To the Graduate Council:

I am submitting herewith a thesis written by Sybil Ann Blalock entitled "Computation of the Cosine of a Matrix." I recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Mathematics.

Steven M. Serbin

Steven M. Serbin, Major Professor

We have read this thesis and recommend its acceptance:

Yueh-er Kuo

A. Samuel Jordan

Accepted for the Council:

L. Evans Roth

Vice Chancellor
Graduate Studies and Research

COMPUTATION OF THE COSINE OF A MATRIX

A Thesis

Presented for the

Master of Science

Degree

The University of Tennessee, Knoxville

Sybil Ann Blalock

December 1979

ACKNOWLEDGEMENTS

To Dr. Steven M. Serbin I wish to express my sincerest appreciation for his unlimited time, patience and guidance during the preparation of this thesis. I also wish to thank Mrs. Beth Kelley for the typing of this manuscript.

I offer much thanks to my family and friends, especially my father and mother, whose teachings of the value of an education, along with their continuous encouragement and support have played an important role in obtaining this goal.

ABSTRACT

This thesis concerns the computation of the cosine of a matrix. To approximate the matrix cosine various approaches involving matrix series, matrix characteristic polynomials, differential equations, matrix eigenvalues and matrix decomposition are described. In particular, we develop the double angle method and discuss the effectiveness of this technique when applied to certain approximation schemes. Computations are done primarily with three algorithms, those being in the form of rational approximants or truncated Taylor series. Further, we deal with a method developed by B. N. Parlett to calculate the functions of a matrix. It is of interest to us to see how the double angle method compares with Parlett's method. Results of our experiments are presented and analyzed.

TABLE OF CONTENTS

LIST OF TABLES

TABLE PAGE

INTRODUCTION

In recent years there has been extensive work done to develop new algorithms and improve existing algorithms which would enable one to approximate more accurately the matrix exponential. An excellent survey of the many approaches to this problem has been done by Moler and Van Loan [5]. Of the many successful methods, one which proceeds without the determination of the eigensystem is the scaling and squaring method proposed and analyzed by Ward [10]. One important interest in the computation of the exponential matrix is in the solution of the first-order differential system

$$\text{(1.1)} \qquad \underline{y}'(t) = A\underline{y}(t)$$
$$\underline{y}(0) = y_0$$

where A is a given $N \times N$ matrix of constants and $\underline{y}$ is an N vector. Such a system might arise, for example, in the semi-discretization of the heat equation.

In this thesis, we are interested not in the first order system, but in the second order system

$$\text{(1.2)} \qquad \underline{y}''(t) + A^2\underline{y}(t) = 0$$
$$\underline{y}(0) = y_0, \qquad \underline{y}'(0)$$

which may arise, for example, from the semi-discretization of the wave equation. In theory, the solution of this second order system is given by

$$\underline{y}(t) = \cos At\ \underline{y}_0$$

where we define cos At to be the power series

$$\cos At = \sum_{j=0}^{\infty} \frac{(-1)^j}{(2j)!} t^{2j} A^{2j} \tag{1.3}$$

convergent for all t. In order to express this solution in some applicable form we must therefore be able to compute the matrix function cos At. Consequently we address ourselves to the problem of computation of the cosine of a matrix A.

In principle there are many ways the cosine of a matrix could be computed. In practice, consideration of efficiency, stability and accuracy indicates that a few of the methods are preferable to others. In Chapter I we outline some approximation techniques for the cosine. These techniques may be classified as series methods, polynomial methods, ordinary differential equations methods, and matrix decomposition methods. Effectiveness of the algorithms is concerned with stability, accuracy, efficiency, ease of use and simplicity.

In Chapter II we elaborate on some specific methods. More particularly, we discuss the validity of the double angle method and develop error bounds for this method. Chapter III contains outlines of the methods we have used for computation in this thesis, those being Taylor series, rational approximation and a method developed by B. N. Parlett. Programming aspects and effectiveness of these algorithms are also discussed.

Lastly we provide numerical results involving some particular methods in Chapter IV. Observations are made with regard to the results obtained in our tests. Any comments or conclusions contained in Chapter IV are based solely on the data produced from the test matrices.

CHAPTER I

SURVEY OF METHODS

Let us consider some methods in general for computing the matrix cosine. As mentioned earlier, Moler and Van Loan have examined many approaches to compute the matrix exponential. Analogously, we want to apply some of these methods to the calculation of the matrix cosine. We may categorize the approximation schemes under consideration as series methods, polynomial methods, matrix decomposition methods and differential equation methods.

At this time it is convenient to review some conventions and definitions. Matrices will be denoted by capital letters. The following types of matrices will be referred to in our discussion:

A Hermitian $\leftrightarrow A^* = A$

A normal $\leftrightarrow A^*A = AA^*$

P unitary $\leftrightarrow P^*P = I$

T upper triangular $\leftrightarrow t_{ij} = 0, \qquad i > j$

D diagonal $\leftrightarrow d_{ij} = 0, \qquad i \neq j$

We shall work with the vector Euclidean, or 2-norm defined as

$$\|x\| = \left[\sum_{i=1}^{N} |x_i|^2\right]^{1/2}, \qquad x = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_N \end{bmatrix}$$

For a matrix A the 2 norm is

$$\|A\| = \max_{\|x\|=1} \|Ax\| = [\rho(A^TA)]^{1/2}$$

Note that if A is symmetric, $\|A\| = \rho(A)$, the spectral radius of A.

