STOCHASTICA Vol. X, Nº 3, 1986.

APPROXIMATE SOLUTIONS OF MATRIX DIFFERENTIAL EQUATIONS

L. Jódar, A. Hervás, D. Garcia-Sala

ABSTRACT

A method for solving second order matrix differential equations avoiding the increase of the dimension of the problem is presented. Explicit approximate solutions and an error bound of them in terms of data are given.

Introduction.

Second order matrix differential equations are important in the theory of damped oscillatory systems and vibrational systems, [5],[8]. The standard method for solving the Cauchy problem

$$X^{(2)}+A_1X^{(1)}+A_oX=0;\; X^{(1)}(0)=C_1,\; X(0)=C_o,\; -\infty<t<+\infty \qquad (1)$$

where A_i and C_i, for $i=o,1$, are nxn complex matrices, is based on the application of the change $X=Y_1$, $X^{(1)}=Y_2$ and the transformation of problem (1) into the following extended linear system

$$Y^{(1)}=C_LY;\; Y(0)=\begin{bmatrix} C_o \\ C_1 \end{bmatrix} \;;\; Y=\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \;,\; C_L=\begin{bmatrix} 0 & I \\ -A_o & -A_1 \end{bmatrix} \qquad (2)$$

Thus the solution of (1) is given by the first entry of the function

$$Y(t)=\exp(tC_L)\begin{bmatrix} C_o \\ C_1 \end{bmatrix} \qquad (3)$$

Although there are interesting methods for computing the exponential function of a matrix, [10], [11], the above method for solving (1) has the numerical inconvinient of the increase of the original dimension of the problem and the lack of an estimation of the approximation error in terms of data problem, when one computes the expression (3).

In a recent paper [7], a method for solving the algebraic matrix equation

$$X^2 + A_1X + A_o = 0 \qquad (4)$$

is presented. In this paper and in an analogous way to the scalar case, we provide an explicit expression of the solution of (1) in terms of a pair of solutions of the algebraic matrix equation (4), avoiding the increase of the original dimension of the problem. We apply the result in order to obtain an iterative algorithm for solving problem (1) by constructing approximate solutions of this problem in terms of approximate solutions of equation (4), when coefficients of equation (4) satisfy certain conditions. Also, an estimation of the approximation error in terms of data problem is given.

The resolution problem of equation (4) is related to the problem of the existence of a linear factorization of the matrix polynomial $L(z)=z^2I+A_1z+A$. So, if the companion matrix C_L is diagonable, then $L(z)$ admits a linear factorization of the type $L(z)=(zI-V_1)(zI-V_2)$, for certain matrices V_1 and V_2, and V_2 is a solution of (4). In this case an explicit expression of solution of (4) is given in [2] .

Nevertheless, it is interesting to remark that the standard method is always available, but equation (4) may be unsolvable. For instance, if A_o and A_1 are nxn complex matrices such that $A_oA_1=A_1A_o$ and the minimal polynomial of the matrix $A_1^2-4A_o$ has a double root at the origin, then (4) is unsolvable, [9].

We recall some concepts and properties of matrix norms that will be used below and whose proofs may be found in [6]. If A is an nxn complex matrix, we denote by $\|A\|$ the operator norm of A, defined by

$$\|A\| = \sup_{x\neq o} \frac{\|Ax\|_2}{\|x\|_2}$$

where $\| \ \|_2$ represents the usual euclidean norm. If A,B are nxn complex matrices it follows that

$$\|AB\| \leqslant \|A\| \, \|B\| \tag{5}$$

$$n^{-\frac{1}{2}} \max_j \{ \sum_{i=1}^{n} |a_{ij}| \} \leqslant \|A\| \leqslant n^{\frac{1}{2}} \max_j \{ \sum_{j=1}^{n} |a_{ij}| \} \tag{6}$$

$$n^{-\frac{1}{2}} \max_i \{ \sum_{j=1}^{n} |a_{ij}| \} \leqslant \|A\| \leqslant n^{\frac{1}{2}} \max_i \{ \sum_{j=1}^{n} |a_{ij}| \} \tag{7}$$

