^m + r_1\lambda^{m-1} + r_2\lambda^{m-2} + \ldots + r_m$

$$\det[R - \lambda I] = \det[-A_2 - \left(\lambda - \frac{k^2}{4}\right)I]$$

$$= \left(\lambda - \frac{k^2}{4}\right)^m - \beta_1\left(\lambda - \frac{k^2}{4}\right)^{m-1} + \beta_2\left(\lambda - \frac{k^2}{4}\right)^{m-2} + \ldots(-1)^m\beta_m \ .$$

Comparing coefficients of λ^i

$$r_n = (-1)^n \sum_{i=0}^{n} {}^{m-n+i}C_i \ \beta_{n-i} \left(\frac{k^2}{4}\right)^i \qquad n = 1,2,\ldots,m \ .$$

A set of equations connecting the characteristic coefficients of X and A_2 may therefore be obtained from

$$\sum_{j+k=2i} z_j z_k (-1)^j = r_i \qquad i = 1,2,\ldots,m, \quad j,k = 0,1,\ldots,m$$

where $z_n = \sum_{i=0}^{n} {}^{m-n+i}C_i \ a_{n-i} \left(\frac{k}{2}\right)^i (-1)^i \qquad n = 1,2,\ldots,m$.

and $r_n = (-1)^n \sum_{i=0}^{n} {}^{m-n+i}C_i \ \beta_{n-i} \left(\frac{k^2}{4}\right)^i \qquad n = 1,2,\ldots,m.$

For $m > 4$ the computational difficulties involved in computing the direct relationship between the a_i and β_i become prohibitive. The equations are given for $m = 2,3,4$ as follows.

2×2 Case.

$$a_1^2 - 2a_2 - a_1k = \beta_1$$

$$a_2^2 - a_1a_2k + a_2k^2 = \beta_2 \quad .$$

3×3 Case.

$$a_1^2 - 2a_2 - a_1k = \beta_1$$

$$a_2^2 - 2a_1a_3 + [3a_3 - a_1a_2]k + a_2k^2 = \beta_2$$

$$a_3^2 - a_2a_3k + a_1a_3k^2 - a_3k^3 = \beta_3 \quad .$$

4×4 Case.

$$a_1^2 - 2a_2 - a_1k = \beta_1$$

$$a_2^2 - 2a_1a_3 + 2a_4 + [3a_3 - a_1a_2]k + a_2k^2 = \beta_2$$

$$a_3^2 - 2a_2a_4 + [3a_1a_4 - a_2a_3]k + [a_1a_3 - 4a_4]k^2 - a_3k^3 = \beta_3$$

$$a_4^2 - a_3a_4k + a_2a_4k^2 - a_1a_4k^3 + a_4k^4 = \beta_4$$

II The equation $X^2 + A_1X + A_2 = 0$ where $A_1A_2 = A_2A_1$.

Consider the case where X, A_1, A_2 are m×m matrices.

Let the characteristic equation of X be $\lambda^m + a_1\lambda^{m-1} + ... + a_m = 0$.

Then X satisfies the two equations

$$X^m + a_1X^{m-1} + a_2X^{m-2} + ... + a_{m-1}X + a_mI = 0$$

$$X^2 + A_1X + A_2 = 0 .$$

The process described in Appendix A, Theorem III may be applied until a linear equation in X is obtained

$$J_{m-1}X + K_{m-1} = 0 \quad \text{where} \quad J_1 = a_1I - A_1 \qquad K_1 = a_2I - A_2$$

$$\text{and} \qquad J_{i+1} = K_i - J_iA_1 \quad K_{i+1} = a_{i+2}I - J_iA_2 .$$

Continuing the process, a second linear equation may be obtained

$$[K_{m-1} - J_{m-1}A_1]X - J_{m-1}A_2 = 0$$

and since A_1 and A_2 commute, X may be eliminated completely to obtain

$$K_{m-1}^2 - J_{m-1}A_1K_{m-1} + J_{m-1}^2A_2 = 0 .$$

In the method of Chapter 6, the final equation obtained after the complete elimination of X was a matrix polynomial equation in the known matrix P and relationships could be obtained by comparison of the scalar coefficients with the known characteristic coefficients of P.

In this case however, the final equation involves the two

matrices A_1, A_2 and hence comparison with the characteristic coefficients is not possible.

In the 2×2 case, it is possible, using the known matrices A_1, A_2 to derive the set of constituent equations and solve for a_1, a_2.

For example, if $A_1 = \begin{pmatrix} -1 & 2 \\ 1 & -1 \end{pmatrix}$ and $A_2 = \begin{pmatrix} -4 & -4 \\ -2 & -4 \end{pmatrix}$

then $J_1 = \begin{pmatrix} a_1+1 & -2 \\ -1 & a_1+1 \end{pmatrix}$ $K_1 = \begin{pmatrix} a_2+4 & 4 \\ 2 & a_2+4 \end{pmatrix}$

and the final equation $K_1^2 - J_1A_1K_1 + J_1^2 A_2 = 0$

becomes

$$\begin{pmatrix} a_2^2+8a_2+24 & 8a_2+32 \\ 4a_2+16 & a_2^2+8a_2+24 \end{pmatrix} - \begin{pmatrix} -a_1a_2-3a_2-4 & 2a_1a_2+4a_1+4a_2+4 \\ a_1a_2+2a_1+2a_2+2 & -a_1a_2-3a_2-4 \end{pmatrix}$$

$$+ \begin{pmatrix} -4a_1^2-4 & -4a_1^2+8a_1+4 \\ -2a_1^2+4a_1+2 & -4a_1^2-4 \end{pmatrix} = 0 \quad .$$

The constituent equations from this are

$$a_2^2 + 11a_2 + a_1a_2 - 4a_1^2 + 24 = 0 \qquad (1)$$

$$4a_2 - 2a_1a_2 + 4a_1 - 4a_1^2 + 32 = 0 \qquad (2)$$

$$2a_2 - a_1a_2 + 2a_1 - 2a_1^2 + 16 = 0 \qquad (3)$$

$$a_2^2 + 11a_2 + a_1a_2 - 4a_1^2 + 24 = 0 \qquad (4)$$

In this particular case there is linear dependence between the equations since equations (1) and (4) are the same and equation (2) is 2 × equation (3).

From equations (1) and (3), using elimination methods, the following equations are obtained

$$[3a_2 - 4]a_1 + a_2^2 + 7a_2 - 8 = 0 \qquad (5)$$

$$[a_2 + 1][a_2 + 8][a_2^2 - 8a_2 + 8] = 0 \quad . \qquad (6)$$

There are 4 solutions for equation (6) : -1. -8, $4+2\sqrt{2}$, $4-2\sqrt{2}$. Substitution of these values in equation (5) gives 4 corresponding values for a_1 : -2, 0, $-1-3\sqrt{2}$, $-1+3\sqrt{2}$.