The common characteristic of series methods for the exponential is the application of an approximation scheme for the scalar function e^t being adapted to matrices. Similarly we would like to utilize approximation techniques for the scalar function $\cos t$ to obtain approximations for the cosine matrix. These methods also do not depend upon the eigensystem of the matrix. To begin, one approach is simply a Taylor series approximation. That is, we take the defining series expansion

$$\cos A = I - \frac{A^2}{2!} + \frac{A^4}{4!} - \ldots + (-1)^m \frac{A^{2m}}{(2m)!} + \ldots$$

truncated at some designated point and accept this as an approximation. One may encounter a problem of inefficiency due to repeated matrix multiplications. That is, the order 2m may have to be large before we obtain a good approximation. The accuracy of the method may be affected by the norm $\|A\|$. For example, if $\|A\|$ is large, the approximation may not be accurate. Also, there may be a problem with subtractive cancellation.

A second series method involves the Padé approach to determine coefficients $\{\alpha_i\}$ and $\{\beta_i\}$ in the rational approximation

$$R_{mn}(A) = \frac{\alpha_0 I + \alpha_1 A^2 + \ldots + \alpha_m A^{2m}}{\beta_0 I + \beta_1 A^2 + \ldots + \beta_n A^{2n}} = \frac{P_{mn}(A)}{Q_{mn}(A)}$$

which would be interpreted rigorously as $Q_{mn}(A)^{-1}P_{mn}(A)$. Replacing the operator A by the scalar z, one can analyze the error in this approximation by considering

$$E(z) \equiv \left[\sum_{K=0}^{\infty} \frac{(-z)^k}{(2k)!}\right] Q_{mn}(z) - P_{mn}(z) = O(z^\nu)$$

for some ν [8]. Normally we try to make ν as large as possible to provide greatest accuracy. Although accuracy plays an important role it may also be of interest to examine methods for which ν is not optimal but for which there are other desirable features. For example, smaller computational effort or less restrictive stability requirements may outweigh increased accuracy provided by making ν as large as possible.

Thirdly, we may approximate $\cos A$ by the quotient of polynomials of a special form

$$\text{(1.4)} \qquad R(A) = \frac{\sum_{j=0}^{\alpha} \phi_j(\beta)A^{2j}}{(I + \beta A^2)^\alpha} \equiv \frac{P}{Q}$$

For our purposes we will limit the discussion to cases where $\gamma \leq 2$. Information on schemes for approximate solution of differential equations of the form (1.2) corresponding to (1.4) may be found in [2]. This reference contains some numerical results in addition to an analysis of the stability and accuracy of these schemes.

As $\|A\|$ increases or as the spread of the eigenvalues increases, difficulties in computing the Padé and Taylor approximants also increase. For the exponential matrix a reliable method has been suggested to control these difficulties by exploiting the property

$$e^A = (e^{A/m})^m.$$

The idea is to choose m to be a power of 2 for which $e^{A/m}$ can be accurately computed and then form $(e^{A/m})^m$ by repeated squaring. This property is unique to the exponential function, but it does suggest the possibility of applying another identity. The trigonometric identity which we shall employ is the double angle formula applied to matrices, i.e.

$$\cos 2A = 2 \cos^2 A - I$$

which we shall prove later. Again, the principle would be to choose m as an appropriate power of 2 so that $\cos A/m$ can be efficiently computed and then form $\cos A$ by repeated use of the double angle formula. Practical criteria for choosing m may be based upon the norm $\|A\|$, i.e. choose m such that $\|A/m\| \leq 1$. Validity of this method will be established and various error bounds will be developed and discussed.

Polynomial Methods

By means of interpolatory polynomials we may offer certain approximations for the $\cos A$. Let the characteristic polynomial be

(1.5) $$C(z) = \det (zI - A) = z^n - \sum_{k=0}^{n-1} C_k z^k.$$

The Cayley Hamilton theorem gives us $C(A) = 0$ so that

$$A^n = C_0 I + C_1 A + \ldots + C_{n-1} A^{n-1}.$$

From this result it is possible to express any power of A in terms of $I, A, \ldots, A^{n-1}$; in particular

(1.6) $$A^{2k} = \sum_{k=0}^{n-1} \beta_{2k,j} A^j$$

By definition

(1.7) $$\cos At = \sum_{k=0}^{\infty} \frac{(-1)^k t^{2k} A^{2k}}{(2k)!}$$

so replacing A^{2k} with the expression (1.6)

$$\cos At = \sum_{k=0}^{\infty} \frac{(-1)^k t^{2k}}{(2k)!} \left[\sum_{j=0}^{n-1} \beta_{2k,j} A^j \right]$$

$$= \sum_{j=0}^{n-1} \left[\sum_{k=0}^{\infty} \frac{(-1)^k t^{2k}}{(2k)!} \beta_{2k,j} \right] A^j$$

$$= \sum_{j=0}^{n-1} \alpha_j(t) A^j .$$

Polynomial methods we shall mention involve a similar use of the characteristic polynomial.