Expressions (6)-(7) allow us an available estimation of the norm $\|A\|$ because the expression of $\|A\|$ is not interesting from a computational point of view. If A and A+E are nonsingular nxn complex matrices, then from the Banach lemma, [6], p.28, one gets

$$\|(A+E)^{-1}-A^{-1}\| \leqslant \|E\| \|A^{-1}\| \, \|(A+E)^{-1}\| \tag{8}$$

Finally from corollary (50.5) of [1], if A is a nonsingular matrix and B is a matrix such that $\|B-A\| < \|A^{-1}\|^{-1}$, then B is also nonsingular. Hence the set of all nonsingular matrices in C^{nxn} is an open set of C^{nxn} endowed with the $\| \ \|$-topology.

A numerical method for solving matrix differential equations

We begin this section with a result that provides an explicit expression of the solution of problem (1) in terms of a pair of solutions of equation (4) satisfying certain additional condition.

Theorem 1. Let X_o, X_1 be two solutions of equation (4) such that $X_1 - X_o$ is nonsingular in $C^{n\times n}$, then the only solution of (1) is given by the expression

$$X(t)=\exp(tX_o)P + \exp(tX_1)Q \qquad (9)$$

where

$$P=I+(X_1-X_o)^{-1}(C_o-C_1); \qquad Q=(X_1-X_o)^{-1}(C_1-X_oC_o) \qquad (10)$$

Proof. Considering the standard method for solving (1) it is clear from the uniqueness property of solutions of a Cauchy problem, that there exists only one solution of (1). For any matrices P and Q in $C^{n\times n}$, it is clear that the matrix function $Y(t)=\exp(tX_o)P+\exp(tX_1)Q$, satisfies the matrix differential equation arising in (1). If we impose that Y(t) satisfies the conditions $Y^{(1)}(0)=C_1$, $Y(0)=C_o$, it follows that matrices P and Q must verify the following matrix system

$$Y(0)=P + Q = C_o$$

$$Y^{(1)}(0)=X_oP + X_1Q = C_1$$

or

$$\begin{bmatrix} I & I \\ X_o & X_1 \end{bmatrix} \begin{bmatrix} P \\ Q \end{bmatrix} = \begin{bmatrix} C_o \\ C_1 \end{bmatrix} \qquad (11)$$

From the invertibility of X_1-X_o, it is easy to show that the coefficient matrix S arising in (11) is nonsingular and

$$S^{-1}=\begin{bmatrix} I & I \\ X_o & X_1 \end{bmatrix}^{-1} = \begin{bmatrix} I+(X_1-X_o)^{-1} & -(X_1-X_o)^{-1} \\ -(X_1-X_o)^{-1}X_o & (X_1-X_o)^{-1} \end{bmatrix}$$

Solving (11) it follows that $\begin{bmatrix} P \\ Q \end{bmatrix} = S^{-1}\begin{bmatrix} C_o \\ C_1 \end{bmatrix}$ are given by (10). As by construction Y(t) satisfies the Cauchy conditions arising in (1), one concludes that the expression (9) defines the only solution of (1).

Corollary 1. Let us consider problem (1) and let us suppose that there exist a pair of solutions X_o, X_1 of (4) such that X_1-X_o is nonsingular, and let $\{Z_k\},\{W_k\}$ be sequences of matrices $\|\;\|$-convergent to X_o and X_1, respectively. Then for each real number t, the sequence of matrix functions defined by

$$\begin{gathered} X_n(t)=\exp(tZ_n)P_n+\exp(tW_n)Q_n \\ P_n=I+(W_n-Z_n)^{-1}(C_o-C_1); \quad Q_n=(W_n-Z_n)^{-1}(C_1-Z_nC_o),\ n\geq o \end{gathered} \tag{13}$$

is $\|\;\|$-convergent when $n\to\infty$, to the only solution X(t) of problem (1).