From these four pairs of solutions, four matrix solutions may be obtained from $J_1X + K_1 = 0$ where $J_1 = a_1I - A_1$, $K_1 = a_2I - A_2$. These are

$$X = \begin{pmatrix} 1 & 2 \\ 1 & 1 \end{pmatrix} \qquad X = \begin{pmatrix} 0 & -4 \\ -2 & 0 \end{pmatrix} \qquad X = \frac{1}{4}\begin{pmatrix} 2+6\sqrt{2} & -4+2\sqrt{2} \\ -2+\sqrt{2} & 2+6\sqrt{2} \end{pmatrix}$$

$$X = \frac{1}{4}\begin{pmatrix} 2-6\sqrt{2} & -4-2\sqrt{2} \\ -2-\sqrt{2} & 2-6\sqrt{2} \end{pmatrix} \quad .$$

In the 3×3 case, the final equation obtained when X has been eliminated is

$$K_2^2 - J_2A_1K_2 + J_2^2A_2 = 0 \quad \text{where} \quad J_1 = a_1I - A_1 \qquad K_1 = a_2I - A_2$$

$$J_2 = K_1 - J_1A_1 \qquad K_2 = a_3I - J_1A_2 \quad .$$

This is equivalent to

$$A_2^3 + (a_1^2 - 2a_2)A_2^2 + (a_2^2 - 2a_1a_3)A_2 + a_3^2 I - a_3A_1^3 + a_1a_3A_1^2$$

$$- a_2a_3A_1 - a_1A_1A_2^2 + a_2A_1^2A_2 + (3a_3 - a_1a_2)A_1A_2 = 0 .$$

The matrices A_1, A_2 are known matrices, and therefore the original matrix equation containing 9 unknown elements in the matrix X has been replaced by a matrix equation containing only 3 unknowns, a_1, a_2, a_3 which could be calculated using numerical methods.

In the m×m case the matrix equation $X^2 + A_1X + A_2 = 0$ where $A_1A_2 = A_2A_1$ contains m^2 unknowns in the matrix X. Using elimination methods this may be replaced by

$$K_{m-1}^2 - J_{m-1}A_1K_{m-1} + J_{m-1}^2A_2 = 0 \quad \ldots \; J_1 = a_1I - A_1 \qquad K_1 = a_2I - A_2$$

$$J_{i+1} = K_i - J_iA_1 \qquad K_{i+1} = a_{i+2}I - J_iA_2$$

$$i = 1 \text{ to } m-2$$

which contains only m unknowns $a_1, a_2, \ldots, a_m$ and numerical methods could be applied to obtain a solution.

III The equation $X^2 + A_1X + A_2 = 0$ where X, A_1, A_2 are 2×2 matrices.

The problems involved in the application of the method of Chapter 6 to the general unilateral quadratic matrix equation have not as yet been resolved.

It can however, be applied to the equation involving 2×2 matrices and is described in this section.

Let the characteristic equation of X be $\lambda^2 + a_1\lambda + a_2 = 0$. Then X satisfies the two equations

$$X^2 + A_1X + A_2 = 0 \qquad (1)$$

$$X^2 + a_1X + a_2I = 0 \quad . \qquad (2)$$

Eliminating X^2, two linear equations in X may be obtained

$$[A_1 - a_1I]X + [A_2 - a_2I] = 0 \qquad (3)$$

$$[A_2 - a_2I]X + [a_1A_2 - a_2A_1] = 0 \quad . \qquad (4)$$

Since A_1 and A_2 do not commute, X cannot be eliminated directly using elimination methods.

However, from equation (3) $X = [A_1 - a_1I]^{-1}[a_2I - A_2]$, provided a_1 is not an eigenvalue of A_1.

But $\det X = a_2$

$$\therefore \quad a_2 = \frac{\det[a_2I - A_2]}{\det[A_1 - a_1I]} = \frac{a_2^2 + \beta_1a_2 + \beta_2}{a_1^2 + \alpha_1a_1 + \alpha_2}$$

where α_1, α_2 and β_1, β_2 are the characteristic coefficients of A_1 and A_2 respectively.

Hence $\quad a_2^2 + \beta_1a_2 + \beta_2 = a_1^2a_2 + \alpha_1a_1a_2 + \alpha_2a_2$.

Also $\quad a_1 = -\text{Trace } X = -\text{Trace } \left\{[A_1 - a_1I]^{-1}[a_2I - A_2]\right\}$

$$\therefore \quad -a_1 = \frac{-\alpha_1a_2 - \beta_1a_1 - 2a_1a_2 + t}{a_1^2 + \alpha_1a_1 + \alpha_2} \quad \text{where } t = \text{Tr}(A_1A_2) - \text{Tr } A_1 . \text{Tr } A_2$$

$$\therefore \quad -a_1^3 - \alpha_1 a_1^2 - \alpha_2 a_1 = -\alpha_1 a_2 - \beta_1 a_1 - 2a_1 a_2 + t \quad .$$

Hence two equations have been obtained in a_1, a_2

$$a_1^3 + \alpha_1 a_1^2 + \alpha_2 a_1 - \alpha_1 a_2 - \beta_1 a_1 - 2a_1 a_2 + t = 0 \qquad (5)$$

$$a_1^2 a_2 + \alpha_1 a_1 a_2 + \alpha_2 a_2 - a_2^2 - \beta_1 a_2 - \beta_2 = 0 \qquad (6)$$

and α_1, α_2, β_1, β_2, t can all be evaluated from the coefficient matrices.

If $[A_1 - a_1 I]$ is singular, the same process could be applied with equation (4)

and $a_2 = \dfrac{\det[a_2 A_1 - a_1 A_2]}{\det[A_2 - a_2 I]}$ $\qquad a_1 = -\text{Trace}\left\{[A_2 - a_2 I]^{-1}[a_2 A_1 - a_1 A_2]\right\}$

provided a_2 is not an eigenvalue of A_2.

Alternatively the constituent equations may be formed from

$$[A_2 - a_2 I][A_1 - a_1 I]^{-1}[a_2 I - A_2] = a_2 A_1 - a_1 A_2 \quad \text{if} \quad \det[A_1 - a_1 I] \neq 0$$

or $\quad [A_1 - a_1 I][A_2 - a_2 I]^{-1}[a_2 A_1 - a_1 A_2] = a_2 I - A_2 \quad \text{if} \quad \det[A_2 - a_2 I] \neq 0 \ .$

Let $\quad A_1 = \begin{pmatrix} p_1 & p_2 \\ p_3 & p_4 \end{pmatrix} \qquad A_2 = \begin{pmatrix} q_1 & q_2 \\ q_3 & q_4 \end{pmatrix} \ .$

Then substituting in

$$[A_2 - a_2 I][A_1 - a_1]^{-1}[a_2 I - A_2] = a_2 A_1 - a_1 A_2 \ .$$

$$\begin{pmatrix} q_1-a_2 & q_2 \\ q_3 & q_4-a_2 \end{pmatrix}\begin{pmatrix} \frac{p_4-a_1}{D} & \frac{-p_2}{D} \\ \frac{-p_3}{D} & \frac{p_1-a_1}{D} \end{pmatrix}\begin{pmatrix} a_2-q_1 & -q_2 \\ -q_3 & a_2-q_4 \end{pmatrix}$$

$$= \begin{pmatrix} a_2p_1-a_1q_1 & a_2p_2-a_1q_2 \\ a_2p_3-a_1q_3 & a_2p_4-a_1q_4 \end{pmatrix}$$

where $D = a_1^2 - (p_1 + p_4)a_1 + p_1p_4 - p_2p_3$.