Lagrange interpolation yields an approximating polynomial such as

$$\cos At = \sum_{k=1}^{n} \prod_{\substack{j=1 \\ j \neq k}}^{n} \left(\frac{A - \lambda_j I}{\lambda_k - \lambda_j} \right) \cos t\lambda_k$$

where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A.

Newton's form of the interpolating polynomial is

$$\cos At = (\cos \lambda_1 t) I + \sum_{k=2}^{n} \prod_{i=1}^{k-1} (A - \lambda_i I) \phi[\lambda_1, \ldots, \lambda_k]$$

where $\phi[\lambda_1, \ldots, \lambda_k]$ is the divided difference of order $k - 1$ which uses data points $\lambda_1, \ldots, \lambda_k$. This expression is more economical in terms of matrix multiplications, requiring (n - 1) matrix multiplications and one extra matrix array for temporary storage of the partial products [7].

There are several ways to evaluate the matrices

$$A_k = \prod_{\substack{j=1 \\ j \neq k}}^{n} \frac{A - \lambda_j I}{(\lambda_k - \lambda_j)}$$

which are required in Lagrange interpolation. One method involves the Vandermonde matrix of eigenvalues

$$V = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \lambda_1 & \lambda_2 & \cdots & \lambda_n \\ \cdot & \cdot & & \\ \cdot & \cdot & & \\ \cdot & \cdot & & \\ \lambda_1^{n-1} & \lambda_2^{n-1} & \cdots & \lambda_n^{n-1} \end{bmatrix}$$

If ν_{jk} is the (j,k) entry of V^{-1} then

$$A^k = \sum_{j=1}^{n} \nu_{kj} A^{j-1}$$

and

$$\cos At = \sum_{k=1}^{n} (\cos t\lambda_k) A^k.$$

Techniques based on interpolation have several disadvantages. These techniques require $O(n^4)$ basic arithmetic operations and extra storage needed for A and cos A. So except for small n these methods become quite expensive. Another weakness is that the eigenvalues of A must be known and problems may arise when the eigenvalues are nearly confluent even though there are special forms for interpolatory polynomials in the confluent case.

Matrix Decomposition Methods

Matrix decomposition methods are based on similarity transformations of the form

$$A = SBS^{-1}.$$

In this case it follows from the definition of $\cos At$ that

$$\cos At = S \cos Bt\, S^{-1}$$

so we want to find an S for which $\cos Bt$ is fairly simple to compute. These methods may be found to be quite efficient, but they suffer the difficulties involving confluent or nearly confluent eigenvalues.

One possibility is to take S as the matrix whose columns are eigenvectors of A, that is

$$S = \{s_1 \mid \ldots \mid s_n\}$$

where

$$As_j = \lambda_j s_j \qquad j = 1, \ldots, n.$$

Then by writing these equations as

$$AS = SD,$$

where $D = \text{diag}\,(\lambda_1, \ldots, \lambda_n)$, we may define $\cos At$ by

$$\cos At = S \cos Dt\, S^{-1}.$$

Note that S must be nonsingular. In theory, A must therefore have a complete set of linearly independent eigenvectors. If a complete set of linearly independent eigenvectors does not exist then S is not invertible and the whole algorithm breaks down.

With regard to the differential equation

$$\ddot{x} + A^2 x = 0 \tag{1.8}$$

we may express the solution in terms of the eigenvectors. First the initial condition is written as a linear combination of the eigenvectors

$$x(0) = \sum_{i=1}^{n} \alpha_j s_j \tag{1.9}$$

and then the solution x(t) is given by

(1.10) $$x(t) = \sum_{j=0}^{n} \alpha_j (\cos \lambda_j t) s_j .$$

The coefficients α_j are found by solving a set of linear equations $S\alpha = x(0)$. The solution (1.10) may be verified by substituting x(t) into (1.8). In theory, difficulty occurs when A does not have a complete set of linearly independent eigenvectors. For then there is no invertible matrix of eigenvectors S and the algorithm breaks down.

Any square matrix A, whether complex or real, is unitarily similar to an upper triangular complex matrix C [6]:

$$A = UCU^*.$$

The matrix C will be upper triangular if A has real eigenvalues. Otherwise we ask for a modification of C, called the real Schur form, which is block upper triangular or quasi-triangular. Today standard programs are available for effecting the Schur decomposition. Since $f(A) = Uf(C)U^*$, for any regular function f, the usefulness of this decomposition depends on the cost of forming f(C). Computing functions of triangular or quasi-triangular matrices is dealt with in a recent paper by Parlett [5]. Again, trouble occurs in the case of nearly confluent eigenvalues. Parlett develops the method in greater detail, covering the treatment of multiple eigenvalues.

Differential Equation Methods

Sophisticated and powerful methods for the numerical solution of general nonlinear differential equations have been developed in recent years. These methods are easy to use when computing e^{tA} and require very little additional programming or other thought [5].