Proof. The result is a consequence of th. 1, the continuity of the exponential function and the property (5).

In order to obtain an effective computation of the solution of (1) we need an explicit definition of the sequences $\{Z_k\}$, $\{W_k\}$, for $k\geq o$, as well as an estimation of the approximation error for the approximate solutions in terms of data problem. Next result provides a complete information in this sense for a class of matrix differential equations of type (1).

Theorem 2. Let us consider problem (1) where A_1 is nonsingular and such that

$$d=(1-4\ \|A_1^{-1}\|\ \|A_1^{-1}A_o\|)^{1/2}>0 \tag{14}$$

Let $a=(1-d)/(2\|A_1^{-1}\|)$ and $b=(2\ \|A_1^{-1}\|)^{-1}$, then it follows that

(i) There exist a pair X_o, X_1 of solutions of equation (4) such that X_1-X_o is non-singular and $\|X_1-X_o\|>d(\|A_1^{-1}\|)^{-1}$.

(ii) Let $F:C^{nxn}\to C^{nxn}$, $G:C^{nxn}\to C^{nxn}$, be matrix functions defined by the expressions $F(Z)=-A_1^{-1}A_o-A_1^{-1}Z^2$, and $G(Z)=-A_1^{-1}A_o-Z^2A_1^{-1}$, then the sequences $\{Z_k\}$, $\{W_k\}$, defined by the recurrent expressions

$$Z_o=0,\ Z_{k+1}=F(Z_k),\ k\geq o\ ;\ V_o=0,\ V_{k+1}=G(V_k),\ W_k=-A_1-A_1V_kA_1^{-1} \tag{15}$$

satisfy $\|Z_k\|\leq b$, $\|V_k\|\leq b$, for all $k\geq o$. If X_o and X_1 are the matrices given in (i) one gets

$$X_o=\lim_{k\to\infty} Z_k\quad,\quad X_1=\lim_{k\to\infty} W_k$$

in the $\|\,\|$-topology, and $\|Z_k\|\leq b, \|V_k\|\leq b, \|X_o\|\leq a, \|X_1\|\leq\|A_1\|(1+a\|A_1^{-1}\|)$.

(iii) If $X_n(t)$ is given by (13) where W_n and Z_n are the matrices defined by (15), then this sequence of matrix functions converges when $n\to\infty$, for each fixed t, to the only solution $X(t)$ of problem (1). If $k(A_1)=\|A_1\|\|A_1^{-1}\|$, $w=\max\{a,\|A_1\|(1+ak(A_1))\}$,

$M(t)=1+t^2(1+2\|C_o-C_1\|\|A_1^{-1}\|/d+(1+k(A_1))(\|C_1\|+a\|C_o\|)+\|C_o\|\|A_1^{-1}\|/d+t^2\|A_1^{-1}\|k(A_1)(\|C_1\|+a\|$

and $e_n=(2a\|A_1^{-1}\|)^n\ \|A_1^{-1}\|/(1-2a\ \|A_1^{-1}\|)$, then for n advanced one has

$$E_n(t)=\|X(t)-X_n(t)\|\ \leq\ 2M(t)\exp(2wt)e_n \tag{16}$$

<u>Proof.</u> Premultiplying the equation (4) by A_1^{-1} one gets the equation

$$A_1^{-1}X^2+X+A_1^{-1}A_o\ =\ 0 \tag{17}$$

Now, let us consider the associated equation

$$X^2A_1^{-1}+X+A_1^{-1}A_o = 0 \tag{18}$$

and the scalar function $f:[0,b] \to [0,\infty[$, defined by the expression

$$f(s)=\|A_1^{-1}\|\ s^2+\|A_1^{-1}A_o\| \tag{19}$$

then $f(b)<b$ and $f^{(1)}(b)=1$. From th.2.1, and section 3 of [4], it follows that the matrix sequences $\{Z_k\}$ and $\{W_k\}$ defined by (15), converges to solutions X_o and U_1 of equations (17) and (18) respectively, with $\|X_o\| \leqslant a$, $\|U_1\| \leqslant a$. Note that the matrix X_o is a solution of (4). From section 3 of [4], considering the matrix $X_1=-A_1-A_1U_1A_1^{-1}$, is a solution of (4) such that X_1-X_o is nonsingular and