The four equations obtained are

Equation (A).

$$[2p_4q_1 - p_3q_2 - p_2q_3 - p_1(p_1p_4 - p_2p_3)]a_2 + [q_1^2 + q_2q_3 + q_1(p_1p_4 - p_2p_3)]a_1$$

$$+ [- p_1q_1 - p_4q_1]a_1^2 + q_1a_1^3 + [- 2q_1 + p_1(p_1 + p_4)]a_1a_2 - p_1a_1^2a_2$$

$$+ a_1a_2^2 - p_4a_2^2 - p_4q_1^2 + p_3q_1q_2 + p_2q_1q_3 - p_1q_2q_3 = 0.$$

Equation (B).

$$[p_4q_2 - p_2q_1 + p_1q_2 - p_2q_4 - p_2(p_1p_4 - p_2p_3)]a_2$$

$$+[q_1q_2 + q_2q_4 + q_2(p_1p_4 - p_2p_3)]a_1 + [- p_1q_2 - p_4q_2]a_1^2$$

$$+ q_2a_1^3 + [- 2q_2 + p_2(p_1 + p_4)]a_1a_2 - p_2a_1^2a_2$$

$$+ p_2a_2^2 - p_4q_1q_2 + p_3q_2^2 + p_2q_1q_4 - p_1q_2q_4 = 0.$$

Equation (C).

$$[p_4p_3 - p_3q_4 - p_3q_1 + p_1q_3 - p_3(p_1p_4 - p_2p_3)]a_2$$

$$+ [q_1q_3 + q_3q_4 + q_3(p_1p_4 - p_2p_3)]a_1 + [- p_1q_3 - p_4q_3]a_1^2$$

$$+ q_3a_1^3 + [- 2q_3 + p_3(p_1 + p_4)]a_1a_2 - p_3a_1^2a_2$$

$$+ p_3a_2^2 - p_4q_1q_3 + p_3q_1q_4 + p_2q_3^2 - p_1q_3q_4 = 0 \quad .$$

Equation (D).

$$[2p_1q_4 - p_3q_2 - p_2q_3 - p_4(p_1p_4 - p_2p_3)]a_2 + [q_4^2 + q_2q_3 + q_4(p_1p_4 - p_2p_3)]a_1$$

$$+ [- p_1q_4 - p_4q_4]a_1^2 + q_4a_1^3 + [- 2q_4 + p_4(p_1 + p_4)]a_1a_2 - p_4a_1^2a_2$$

$$+ a_2^2a_1 - p_1a_2^2 - p_4q_2q_3 + p_3q_2q_4 + p_2q_3q_4 - p_1q_4^2 = 0 \quad .$$

For matrices of order greater than 2×2 the problem of eliminating X completely arises at the final stage when two linear equations in X have been obtained.

As shown in section II the two linear equations in X which are obtained are

$$J_{m-1}X + K_{m-1} = 0$$

$$[K_{m-1} - J_{m-1}A_1]X - J_{m-1}A_2 = 0 \quad .$$

However since A_1 and A_2 do not commute in general, X is not eliminated completely by continuing the process.

The only possibility is to express X using an inverse matrix from one equation and substitute in the other. The equations obtained are

$$[J_{m-1}A_1 - K_{m-1}]J_{m-1}^{-1} K_{m-1} - J_{m-1} A_2 = 0 \quad \text{provided det } J_{m-1} \neq 0$$

$$J_{m-1}[K_{m-1} - J_{m-1}A_1]^{-1} J_{m-1}A_2 + K_{m-1} = 0 \text{ provided det}[K_{m-1} - J_{m-1}A_1] \neq 0 .$$

Again these matrix equations involve only m unknowns, $a_1 a_2,\ldots,a_m$ rather than the m^2 unknowns contained in the original equation and so might have some advantages if solutions are sought by numerical methods.

Section β.

Extension of the method of Chapter 6 to the equation $X^n = P$.

I The equation $X^3 = P$.

In Chapter 6 a set of equations was obtained relating the characteristic coefficients of X and P in the equation $X^2 = P$. Some progress has been made in attempting to do the same for the equation $X^3 = P$ though an explicit formula has not as yet been found. An attempt to find this is shown in part II of this section.

Let the characteristic polynomials of X and P be $\sum_{i=0}^{m} a_i\lambda^{m-i}$ and $\sum_{i=0}^{m} \alpha_i\lambda^{m-i}$ respectively, where $a_o = \alpha_o = 1$.

Then X satisfies the equation

$$X^m + a_1X^{m-1} + a_2X^{m-2} + \ldots. + a_{m-1}X + a_mI = 0.$$

Since X also satisfies the equation $X^3 = P$, then substituting P for X^3 gives a quadratic equation in X, whose coefficients are scalar polynomials in P. This is

$$A_oX^2 + A_1X + A_2 = 0 \quad \text{where} \quad A_{2-r} = \sum_{i=0}^{[\frac{m-r}{3}]} a_{m-r-3i} P^i, \quad r = 0,1,2 \quad .$$

A second quadratic equation may be obtained by multiplying this equation on the right by X and substituting P for X^3 again.

Hence for m×m matrices, X satisfies the two equations

$$A_oX^2 + A_1X + A_2 = 0 \tag{1}$$

$$A_1X^2 + A_2X + A_oP = 0 \tag{2}$$

where A_o, A_1, A_2 are scalar polynomials in P defined by the given formula.

Eliminating X^2 gives two linear equations in X

$$[A_1^2 - A_oA_2]X + [A_1A_2 - A_o^2P] = 0 \tag{3}$$

$$[2A_oA_1A_2 - A_o^3P - A_1^3]X + [A_oA_2^2 - A_1^2A_2] = 0 \; . \tag{4}$$

Eliminating X completely gives

$$A_o^2[A_o^3P^2 + A_1^3P + A_2^3 - 3A_oA_1A_2P] = 0 \quad .$$

This is a polynomial in P of degree $m + 2[\frac{m-2}{3}]$.

The factors A_o^2 and the expression in square brackets are irreducible factors of degree $2[\frac{m-2}{3}]$ and degree m respectively. Since P satisfies a polynomial equation of degree > m then the minimum polynomial of P must be a factor of this polynomial.

As in Chapter 6, we concentrate on the case where P is non derogatory and take the expression in square brackets to be the minimum and characteristic polynomial of P.

Hence, for non derogatory P, the characteristic coefficients α_i, $i = o,1,\ldots,m$, may be equated with the coefficients in the equation

$$\left\{a_{m-2}I + a_{m-5}P + \ldots a_{(m-2)-3[\frac{m-2}{3}]} P^{[\frac{m-2}{3}]}\right\}^3 P^2$$

$$+ \left\{a_{m-1}I + a_{m-4}P + \ldots\ldots + a_{(m-1)-3[\frac{m-1}{3}]} P^{[\frac{m-1}{3}]}\right\}^3 P$$

$$+ \left\{a_m I + a_{m-3}P + \ldots\ldots + a_{m-3[\frac{m}{3}]} P^{[\frac{m}{3}]}\right\}^3$$

$$- 3\left\{a_{m-2}I + a_{m-5}P + \ldots\right\}\left\{a_{m-1}I + a_{m-4}P + \ldots\right\}$$

$$\left\{a_m I + a_{m-3}P + \ldots\right\} P = 0 \ .$$

The method is illustrated in the following examples.