It is possible to reduce the system (1.2) to (1.1) by defining the 2N-vector

$$\underline{u}(t) = \begin{bmatrix} \underline{y}(t) \\ \underline{y}'(t) \end{bmatrix} \tag{1.11}$$

so that

$$\underline{u}'(t) = \begin{bmatrix} 0 & I \\ -A^2 & 0 \end{bmatrix} \underline{u}(t). \tag{1.12}$$

Then we can apply any of the o.d.e. programs available for solving (1.1).

Another approach that is a standard idea in differential equations is to approximate $\cos A$ by using a linear recurrence relation. We define

$$Z_j = \cos \frac{jA}{N}, \qquad Z_0 = I$$

Then,

$$Z_{j+1} - 2Z_1 Z_j + Z_{j-1} = 0. \tag{1.13}$$

The Z_j's are generated by (1.13), and we may verify this identity using basic trigonometric identities for $\cos(\alpha \pm \beta)$:

$$\cos \frac{(j+1)A}{N} - 2 \cos \frac{A}{N} \cos \frac{jA}{N} + \cos \frac{(j-1)A}{N}$$

$$= \cos \left(\frac{jA}{N} + \frac{A}{N}\right) - 2 \cos \frac{A}{N} \cos \frac{jA}{N} + \cos \left(\frac{jA}{N} - \frac{A}{N}\right)$$

$$= \cos \frac{jA}{N} \cos \frac{A}{N} - \sin \frac{jA}{N} \sin \frac{A}{N} - 2 \cos \frac{A}{N} \cos \frac{jA}{N}$$
$$+ \cos \frac{jA}{N} \cos \frac{A}{N} - \sin \frac{jA}{N} \sin \left(\frac{-A}{N}\right)$$

$$= \cos \frac{jA}{N} \cos \frac{A}{N} - \sin \frac{jA}{N} \sin \frac{A}{N} - 2 \cos \frac{A}{N} \cos \frac{jA}{N}$$
$$+ \cos \frac{jA}{N} \cos \frac{A}{N} + \sin \frac{jA}{N} \sin \frac{A}{N}$$

$$= 0.$$

Now we approximate Z_1 by $R(\frac{A}{N}) = W_1$, some rational approximation, and generate W_j's with the linear recurrence

(1.14) $$W_{j+1} = 2W_1W_j - W_{j-1}$$

Then W_N approximates $Z_N = \cos A$.

The accuracy of this technique depends on the accuracy of W_1 and on the value of N. If N is large then the arithmetic operations needed to calculate W_N will be costly. In addition, repeated matrix multiplications increase the effects of round-off error.

Other Methods

As discussed earlier, the solution of (1.2),

(1.15) $$y(t) = \cos At\, y_0$$

can be obtained by noting that the second order matrix differential system is equivalent to a first order system as in (1.12), for which the solution is

$$\underline{u}(t) = e^{tB}\underline{u}(0) \qquad \underline{u}(0) = \begin{bmatrix} \underline{y}(0) \\ \underline{y}'(0) \end{bmatrix} \tag{1.16}$$

where

$$B = \begin{bmatrix} 0 & I \\ -A^2 & 0 \end{bmatrix}.$$

To calculate $\cos At$ we compute the exponential series e^{tB} then substitute in (1.16), from which we find $\cos At$. There is a procedure described by T. M. Apostol [1] for calculating $\cos At$ without the need of computing the exponential in (1.16). The method is much simpler than computing $\underline{u}(t)$ because the matrices in (1.16) are twice the size of those required by this procedure. Explicit formulas for solutions of the second-order matrix differential equation

$$\underline{y}''(t) = A^2\underline{y}(t)$$

with prescribed initial values $\underline{y}(0)$ and $\underline{y}'(0)$ are given by Apostol. It should be noted that this method may be extended for the matrix differential equation of order r,

$$Y^{(r)}(t) = AY(t)$$

with prescribed initial values $Y(0), Y'(0), \ldots, Y^{(r-1)}(0)$.

We shall consider the matrix differential system (1.2). In particular, let us outline Apostol's procedure where A is either 2×2

or 3×3. First, suppose A is 2×2 having eigenvalues λ_1, λ_2. Then the solution to (1.2) is given by

$$Y(t) = y_1(t)I + y_2(t)(A^2 - \lambda_1^2 I) \equiv C(t) \tag{1.17}$$

where y_1 and y_2 are determined by the following triangular system of second-order linear differential equations with constant coefficients

$$\begin{aligned} &y_1'' + \lambda_1^2 y_1 = 0 && y_1(0) = 1 && y_1'(0) = 0 \\ &y_2'' + \lambda_2^2 y_2 = -y_1 && y_2(0) = 0 && y_2'(0) = 0 \,. \end{aligned} \tag{1.18}$$