$$A_1^{-1}(X_1-X_o)=I-D;\ D=-U_1A_1^{-1}-A_1^{-1}X_o,\ \ \|X_1\|\leqslant\|A_1\|(1+a\|A_1^{-1}\|),\|D\|<1-d \tag{20}$$

From (20) it follows that $\|A_1^{-1}\|\|X_1-X_o\| \geqslant |1-\|D\||>d$. Hence (i) and (ii) are proved. (iii) From corollary 1, the matrix functions $X_n(t)$ defined by (13) where Z_n, V_n, are given by (15), converges when $n\to\infty$, and t is fixed, to the only solution $X(t)$ of problem (1). Considering the difference $X(t)-X_n(t)$ it follows that

$$X(t)-X_n(t)=\exp(tX_o)(P-P_n)+(\exp(tX_o)-\exp(tZ_n))P_n+\exp(tX_1)(Q-Q_n)+(\exp(tX_1)-\exp(tW_n))Q_n \tag{21}$$

From (10), (13) it follows that

$$P-P_n=((X_1-X_o)^{-1}-(W_n-Z_n)^{-1}) \tag{22}$$

$$Q-Q_n=((X_1-X_o)^{-1}-(W_n-Z_n)^{-1})(C_1-X_oC_o)+(W_n-Z_n)^{-1}(C_1-X_oC_o-C_1+Z_nC_o)=$$
$$=((X_1-X_o)^{-1}-(W_n-Z_n)^{-1})(C_1-X_oC_o)+(W_n-Z_n)^{-1}(Z_n-X_o)C_o \tag{23}$$

From the convergence of Z_n and W_n to X_o and X_1 respectively, and from (i), there exists an integer n_o such that for $n \geq n_o$ one gets $\| W_n - Z_n \| \geq \| X_1 - X_o \| /2 > (\| A_1^{-1} \|)^{-1} d/2$. Thus for $n \geq n_o$ we have

$$\| (W_n - Z_n)^{-1} \| \leq 2 (\| X_1 - X_o \|)^{-1} < 2 \| A_1^{-1} \| / d \qquad (24)$$

From (8), (24) and (i), for $n \geq n_o$ it follows that

$$\| (X_1 - X_o)^{-1} - (W_n - Z_n)^{-1} \| \leq \| (X_1 - X_o)^{-1} \| \, \| (W_n - Z_n)^{-1} \| \, \| (X_1 - W_n) + (X_o - Z_n) \| \leq \qquad (25)$$
$$\leq 2 (\| X_1 - W_n \| + \| X_o - Z_n \|)$$

Considering the function f given by (19) and from th.2.1-(4), of [4], one gets

$$\| X_o - Z_n \| \leq e_n = (2a \| A_1^{-1} \|)^n \| A_1^{-1} \| / (1 - 2a \| A_1^{-1} \|) \; ; \; \| U_1 - V_n \| \leq e_n \qquad (26)$$

In accordance with the notation of [6], p. 25, let $k_1 = k(A_1) = \| A_1 \| \| A_1^{-1} \|$, the condition number of A_1, and taking into account (15) and (26) one gets

$$\| X_1 - W_n \| = \| A_1 V_n A_1^{-1} - A_1 U_1 A_1^{-1} \| \leq e_n k(A_1) \qquad (27)$$

From (25)-(27) it follows that

$$\| (X_1 - X_o)^{-1} - (W_n - Z_n)^{-1} \| \leq 2 e_n (1 + k(A_1)), \; n \geq n_o \qquad (28)$$

where e_n is defined by (26).