2×2 Case.

$$A_o = \sum_{i=0}^{o} a_{o-3i}P^i = a_o I, \qquad A_1 = \sum_{i=0}^{o} a_{1-3i}P^i = a_1 I$$

$$A_2 = \sum_{i=0}^{o} a_{2-3i}P^i = a_2 I \ .$$

Substituting in $A_o^3P^2 + A_1^3P + A_2^3 - 3A_oA_1A_2P = 0$ (5)

gives $P^2 + a_1^3 P + a_2^3 I - 3a_1 a_2 P = 0$

$\therefore \quad P^2 + (a_1^3 - 3a_1 a_2)P + a_2^3 I = 0 \quad .$

Equating the coefficients with the characteristic coefficients of P gives the two equations

$$a_1^3 - 3a_1 a_2 = \alpha_1$$

$$a_2^3 = \alpha_2 \quad .$$

3×3 Case.

$$A_o = \sum_{i=0}^{o} a_{1-3i} P^i = a_1 I, \qquad A_1 = \sum_{i=0}^{o} a_{2-3i} P^i = a_2 I,$$

$$A_2 = \sum_{i=0}^{1} a_{3-3i} P^i = a_3 I + a_o P \ .$$

Substitution in (5) gives

$$a_1^3 P^2 + a_2^3 P + (a_3 I + P)^3 - 3a_1 a_2 (a_3 I + P)P = 0$$

$$\therefore \quad P^3 + (a_1^3 - 3a_1 a_2 + 3a_3)P^2 + (a_2^3 - 3a_1 a_2 a_3 + 3a_3^2)P + a_3^3 I = 0 \ .$$

Equating coefficients, the equations obtained are

$$a_1^3 - 3a_1 a_2 + 3a_3 = \alpha_1$$

$$a_2^3 - 3a_1 a_2 a_3 + 3a_3^2 = \alpha_2$$

$$a_3^3 = \alpha_3 \quad .$$

4×4 Case.

$$A_o = \sum_{i=0}^{o} a_{2-3i}P^i = a_2 I, \qquad A_1 = \sum_{i=0}^{1} a_{3-3i}P^i = a_3 I + a_o P,$$

$$A_2 = \sum_{i=0}^{1} a_{4-3i}P^i = a_4 I + a_1 P \ .$$

Substitution in (5) gives

$$a_2^3 P^2 + (a_3 I + P)^3 P + (a_4 I + a_1 P)^3 - 3a_2(a_3 I + P)(a_4 I + a_1 P)P = 0$$

$$\therefore \quad P^4 + (a_1^3 - 3a_1 a_2 + 3a_3)P^3 + (a_2^3 - 3a_1 a_2 a_3 + 3a_3^2 - 3a_2 a_4 + 3a_1^2 a_4)P^2$$

$$+ (a_3^3 - 3a_2 a_3 a_4 + 3a_1 a_4^2)P + a_4^3 I = 0 \ .$$

Equating coefficients, the equations obtained are

$$a_1^3 - 3a_1 a_2 + 3a_3 = \alpha_1$$

$$a_2^3 - 3a_1 a_2 a_3 + 3a_3^2 - 3a_2 a_4 + 3a_1^2 a_4 = \alpha_2$$

$$a_3^3 - 3a_2 a_3 a_4 + 3a_1 a_4^2 = \alpha_3$$

$$a_4^3 = \alpha_4 \ .$$

5×5 Case.

$$A_o = \sum_{i=0}^{1} a_{3-3i}P^i = a_3 I + a_o P, \qquad A_1 = \sum_{i=0}^{1} a_{4-3i}P^i = a_4 I + a_1 P$$

$$A_2 = \sum_{i=0}^{1} a_{5-3i}P^i = a_5 I + a_2 P \ .$$

Substitution in (5) gives

$$(a_3I + P)^3P^2 + (a_4I + a_1P)^3P + (a_5I + a_2P)^3$$

$$- 3(a_3I + P)(a_4I + P)(a_5I + a_2P)P = 0 .$$

Equating coefficients the equations obtained are

$$a_1^3 - 3a_1a_2 + 3a_3 = \alpha_1$$

$$a_2^3 - 3a_1a_2a_3 + 3a_3^2 - 3a_2a_4 + 3a_1^2a_4 - 3a_1a_5 = \alpha_2$$

$$a_3^3 - 3a_2a_3a_4 + 3a_1a_4^2 - 3a_1a_3a_5 - 3a_4a_5 + 3a_2^2a_5 = \alpha_3$$

$$a_4^3 - 3a_3a_4a_5 + 3a_2a_5^2 = \alpha_4$$

$$a_5^3 = \alpha_5 .$$

6×6 Case.

$$A_o = \sum_{i=0}^{1} a_{4-3i}P^i = a_4 I + a_1P , \qquad A_1 = \sum_{i=0}^{1} a_{5-3i}P^i = a_5I + a_2P$$

$$A_2 = \sum_{i=0}^{2} a_{6-3i}P^i = a_6I + a_3P + a_oP^2 .$$

Substitution in (5) gives

$$(a_4I + a_1P)^3P^2 + (a_5I + a_2P)^3P + (a_6I + a_3P + a_oP^2)^3$$

$$- 3(a_4I + a_1P)(a_5I + a_2P)(a_6I + a_3P + a_oP^2)P = 0.$$

Equating coefficients the equations obtained are

$$a_1^3 - 3a_1a_2 + 3a_3 = \alpha_1$$

$$a_2^3 - 3a_1a_2a_3 + 3a_3^2 - 3a_2a_4 + 3a_1^2a_4 - 3a_1a_5 + 3a_6 = \alpha_2$$

$$a_3^3 - 3a_2a_3a_4 + 3a_1a_4^2 - 3a_1a_3a_5 - 3a_4a_5 + 3a_2^2a_5 + 6a_3a_6 - 3a_1a_2a_6 = \alpha_3$$

$$a_4^3 - 3a_3a_4a_5 + 3a_2a_5^2 - 3a_1a_5a_6 - 3a_2a_4a_6 + 3a_6^2 + 3a_3^2a_6 = \alpha_4$$

$$a_5^3 + 3a_3a_6^2 - 3a_4a_5a_6 = \alpha_5$$

$$a_6^3 = \alpha_6 \quad .$$

In Chapter 6 the equations relating the characteristic coefficients of X and P in the equation $X^2 = P$ were obtained by using the companion form of X.

Similarly the relations between the characteristic coefficients of X and P in the equation $X^3 = P$ may be obtained by using the companion form of X and evaluating $\det[C_x^3 - \lambda I]$ which can be reduced to the evaluation of a 3×3 determinant by row and column operations.

The process is illustrated in the 5×5 case as follows.

Let the characteristic polynomials of X and P be

$$\sum_{i=0}^{5} a_i\lambda^{5-i}, \qquad \sum_{i=0}^{5} \alpha_i\lambda^{5-i}, \qquad a_o = \alpha_o = 1 \ .$$

Let the companion form of X be C_x.