Now, let us show that C(t) determined by (1.17) and (1.18) satisfies the differential equation $C''(t) + A^2C(t) = 0$ and the initial conditions $C(0) = I$ and $C'(0) = 0$. Differentiating C(t) yields

$$C'(t) = y_1'(t) + y_2'(t)(A^2 - \lambda_1^2 I)$$

and

$$C''(t) = y_1''(t)I + y_2''(t)(A^2 - \lambda_1^2 I).$$

Then

$$C''(t) + A^2C(t) = y_1''(t)I + y_2''(t)(A^2 - \lambda_1{}^2I) + A^2[y_1(t) + y_2(t)(A^2 - \lambda_1{}^2I)]$$

$$= -\lambda_1{}^2y_1I + (-y_1(t) - \lambda_2{}^2y_2)(A^2 - \lambda_1{}^2I)$$

$$+ A^2y_1(t) + A^2y_2(t)(A^2 - \lambda_1{}^2I)$$

$$= (A^2 - \lambda_1{}^2I)y_1(t) - (A^2 - \lambda_1{}^2I)y_1(t)$$

$$- (A^2 - \lambda_1{}^2I)\lambda_2{}^2y_2(t) + A^2y_2(t)(A^2 - \lambda_1{}^2I)$$

$$= (A^2 - \lambda_1{}^2I)(A^2 - \lambda_2{}^2I)y_2(t)$$

$$= 0$$

by Cayley Hamilton theorem. Using the conditions in (1.18) we have

$$C(0) = y_1(0)I + y_2(0)(A^2 - \lambda_1{}^2I) = I$$

and

$$C'(0) = y_1'(0)I + y_2'(0)(A^2 - \lambda_1{}^2I) = 0$$

so that C satisfies the initial conditions. Further, this approach also handles the case of multiple eigenvalues; we suppose that $\lambda_1 = \lambda_2 = \lambda \neq 0$. The solutions to (1.18) are

$$y_1(t) = \cos \lambda t$$

$$y_2(t) = -\frac{t}{2\lambda} \sin \lambda t$$

Hence we have

$$C(t) = (\cos \lambda t)I - \frac{t}{2\lambda} \sin \lambda t(A^2 - \lambda^2 I). \tag{1.19}$$

To verify that this C satisfies (1.2) differentiate C(t) to get

$$C'(t) = -\lambda \sin \lambda tI - \frac{t}{2} \cos \lambda t(A^2 - \lambda^2 I) - \frac{1}{2\lambda} \sin \lambda t(A^2 - \lambda^2 I)$$

and

$$C''(t) = -\lambda^2 \cos \lambda tI - \frac{1}{2} \cos \lambda t(A^2 - \lambda^2 I) + \frac{t\lambda}{2} \sin \lambda t(A^2 - \lambda^2 I)$$
$$- \frac{1}{2} \cos \lambda t(A^2 - \lambda^2 I).$$

Substituting in the differential equation yields

$$\begin{aligned}
C''(t) + A^2 C(t) &= -\lambda^2 \cos \lambda tI - \frac{1}{2} \cos \lambda t(A^2 - \lambda^2 I) \\
&\quad + \frac{t\lambda}{2} \sin \lambda t(A^2 - \lambda^2 I) - \frac{1}{2} \cos \lambda t(A^2 - \lambda^2 I) \\
&\quad + A^2(\cos \lambda t)I - \frac{t}{2\lambda} A^2 \sin \lambda t(A^2 - \lambda^2 I) \\
&= (A^2 - \lambda^2 I) \cos \lambda t - (A^2 - \lambda^2 I) \cos \lambda t \\
&\quad + (t \sin \lambda t)(A^2 - \lambda^2 I)\left(\frac{\lambda I}{2} - \frac{A^2}{2\lambda}\right) \\
&= \frac{-1}{2\lambda} t \sin\lambda t(A^2 - \lambda^2 I)(A^2 - \lambda^2 I) \\
&= 0.
\end{aligned}$$

At $t = 0$,

$$C(0) = I, \qquad C'(0) = 0$$

satisfying the initial conditions. Thus, (1.19) is a solution to (1.2) when $\lambda_1 = \lambda_2 \neq 0$.

In the case where A is 3×3 the solution to (1.2) is given by

(1.20)
$$C(t) = y_1(t)I + y_2(t)(A^2 - \lambda_1^2 I) + y_3(t)(A^2 - \lambda_1^2 I)(A^2 - \lambda_2^2 I)$$

with y_1, y_2, y_3 being determined by

(1.21)
$$\begin{array}{lll} y_1'' + \lambda_1^2 y_1 = 0 & y_1(0) = 1 & y_1'(0) = 0 \\ y_2'' + \lambda_2^2 y_2 = -y_1 & y_2(0) = 0 & y_2'(0) = 0 \\ y_3'' + \lambda_3^2 y_3 = -y_2 & y_3(0) = 0 & y_3'(0) = 0. \end{array}$$