From (23), (24), (27) and (28) it follows that

$$\| Q - Q_n \| \leq 2 e_n ((1 + k(A_1)) (\| C_1 \| + a \| C_o \|) + \| C_o \| \| A_1^{-1} \| / d), \text{ if } n \geq n_o \qquad (29)$$

As $\| X_o \| \leq a, \| X_1 \| \leq \| A_1 \| (1 + a k(A_1))$, from the mean value th., [3], p. 158, it follows that

$$\|\exp(tX_o)-\exp(tZ_n)\| \leq t^2\exp(2at)\|X_o-Z_n\| \leq t^2\exp(2at)e_n \tag{30}$$

$$\|\exp(tX_1)-\exp(tW_n)\| \leq t^2\exp(2\|A_1\|t(1+k(A_1)a))\|X_1-W_n\| \leq t^2k(A_1)\exp(2\|A_1\|t(1+k(A_1)a))e_n \tag{31}$$

From the expressions (13) and (24), for $n \geq n_o$ we have

$$\|P_n\| \leq 1+2\|C_o-C_1\| \ \|A_1^{-1}\|/d \ ; \ \|Q_n\| \leq 2\|A_1^{-1}\|(\|C_1\|+a\|C_o\|)/d \tag{32}$$

Let $w=\max\{a,\|A_1\|(1+ak(A_1))\}$, then from (21)-(32), it follows that

$$\|X(t)-X_n(t)\| \leq 2e_n\exp(2wt)\{1+t^2(1+2\|C_o-C_1\|\|A_1^{-1}\|/d)+(1+k(A_1))(\|C_1\|+a\|C_o\|)+\|C_o\|\|A_1^{-1}\|/d)\}$$

$$+2e_n\exp(2wt) \ t^2(\|C_1\|+a \ \|C_o\|)\|A_1^{-1}\|k(A_1)/d$$

Thus the result is proved.

References.

[1] BERBERIAN, S.K., Lectures on functional analysis and operator theory, Springer Verlag, New York (1974).

[2] DENNIS, JR., J.E., TRAUBB, J.F. and WEBER, R.P., "The algebraic theory of matrix polynomials", SIAM J. Numer. Anal. 6, 831-845 (1976).

[3] DIEUDONNÉ, J., Fundamentos de análisis moderno. Ed. Reverté, Barcelona, España (1966).

[4] EISENFELD J. "Operator Equations and Nonlinear Eigenparameter Problems", J. Funct. Anal. 4, 475-490 (1973).

[5] GOHBERG, I.C., LANCASTER, P. and RODMAN, L., Matrix Polynomials, Academic Press, New York, (1982).

[6] GOLUB, G. and VAN LOAN, C.P., Matrix Computations, John Hopkins Univ. Press, Baltimore, M.D. (1983).

[7] JÓDAR, L. "Boundary value problems for second order operator differential equations", Linear Algebra Appls. 83, 29-38 (1986).

[8] LANCASTER, P., Lambda Matrices and Vibrating Systems, Pergamon Elmsford, New York (1966).

[9] LOVASS-NAGY, L. and POWERS D.L. "A note on block diagonalization of some partioned matrices", Linear Algebra Appls. 5, 339-346 (1972).

[10] RUNCKELL, H. J. and PITTELKOW, P. "Practical computation of matrix functions", Linear Algebra Appls. 49, 161-178 (1983).

[11] VAN LOAN, C.P. "Computing integrals involving the matrix exponential", IEEE Trans. Autom. Control 23, 395-404 (1974).

Departamento de Matemática Aplicada.
Universidad Politécnica de Valencia.
Apdo. 22.012, Valencia, Spain.

Manuscript received in November 13, 1987 and in final form January 14, 1988