Then $C_x = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ -a_5 & -a_4 & -a_3 & -a_2 & -a_1 \end{pmatrix}$

$$C_x^3 = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ -a_5 & -a_4 & -a_3 & -a_2 & -a_1 \\ a_1a_5 & (a_1a_4-a_5) & (a_1a_3-a_4) & (a_1a_2-a_3) & (a_1^2-a_2) \\ (a_2a_5-a_1^2a_5) & (a_1a_5+a_2a_4-a_1^2a_4) & (a_1a_4-a_5+a_2a_3-a_1^2a_3) & (a_1a_3-a_4+a_2^2-a_1^2a_2) & 2a_1a_2-a_3-a_1^3 \end{pmatrix}$$

$\det[C_x^3 - \lambda I]$ after row operations simplifies to

$$\det \begin{pmatrix} -\lambda & 0 & 0 & 1 & 0 \\ 0 & -\lambda & 0 & 0 & 1 \\ -a_5 & -a_4 & -a_3-\lambda & -a_2 & -a_1 \\ 0 & -a_5 & -a_4-a_1\lambda & -a_3-\lambda & -a_2 \\ 0 & 0 & -a_5-a_2\lambda & -a_4-a_1\lambda & -a_3-\lambda \end{pmatrix}$$

$$= \det \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & -a_3-\lambda & -a_5-a_2\lambda & -a_4-a_1\lambda \\ 0 & 0 & -a_4-a_1\lambda & (-a_3-\lambda)\lambda & -a_5-a_2\lambda \\ 0 & 0 & -a_5-a_2\lambda & (-a_4-a_1\lambda)\lambda & (-a_3-\lambda)\lambda \end{pmatrix}$$

$$= [- a_3 - \lambda]^3\lambda^2 + [- a_4 - a_1\lambda]^3 + [- a_5 - a_2\lambda]^3$$

$$- 3[- a_3 - \lambda][- a_4 - a_1\lambda][- a_5 - a_2\lambda]\lambda \quad .$$

Setting $\det[C_x^3 - \lambda I] = 0$

then

$$- [a_3 + \lambda]^3\lambda^2 - [a_4 + a_1\lambda]^3 - [a_5 + a_2\lambda]^3 + 3[a_3 + \lambda][a_4 + a_1\lambda]$$

$$[a_5 + a_2\lambda]\lambda = 0$$

$$\Rightarrow [a_3 + \lambda]^3\lambda^2 + [a_4 + a_1\lambda]^3\lambda + [a_5 + a_2\lambda]^3 - 3[a_3 + \lambda][a_4 + a_1\lambda]$$

$$[a_5 + a_2\lambda]\lambda = 0 \quad .$$

This is a polynomial in λ of degree 5 and since $\det[C_x^3 - \lambda I] = \det[P - \lambda I]$ then the coefficients may be equated to the characteristic coefficients of P.

$$\lambda^5 + (a_1^3 - 3a_1a_2 + 3a_3)\lambda^4 + (a_2^3 - 3a_1a_2a_3 + 3a_3^2 - 3a_2a_4 + 3a_1^2a_4 - 3a_1a_5)\lambda^3$$

$$+ (a_3^3 - 3a_2a_3a_4 + 3a_1a_4^2 - 3a_1a_3a_5 - 3a_4a_5 + 3a_2^2a_5)\lambda^2$$

$$+ (a_4^3 - 3a_3a_4a_5 + 3a_2a_5^2)\lambda + a_5^3 = 0 \quad .$$

This gives the same set of equations connecting the characteristic coefficients of X and P as those obtained by the elimination method.

Some difficulty has been encountered in attempting to find

an explicit formula connecting the characteristic coefficients of X and P, but the set of equations may be formed for any given m, by formation of the polynomials A_i, $i = 0,1,2$ and computation of the polynomial in P of degree m from equation (5).

II An attempt to find an explicit formula connecting the characteristic coefficients of X and P in the equation $X^3 = P$.

It has been shown in part I that if P is non derogatory its characteristic coefficients may be equated with the coefficients of P in the equation

$$A_o^3 P^2 + A_1^3 P + A_2^3 - 3A_o A_1 A_2 P = 0$$

where

$$A_{2-r} = \sum_{i=0}^{[\frac{m-r}{3}]} a_{m-r-3i} P^i, \qquad r = 0,1,2 \quad .$$

Consider the case where m is a multiple of 3.

Then

$$A_o = \left\{ a_1 P^{(\frac{m-3}{3})} + a_4 P^{(\frac{m-3}{3})-1} + a_1 P^{(\frac{m-3}{3})-2} + \ldots + a_{m-5} P + a_{m-2} I \right\}$$

$$A_1 = \left\{ a_2 P^{(\frac{m-3}{3})} + a_5 P^{(\frac{m-3}{3})-1} + a_8 P^{(\frac{m-3}{3})-2} + \ldots + a_{m-4} P + a_{m-1} I \right\}$$

$$A_2 = \left\{ a_o P^{\frac{m}{3}} + a_3 P^{\frac{m}{3}-1} + a_6 P^{\frac{m}{3}-2} + \ldots + a_{m-3} P + a_m I \right\} \quad .$$

The formula for the cube of a polynomial of degree n is

$$[a_o P^n + a_1 P^{n-1} + \ldots + a_n I]^3 = \sum_{s=0}^{3n} \sum_{i+j+k=s} a_i a_j a_k P^{3n-s} \quad ,$$

$$i,j,k = 0,1,2,\ldots,n \quad .$$

From this, it is clear that

$$[a_o P^n + a_3 P^{n-1} + a_6 P^{n-1} + \ldots a_{3n} I]^3 = \sum_{s=0}^{3n} \sum_{i+j+k=3s} a_i a_j a_k P^{3n-s} ,$$

$$i,j,k = 0,3,6.,,,3n$$

and further

$$[a_r P^n + a_{r+3} P^{n-1} + a_{r+6} P^{n-2} + \ldots a_{r+3n} I]^3 = \sum_{s=0}^{3n} \sum_{i+j+k=3r+3s} a_i a_j a_k P^{3n-s} ,$$

$$i,j,k = r,r+3,r+6,\ldots,r+3n .$$

Using the last formula it is possible to obtain expressions for A_o^3, A_1^3, A_2^3 .

For A_o^3, setting $r = 1$ and $n = \frac{m-3}{3}$.

Then $$A_o^3 = \sum_{s=0}^{m-3} \sum_{i+j+k=3+3s} a_i a_j a_k P^{(m-3)-s}$$

$$= \sum_{s=1}^{m-2} \sum_{i+j+k=3s} a_i a_j a_k P^{(m-2)-s} \qquad i,j,k = 1,4,7,\ldots,m-2 .$$

For A_1^3, setting $r = 2$ and $n = \frac{m-3}{3}$

$$A_1^3 = \sum_{s=0}^{m-3} \sum_{i+j+k=6+3s} a_i a_j a_k P^{(m-3)-s}$$

$$= \sum_{s=2}^{m-1} \sum_{i+j+k=3s} a_i a_j a_k P^{(m-1)-s} \qquad i,j,k = 2,5,8,\ldots,m-1 .$$

For A_2^3, setting $r = 0$ and $n = \frac{m}{3}$

$$A_2^3 = \sum_{s=0}^{m} \sum_{i+j+k=3s} a_i a_j a_k P^{m-s} \qquad i,j,k = 0,3,6,\ldots,m .$$

Since A_o does not involve a_o, a_{m-1} or a_m

and A_1 does not involve a_o, a_1 or a_m

The two summations for A_o^3 and A_1^3 may be set from s = 0 to m.

$$\therefore \quad A_o^3 P^2 = \sum_{s=0}^{m} \sum_{i+j+k=3s} a_i a_j a_k P^{m-s} \qquad i,j,k = 1,4,7,\ldots,m-2$$

$$A_1^3 P = \sum_{s=0}^{m} \sum_{i+j+k=3s} a_i a_j a_k P^{m-s} \qquad i,j,k = 2,5,8,\ldots,m-1$$

$$A_2^3 = \sum_{s=0}^{m} \sum_{i+j+k=3s} a_i a_j a_k P^{m-s} \qquad i,j,k = 0,3,6,\ldots,m \ .$$

Though each term can be expressed as $\sum_{s=0}^{m} \sum_{i+j+k=3s} a_i a_j a_k P^{m-s}$

the choice of i,j,k is different for each term.