To show that such a C actually works, differentiate (1.20) and substitute in the following equation using the conditions in (1.21)

$$\begin{aligned} C'' + A^2 C &= y_1'' + y_2''(A^2 - \lambda_1^2 I) + y_3''(A^2 - \lambda_1^2 I)(A^2 - \lambda_2^2 I) \\ &\quad + A^2 y_1 + A^2 y_2(A^2 - \lambda_1^2 I) + A^2 y_3(A^2 - \lambda_1^2 I)(A^2 - \lambda_2^2 I) \\ &= -\lambda_1^2 y_1 + (-y_1 - \lambda_2^2 y_2)(A^2 - \lambda_1^2 I) \\ &\quad + (-y_2 - \lambda_3^2 y_3)(A^2 - \lambda_1^2 I)(A^2 - \lambda_2^2 I) + A^2 y_1 \\ &\quad + A^2 y_2(A^2 - \lambda_1^2 I) + A^2 y_3(A^2 - \lambda_1^2 I)(A^2 - \lambda_2^2 I) \\ &= (A^2 - \lambda_1^2 I)y_1 - (A^2 - \lambda_1^2 I)y_1 + (A^2 - \lambda_2^2 I)(A^2 - \lambda_2^2 I)y_2 \\ &\quad + (A^2 - \lambda_3^2 y_3 I - y_2 I)(A^2 - \lambda_1^2 I)(A^2 - \lambda_2^2 I) \\ &= (A^2 - \lambda_3^2 I)(A^2 - \lambda_2^2 I)(A^2 - \lambda_1^2 I)y_3 \\ &= 0. \end{aligned}$$

Also, at $t = 0$, we have

$$C(0) = I, \qquad C'(0) = 0$$

so the initial conditions are satisfied. We see then, that $C(t)$ defined by (1.19) and (1.20) provides a solution to (1.2).

Further, let us suppose that $\lambda_1 = \lambda_2 = \lambda_3 = \lambda \neq 0$. In this particular case, y_1, y_2, y_3 are given by

$$y_1 = \cos \lambda t$$

$$y_2 = -\frac{t}{2\lambda} \sin \lambda t$$

$$y_3 = \frac{t}{8\lambda^3} \sin \lambda t - \frac{t^2}{8\lambda^2} \cos \lambda t.$$

The solution $C(t)$ becomes

$$C(t) = \cos \lambda t I - \frac{t}{2\lambda} \sin \lambda t(A^2 - \lambda^2 I) + \left[\frac{t}{8\lambda^3} \sin \lambda t - \frac{t^2}{8\lambda^2} \cos \lambda t\right](A^2 - \lambda^2 I)^2. \tag{1.22}$$

Again, as in the 2×2 case, the solution may be verified by simply substituting $C(t)$ into (1.2).

This method may be extended for an $n \times n$ matrix A and offers an alternative to those methods with computational problems when A has multiple eigenvalues. However, for n very large it seems impractical to solve the triangular system of second-order equations analytically.

CHAPTER II

APPROXIMATIONS AND THE DOUBLE ANGLE FORMULA

Now, let us focus our attention on some particular methods chosen to compute the cosine of a matrix. The first method is a Taylor series approximation. We take the series expansion

$$\cos A = I - \frac{A^2}{2!} + \frac{A^4}{4!} - \frac{A^6}{6!} + \ldots + (-1)^j \frac{A^{2j}}{(2j)!} + \ldots$$

and truncate this series at a predetermined point and accept this finite sum as an approximation.

Our second approach leads us to rational approximations. We shall use the Padé approach to determine the coefficients $\{\alpha_i\}$ and $\{\beta_i\}$ where

$$R(A) = \frac{\alpha_0 I + \alpha_1 A^2 + \alpha_2 A^4}{\beta_0 I + \beta_1 A^2 + \beta_2 A^4} .$$

The second rational approximation we shall consider takes the general form [2]

$$R(A) = \frac{\sum_{j=0}^{\alpha} \phi_j(\beta) A^{2j}}{(I + \beta^2 A^2)^{\alpha}} .$$

Lastly we will compute the cosine using an algorithm developed by B. N. Parlett [7] and compare these results with other approximations discussed above.

The Padé and Taylor approximants involve the idea of scaling the matrix A by 2^{-j} for some $j \in \mathbb{Z}^+$ and then using an extension of the double angle formula applied to matrices

$$\cos 2A = 2 \cos^2 A - I$$

to obtain cos A. So we first form the matrix $\frac{A}{m}$ where we choose $m = 2^j$ for some appropriate j. Next the matrix $\cos \frac{A}{m}$ is calculated by one of the methods discussed and finally we employ the double angle formula to obtain our approximation cos A. Therefore the question of convergence of the double angle method arises and so we direct our attention to this point.

Convergence of the Double Angle Method

Let A be a normal matrix, and denote

$$\sigma(A) = \text{spectrum of } A.$$

As before

$$\rho(A) = \text{spectral radius of } A$$

$$\|A\| = \|A\|_2 = \sqrt{\rho(AA^*)} = \sup \frac{\|Ax\|_2}{\|x\|_2} .$$

The significance of normality is in the following result [4]:

Let f be any function which is analytic on a neighborhood of $\sigma(A)$. Then

$$\|f(A)\| = \sup_{\lambda \in \sigma(A)} |f(\lambda)| \tag{2.1}$$

Now we compute our approximation as follows: Let

$$Y_j = \cos \frac{A}{2^{m-j}}$$

where in this section m is used to denote a power of 2. Then

$$Y_{j+1} = \cos 2\,\frac{A}{2^{m-j}} = 2\cos\frac{A}{2^{m-j}} - I.$$

Although this last statement is true for scalars, it may not be quite so obvious for matrices so this will be our first lemma.

LEMMA 1: Let A be a matrix. Then

$$\cos 2A = 2\cos^2 A - I \tag{2.2}$$

PROOF: Let $\cos At$ be defined by the power series

$$\sum_{j=0}^{\infty} \frac{(-t)^j}{(2j)!} A^{2j} \tag{2.3}$$

and $\sin At$ be defined by the power series

$$\sum_{j=0}^{\infty} \frac{(-t)^j}{(2j+1)!} A^{2j+1} \tag{2.4}$$

Define

$$R_K = \sum_{j=K+1}^{\infty} \frac{(-1)^j}{(2j)!} t^{2j} A^{2j}. \tag{2.5}$$

Then

$$\|R_K\| \le \sum_{j=K+1}^{\infty} \frac{1}{(2j)!} (t\|A\|)^{2j}. \tag{2.6}$$

But the series

$$\sum_{j=1}^{\infty} \frac{(t\|A\|)^{2j}}{(2j)!}$$

converges for all t to $\cos(t\|A\|)$; thus $\|R_K\| \to 0$ as $K \to \infty$, and so the series (2.3) converges for all t. Similarly it may be shown that (2.4) also is convergent for all t.

Since these are convergent power series, term-by-term differentiation of (2.3) and (2.4) yields

$$(2.7) \qquad D_t(\cos At) = -A \sin At$$
$$D_t(\sin At) = A \cos At.$$

First we want to show the more general property

$$(2.8) \qquad \cos (A + B)t = \cos At \cos Bt - \sin At \sin Bt$$

when $AB = BA$. Let

$$F(t) = \cos (A + B)t - \cos At \cos Bt + \sin At \sin Bt.$$

Then using (2.7)

$$\begin{aligned} F'(t) &= D_t[\cos (A + B)t] - [D_t(\cos At)] \cos Bt \\ &\quad + \cos (At) [D_t(\cos Bt)] + [D_t(\sin At)] \sin Bt \\ &\quad + \sin At [D_t(\cos Bt)] \\ &= -(A+B) \sin (A+B)t + A \sin At \cos Bt + B \cos At \sin Bt \\ &\quad + A \cos At \sin Bt + B \sin At \cos Bt \\ &= -(A+B) \sin (A+B)t + (A+B) \sin At \cos Bt \\ &\quad + (B+A) \cos At \sin Bt \end{aligned}$$

And

$$F''(t) = -(A + B)^2 \cos (A + B)t + A(A + B) \cos At \cos Bt$$
$$- B(A + B) \sin At \sin Bt + B(B + A) \cos At \cos Bt$$
$$- A(B + A) \sin At \sin Bt$$

$$= -(A + B)^2 \cos (A + B)t$$
$$+ [(A + B)^2 \cos At \cos Bt - (A + B)^2 \sin At \sin Bt].$$

Hence

$$F''(t) + (A + B)^2 F(t) = -(A + B)^2 \cos (A + B)t$$
$$+ (A+B)^2[\cos At \cos Bt - \sin At \sin Bt]$$
$$+ (A + B)^2 \cos (A + B)t$$
$$- (A+B)^2[\cos At \cos Bt - \sin At \sin Bt]$$

$$= 0.$$

Then by the uniqueness theorem for second order equations, since $F(0) = 0 = F'(0)$, we may conclude $F(t) = 0$. Therefore

(2.9) $$\cos (A + B)t = \cos At \cos Bt - \sin At \sin Bt.$$

Now as a special case let $t = 1$, $B = A$ so that

(2.10) $$\cos 2A = \cos^2 A - \sin^2 A$$

and with $t = 1$, $B = -A$

(2.11) $$\cos 0 = (\cos A) \cos (-A) - \sin A \sin (-A).$$

By definition

$$\cos(-A) = \sum_{j=0}^{\infty} \frac{(-1)^j}{(2j)!} (-A)^{2j}$$

$$= \sum_{j=0}^{\infty} \frac{(-1)^j}{(2j)!} A^{2j} = \cos A$$

and

$$\sin(-A) = \sum_{j=0}^{\infty} \frac{(-1)^j}{(2j+1)!} (-A)^{2j+1}$$

$$= \sum_{j=0}^{\infty} \frac{(-1)^j}{(2j+1)!} (-1)^{2j+1} A^{2j+1} = -\sin A.$$

Thus (2.11) becomes

$$\cos 0 = I = \cos^2 A + \sin^2 A$$

or

$$\cos^2 A - I = -\sin^2 A.$$

Using this identity in (2.10) yields

$$\cos 2A = \cos^2 A + \cos^2 A - I$$

$$= 2\cos^2 A - I. \qquad \blacksquare$$

Having obtained the lemma, now we use the approximation

$$U_0 = R(A/2^m)$$

where $R(z)$ is some rational approximation to $\cos z$, and then define the nonlinear recurrence

$$U_{j+1} = 2U_j^2 - I \qquad j = 0, \ldots, m-1 \tag{2.12}$$

Note that the Chebyshev polynomial of the first kind is

(2.13) $$T_2(x) = 2x^2 - 1$$

so we have

(2.14) $$U_{j+1} = T_2(U_j).$$

The Chebyshev polynomial of degree r is given by

(2.15) $$T_r(x) = \cos (r \cos^{-1} x).$$

From this, it follows that in $[-1,1]$ $T_r(x)$ has r zeros at

(2.16) $$x = \cos \frac{(2j+1)\pi}{2r} \qquad j = 0, \ldots, r-1$$

and $r + 1$ extrema of magnitude 1 at

(2.17) $$x = \cos j\pi/r \qquad j = 0, \ldots, r$$

where (2.17) includes the $r - 1$ places where $T'_r(x) = 0$ and the two endpoints. We see that