An expression is now required for $-3A_o A_1 A_2 P$

$$A_o A_1 = \left\{a_1 P^{\frac{m-3}{3}} + a_4 P^{(\frac{m-3}{3})-1} + a_7 P^{(\frac{m-3}{3})-2} + \ldots a_{m-2} I\right\}$$

$$\left\{a_2 P^{\frac{m-3}{3}} + a_5 P^{(\frac{m-3}{3})-1} + \ldots a_{m-1} I\right\}$$

$$= a_1 a_2 P^{\frac{2}{3}m-2} + [a_4 a_2 + a_1 a_5] P^{\frac{2}{3}m-3} + [a_7 a_2 + a_4 a_5 + a_1 a_8] P^{\frac{2}{3}m-4}$$

$$+ [a_{10} a_2 + a_7 a_5 + a_4 a_8 + a_1 a_{11}] P^{\frac{2}{3}m-5} + \ldots.$$

$$+ \sum_{i+j=3(s-1)} a_i a_j P^{\frac{2}{3}m-s} + \ldots + a_{m-2} a_{m-1} I$$

$$\therefore \quad A_o A_1 = \sum_{s=2}^{\frac{2}{3}m} \sum_{i+j=3(s-1)} a_i a_j P^{\frac{2}{3}m-s} \qquad i = 1,4,7\ldots\ldots\ m-2$$

$$j = 2,5,8\ldots\ldots\ m-1$$

$$A_2 P = \left\{ a_o P^{\frac{m}{3}+1} + a_3 P^{\frac{m}{3}} + a_6 P^{\frac{m}{3}-1} + \ldots\ a_{m-3} P^2 + a_m P \right\}$$

$$= \sum_{s=0}^{\frac{m}{3}} a_{3s} P^{\frac{m}{3}+1-s}$$

$$\therefore \quad A_o A_1 A_2 P = a_1 a_2 a_o P^{m-1}$$

$$+ [a_4 a_2 a_o + a_1 a_5 a_o + a_1 a_2 a_3] P^{m-2}$$

$$+ [a_7 a_2 a_o + a_4 a_5 a_o + a_1 a_8 a_o + a_4 a_2 a_3 + a_1 a_5 a_3 + a_1 a_2 a_6] P^{m-3}$$

$$+ [a_{10} a_2 a_o + a_7 a_5 a_o + a_4 a_8 a_o + a_1 a_{11} a_o + a_7 a_2 a_3$$

$$+ a_4 a_5 a_3 + a_1 a_8 a_3 + a_4 a_2 a_6 + a_1 a_5 a_6 + a_1 a_2 a_9] P^{m-4}$$

$$+ \ldots\ldots$$

$$+ \sum_{i+j+k=3s} a_i a_j a_k P^{m-s}$$

$$+ \ldots\ldots$$

$$+ a_{m-2}\ a_{m-1}\ a_m I$$

$$\therefore \quad A_o A_1 A_2 P = \sum_{s=1}^{m} \sum_{i+j+k=3s} a_i a_j a_k P^{m-s} \qquad i = 1,4,7\ldots\ldots m-2$$

$$j = 2,5,8\ldots\ldots m-1$$

$$k = 0,3,6\ldots\ldots m\ .$$

The coefficient of P^{m-s} in the expression

$$A_o^3 P^2 + A_1^3 P + A_2^3 - 3A_o A_1 A_2 P$$

can therefore be obtained from

$$\sum_{i+j+k=3s} a_i a_j a_k \qquad i,j,k = 1,4,7 \ldots\ldots m-2$$

$$+ \sum_{i+j+k=3s} a_i a_j a_k \qquad i,j,k = 2,5,8 \ldots\ldots m-1$$

$$+ \sum_{i+j+k=3s} a_i a_j a_k \qquad i,j,k = 0,3,6 \ldots\ldots m$$

$$- 3 \sum_{i+j+k=3s} a_i a_j a_k \qquad i = 1,4,7 \ldots\ldots m-2$$

$$j = 2,5,8 \ldots\ldots m-1$$

$$k = 0,3,6 \ldots\ldots m \ .$$

III The equation $X^4 = P$.

Let the characteristic polynomials of X and P be

$\sum_{i=0}^{m} a_i \lambda^{m-i}$ and $\sum_{i=0}^{m} \alpha_i \lambda^{m-i}$ respectively where $a_o = \alpha_o = 1$.

Then X satisfies the equation $X^m + a_1 X^{m-1} + a_2 X^{m-2} + \ldots a_{m-1} X + a_m I = 0$

Since X also satisfies the equation $X^4 = P$, then substituting P for X^4 gives a cubic equation in X, whose coefficients are scalar polynomials in P.

This is

$$A_o X^3 + A_1 X^2 + A_2 X + A_3 = 0$$

where $A_{3-r} = \sum_{i=0}^{[\frac{m-r}{4}]} a_{m-r-4i} P^i \qquad r = 0,1,2,3$

for m = 2, define $A_i = a_i I \quad i = 0,1,2 \quad A_3 = 0$.

A second cubic equation may be obtained by multiplying this equation on the right by X and substituting P for X^4 again. Hence for m×m matrices, X satisfies the two equations

$$A_o X^3 + A_1 X^2 + A_2 X + A_3 = 0$$

$$A_1 X^3 + A_2 X^2 + A_3 X + A_o P = 0 .$$

As in the equation $X^3 = P$, X may be eliminated completely since the coefficients commute with each other. The computational problems are greatly increased however.

Eliminating X completely, the equation obtained is

$$[J_2K_1 - J_1K_2][J_2K_2L_1 - J_1K_2L_2] - J_2[J_2L_1 - J_1L_2]^2$$

$$- [J_2K_1 - J_1K_2]^2 L_2 = 0$$

where $J_1 = A_1^2 - A_oA_2$, $K_1 = A_1A_2 - A_oA_3$, $L_1 = A_1A_3 - A_o^2P$

$J_2 = A_1J_1 - A_oK_1$, $K_2 = A_2J_1 - A_oL_1$, $L_2 = A_3J_1$.

Clearly, an attempt to write this equation in terms of A_i, i = 0,...,3 would be very difficult.

IV The equation $X^n = P$.

Let the characteristic polynomials of X and P be

$\sum_{i=0}^{m} a_i\lambda^{m-i}$ and $\sum_{i=0}^{m} \alpha_i\lambda^{m-i}$ respectively where $a_o = \alpha_o = 1$.

Then X satisfies the equation $X^m + a_1X^{m-1} + a_2X^{m-2} + \ldots a_{m-1}X + a_mI = 0$.

Since X also satisfies the equation $X^n = P$, then substituting P for X^n gives an equation in X of degree (n-1)

$$A_oX^{n-1} + A_1X^{n-2} + A_2X^{n-3} + \ldots A_{n-1} = 0 \qquad (1)$$

where $$A_{(n-1)-r} = \sum_{i=0}^{[\frac{m-r}{n}]} a_{m-r-n_i} P^i \qquad i = 0,1,2,\ldots n-1$$

for $m \geq n-1$.

If $m < n-1$ then $A_i = a_iI$ for $i = 0,1,\ldots,m$

and $A_i = 0$ for $m < i \leq n-1$.

A second equation of degree (n-1) may be obtained by multiplying equation (1) on the right by X and substituting P for X^n again. Then X satisfies the two equations

$$A_oX^{n-1} + A_1X^{n-2} + A_2X^{n-3} + \ldots + A_n = 0 \qquad (1)$$

$$A_1X^{n-1} + A_2X^{n-2} + A_3X^{n-3} + \ldots + A_oP = 0 . \qquad (2)$$

Since the coefficient matrices are all polynomials in P and commute with each other, elimination methods may be applied to eliminate X completely. The computational problems however, are very great.

Appendix C

Section α.

In this thesis, various methods of solution of the unilateral matrix equation $A_0X^n + A_1X^{n-1} + \ldots + A_n = 0$ have been discussed. Emphasis has mainly been placed on the case where the equation is monic, i.e. $A_0 = I$. In cases where A_0 is nonsingular, the equation can obviously be made monic by multiplication on the left by A_0^{-1}.

In section 3.3, Method I could only be applied if the equation were first made monic and hence would not be useful in the case of singular A_0 [see however, section β].

Methods such as those described in section 3.2, Methods I and II could still be applied with singular A_0. These methods depend upon finding solutions of the scalar equation $g(\lambda) = 0$ where $g(\lambda) = \det[A_0\lambda^n + A_1\lambda^{n-1} + \ldots + A_n]$. Since the leading coefficient of this determinant is $\det[A_0]$, then if A_0 is singular, the polynomial will be of degree less than mn and hence the number of roots to be found will be less than in the case of nonsingular A_0.

The method described in section 4.2 also depends upon examination of $\det[A_0\lambda^n + A_1\lambda^{n-1} + \ldots + A_n]$ to find possible characteristic polynomials for X. The choice of factors will be reduced since the polynomial $g(\lambda)$ will be of degree less than mn.

Other problems arise in this method if A_0 is singular. If the characteristic polynomial of X is $\sum_{i=0}^{m} a_i\lambda^{m-i}$, $a_0 = 1$ then X satisfies the two equations

$$A_o X^2 + A_1 X + A_2 = 0 \qquad (1)$$

$$X^m + a_1 X^{m-1} + a_2 X^{m-2} + \ldots + a_m I = 0 \quad . \qquad (2)$$

Multiplying equation (2) on the left by A_o and equation (1) on the right by X^{m-2} and subtracting gives equation (3)

$$[a_1 A_o - A_1] X^{m-1} + [a_2 A_o - A_2] X^{m-2} + a_3 A_o X^{m-3} + \ldots + a_m A_o = 0. \qquad (3)$$

The elimination cannot be continued since A_o and A_1 do not in general commute.

It would be necessary to make equation (3) monic by multiplying on the left by $[a_1 A_o - A_1]^{-1}$. If both A_o and $[a_1 A_o - A_1]$ are singular then the process may break down.

For nonsingular $[a_1 A_o - A_1]$ then equation (3) becomes

$$X^{m-1} + B_1 X^{m-2} + B_2 X^{m-3} + \ldots + B_{m-1} = 0 \qquad (4)$$

where $B_1 = [a_1 A_o - A_1]^{-1} [a_2 A_o - A_2]$

and $B_i = [a_1 A_o - A_1]^{-1} [a_{i+1} A_o]$ for $i = 2,3,\ldots,m-1$.

The process could be continued by combining equations (4) and (1) to obtain an equation in X of degree (m-2) which would again have to be made monic to continue the process.

It can be seen that if A_o is singular, the method of section 4.2 can be applied, but it involves the calculation of an inverse matrix at each stage, and if the leading coefficient of the equation obtained becomes singular then the method may break down.

Section β.

Dennis, Traub and Weber [1976] consider solutions of the matrix equation $A_o X^n + A_1 X^{n-1} + \ldots + A_n = 0$. (1)

It is stated by them that if A_o is singular, 'one can shift co-ordinates and reverse the order of the coefficients to get a related problem with a nonsingular leading coefficient'.

e.g. Multiplying equation (1) on the right by X^{-n} gives

$$A_o + A_1 X^{-1} + A_2 X^{-2} + \ldots \; A_{n-1} X^{1-n} + A_n X^{-n} = 0$$

or $$A_n Y^n + A_{n-1} Y^{n-1} + \ldots \; A_1 Y + A_o = 0 \qquad (2)$$

where $Y = X^{-1}$.

Hence if A_o is singular but A_n is nonsingular equation (2) can be solved instead and the solution X for equation (1) obtained from $X = Y^{-1}$.

This method could be applied for singular A_o in solving equation (1), for n = 2, by Method I of section 3.3, since equation (2) could then be made monic provided A_2 was nonsingular.

There are problems which are not mentioned by Dennis, Traub and Weber. Reversing the coefficients would not be an advantage if both A_o and A_n were singular.

For example, if $A_o X^2 + A_1 X + A_2 = 0$

where $$A_o = \begin{pmatrix} 1 & -1 \\ 1 & -1 \end{pmatrix} \quad A_1 = \begin{pmatrix} -2 & 1 \\ 0 & 1 \end{pmatrix} \quad A_2 = \begin{pmatrix} 2 & 4 \\ 1 & 2 \end{pmatrix} .$$

Though A_o and A_2 are singular, solutions may still be obtained by use of latent roots and vectors as described in section 3.2.

$\text{Det}[A_o\lambda^2 + A_1\lambda + A_2] = \lambda(\lambda-3)(2\lambda+1)$ and the latent roots are 0, 3, $-\frac{1}{2}$ with latent vectors $\begin{pmatrix}-2\\1\end{pmatrix}$ $\begin{pmatrix}2\\5\end{pmatrix}$ and $\begin{pmatrix}1\\-1\end{pmatrix}$ respectively.

The solutions obtained are $X = \begin{pmatrix}\frac{1}{2} & 1\\ \frac{5}{4} & \frac{5}{2}\end{pmatrix}$ $X = \begin{pmatrix}\frac{1}{2} & 1\\ \frac{5}{2} & 2\end{pmatrix}$

and $X = \begin{pmatrix}\frac{1}{2} & 1\\ -\frac{1}{2} & -1\end{pmatrix}$.

These solutions could not be obtained by the iterative method described by Dennis, Traub and Weber in section 3.2 since the equation could not be made monic even after the order of the coefficient matrices have been reversed.

This equation could not be solved either by the method described by Roth [p.64] since this requires the equation to be made monic.

Another problem involved in reversing the order of the coefficients is that the multiplication of equation (1) by X^{-n} assumes that there is at least one nonsingular solution and so the method could not lead to a solution X which was singular. However, the iterative method in section 3.2 is designed to lead to the dominant solvent which cannot be singular since all the eigenvalues of a dominant solvent are greater in modulus than those of any other solvent and hence are all greater than zero.

In fact, if A_o is singular and there is a singular solvent X, then A_n would also be singular since

$$\det[A_oX^n + A_1X^{n-1} + \ldots + A_{n-1}X] = \det[-A_n]$$

$$\det[A_o X^{n-1} + A_1 X^{n-2} + \ldots + A_{n-1}][\det X] = -\det[A_n]$$

and if there is a solution X such that det[X] = 0 then $\det[A_n]$ must also be zero, and so it would not be possible to obtain a nonsingular leading coefficient by reversing the order of the coefficients.

Appendix D

The problems of singular matrices can affect many of the methods of solution described in this thesis. The case of the singular leading coefficient has been discussed in Appendix C, but singular matrices can also arise at some stage in applying a particular method, or an iterative method may break down because of the singularity of a matrix.

The particular problems which arise when J_{m-1} is singular in the method described in Chapters 4 and 6 have been discussed [pages 101 to 108]. Problems encountered in other methods due to singular matrices or zero denominators are now described.

Section 2.5. Direct solution of matrix equations.

In the case of 2×2 matrices, the equations $X^2 = A$, $XAX = B$, $X^2 + X - A = 0$ were shown, subject to certain conditions, to have 4 solutions, since the element x_2 could always be found as the root of an equation which was a quadratic in $(x_2)^2$. If the conditions are not fulfilled however, the number of solutions is reduced.

If $a_2 \neq 0$, a quadratic in $(x_2)^2$ is obtained in the solution of $X^2 = A$, on page 28

$$\left\{4|A| - (TrA)^2\right\} x_2^4 + 2a_2^2 (TrA)x_2^2 - a_2^4 = 0 \qquad (1)$$

and hence $$x_2^2 = \frac{a_2^2}{(TrA) \pm 2\sqrt{|A|}}$$ provided that $(TrA) \pm 2\sqrt{|A|} \neq 0$.

If the denominator is zero, however, then $4|A| - (TrA)^2 = 0$ and hence equation (1) becomes

$$2a_2^2(\mathrm{TrA})\, x_2^2 - a_2^4 = 0$$

and there are therefore only 2 possible values for x_2 and hence only 2 possible matrix solutions.

On page 33 a quadratic in $(x_2)^2$ is obtained in the solution of XAX = B, provided that $b_2 \neq 0$.

This is

$$[P^2 - 4|A||B|]\, x_2^4 - 2b_2^2\, Px_2^2 + b_2^4 = 0 \qquad (2)$$

where P = Trace (AB)

and hence $$x_2^2 = \frac{b_2^2}{P \pm 2\sqrt{|A||B|}}$$ provided $P \pm 2\sqrt{|A||B|} \neq 0$.

If the denominator is zero then $P^2 - 4|A||B| = 0$ and equation (2) becomes

$$-2b_2^2\, Px_2^2 + b_2^4 = 0$$

and therefore there are only 2 possible values for x_2.

The quadratic in $(x_2)^2$ obtained in the solution of $X^2 + X - A = 0$ on page 34 is

$$[(\mathrm{TrA})^2 - 4|A|]x_2^4 + a_2^2[-1 - 2(\mathrm{TrA})]x_2^2 + a_2^4 = 0 \qquad (3)$$

provided that $a_2 \neq 0$

and hence $$x_2^2 = \frac{a_2^2}{2}\left\{\frac{(1 + 2(\mathrm{TrA}) \pm \sqrt{1 + 4(\mathrm{TrA}) + 16|A|}}{(\mathrm{TrA})^2 - 4|A|}\right\}$$

provided that $(\mathrm{TrA})^2 - 4|A| \neq 0$.

If the denominator is zero then equation (3) becomes

$$a_2^2[-1 - 2(\mathrm{TrA})]\, x_2^2 + a_2^4 = 0$$

and therefore there are only 2 possible values for x_2 and hence only 2 matrix solutions.

Section 3.2. A survey of some methods of solution of the unilateral Matrix equation.

Method I breaks down if the matrix T obtained by solving $A_oTD^n + A_1TD^{n-1} + \ldots A_nT = 0$ is singular since no matrix of the form $X = TDT^{-1}$ can then be found. This can occur if the λ_i on the diagonal of the matrix D are not in fact the eigenvalues of a solution matrix X.

Section 3.3. The Quadratic Matrix Equation.

Several methods in this section express the solution X in a form which involves the inverse of a matrix. All of them can fail when this matrix is singular.

Method I forms the matrix $R = \begin{bmatrix} 0 & I \\ -A_2 & A_1 \end{bmatrix}$ and evaluates f(R) where $f(\lambda)$ is a factor of $\det[R - \lambda I]$.

If $f(R) = \begin{pmatrix} U & M \\ V & N \end{pmatrix}$ then either $X = -NM^{-1}$ or $X = -VU^{-1}$.

Clearly this method may fail if both M and U are singular.

Similarly Method 2 may break down if both H_{12} and H_{22} are singular since the solution X is expressed as either $X = H_{12}^{-1} H_{11}$ or $X = H_{22}^{-1} H_{21}$ [page 69].

Method 4, using the Schur canonical form of a matrix expresses the solution X as $X = U_{21} U_{11}^{-1}$ and may fail if U_{11} is singular.

Section 5.2. The method of Simple Iteration.

This method involves the rearrangement of a matrix equation $f(X) = 0$ in the form $X = g(X)$ and forming the iterative process $X_{i+1} = g(X_i)$.

Clearly any rearrangement which involves the inverse of a coefficient matrix cannot be applied directly if that coefficient is singular.

e.g. the iterative process $X_{i+1} = - A_1^{-1}[A_2 + X^2]$ formed from $X^2 + A_1X + A_2 = 0$ cannot be applied if A_1 is singular.

The method of simple iteration also may fail at some stage if the rearrangement involves the inverse of X.

e.g. $X_{i+1} = - A_2X_i^{-1} - A_1$ may fail if one of the iterates X_i becomes singular.

When these problems occur it is possible that a different rearrangement may lead to a solution.

Section 5.3. Bernoulli's Algorithm.

This is an iterative process in which the sequence $X_i\ X_{i-1}^{-1}$ converges to the dominant solvent. Though a dominant solvent is nonsingular, the method could fail at some stage if X_i became singular, which could happen if no dominant solvent existed.

Section 5.5. An iterative method for the solution of matrix equations using the characteristic matrix of a solution.

This method uses an iterative process formed from the elimination method described in Chapter 4

$$X^{(i+1)} = - J_{m-1}^{-1} K_{m-1} \text{ where } J_1 = a_1^{(i)} I - A_1, \qquad K_1 = a_2^{(i)} I - A_2$$

$$\text{and} \quad J_{r+1} = K_r - J_r A_1$$

$$K_{r+1} = a_{r+2}^{(i)} I - J_r A_2 \quad .$$

This method may break down at some stage if the matrix J_{m-1} becomes singular. This could happen if the choice of initial values for $a_1 a_2 \ldots a_m$ did not produce a convergent sequence.

This appendix shows that problems of singularity can affect many methods of solution of matrix equations. The singularities may arise from the given coefficient matrices or from the method itself. Iterative processes may break down due to singularities when conditions for convergence are not satisfied.