(2.18) $$\max_{-1\leq x\leq 1} |T_2(x)| = 1.$$

Each U_j is an approximation to

$$Y_j = \cos \frac{A}{2^{m-j}} .$$

In order to develop certain error bounds we would like to show that $\|Y_j\| \leq 1$ and $\|U_j\| \leq 1$. First we must introduce some lemmas to aid us in our proof.

LEMMA 2: Let U be a unitary matrix. Then

$$\|UA\|_2 = \|A\|_2 = \|AU\|_2.$$

PROOF: By definition

$$\|UA\|_2 = \rho^{1/2}[(UA)^*UA] = \rho^{1/2}(A^*U^*UA) = \rho^{1/2}(A^*A) = \|A\|_2.$$

Again by definition

$$\|AU\|_2 = \sup_{x\neq 0} \frac{\|AUx\|_2}{\|x\|_2}$$

$$= \sup_{y\neq 0} \frac{\|Ay\|_2}{\|U^*y\|_2}$$

where $y = Ux$. Since U is unitary

$$\|U^*y\|_2 = \|y\|_2$$

so that

$$\|AU\|_2 = \sup_{y\neq 0} \frac{\|Ay\|_2}{\|U^*y\|_2} = \sup_{y\neq 0} \frac{\|Ay\|_2}{\|y\|_2} = \|A\|_2. \qquad \blacksquare$$

LEMMA 3: Let A be a normal matrix. Then

$$\|A\|_2 = \rho(A).$$

PROOF: Since A is normal there exists U unitary such that $A = UDU^*$, and D is a diagonal matrix. Now

$$\|A\|_2 = \|UDU^*\|_2 = \|DU^*\|_2 = \|D\|_2$$

by lemma 2. By definition

$$\|D\|_2 = \rho^{1/2}(D^*D).$$

But D is diagonal so D*D has eigenvalues $|\lambda_i|^2$. Thus,

$$\rho^{1/2}(D^*D) = \sqrt{\max |\lambda_i|^2} = \max |\lambda_i| = \rho(D).$$

Since the eigenvalues of D are the eigenvalues of A

$$\|A\|_2 = \rho(A). \qquad \blacksquare$$

Now let f be an analytic function defined on the spectrum of A. For simplicity assume that A is normal so it has a complete set of orthonormal eigenvectors, i.e. the columns of U where $A = UDU^*$. We shall have occasion to use the following well-known result, whose proof we included for completeness.

THEOREM 1: If f is an analytic function defined on the spectrum of a normal matrix A, then f(A) is a normal matrix.

PROOF: One way to define the function f(A) is via the interpolating polynomial. We define $f(A) = g(A)$ where

$$g(\lambda) = \sum_{K=1}^{n} f(\lambda_K)\ell_K(\lambda)$$

and

$$\ell_K(\lambda) = \prod_{\substack{i=1 \\ i\neq K}}^{n} \frac{\lambda - \lambda_i}{\lambda_K - \lambda_i}$$

is the fundamental Lagrange polynomial (assuming distinct eigenvalues; the confluent case may similarly be treated). Suppose $Au_K = \lambda_K u_K$. Then

$$f(A)u_K = g(A)u_K$$

$$= (\sum_{j=0}^{n} C_j A^j)u_K = (\sum_{j=0}^{n} C_j \lambda_K^j)u_K$$

$$= g(\lambda_K)u_K$$

$$= f(\lambda_K)u_K .$$

Therefore the eigenvalues of g(A) are $f(\lambda_K)$ with corresponding eigenvectors u_K. This implies, recalling that the columns of U are the eigenvectors, that

$$f(A)U = g(A)U = Ug(D) = Uf(D)$$

or

$$f(A) = Uf(D)U^*$$

so f(A) can be diagonalized by a unitary transformation and is thereby normal. ■

Note that since f(A) is normal, by (2.1)

(2.19) $$\|f(A)\|_2 = \rho(f(A)) = \sup_{\lambda \varepsilon \sigma(A)} |f(\lambda)| .$$

THEOREM 2: Suppose further now that A has real eigenvalues (e.g. A Hermitian). Then, for $j = 0, \ldots, m$,

$$\|Y_j\| \leq 1$$

where

$$Y_j = \cos \frac{A}{2^{m-j}}$$

and, assuming that we can make

$$\|U_0\| \le 1, \qquad U_0 = R(\frac{A}{2m})$$

for some rational approximation $R(z)$ to $\cos z$.

PROOF: For each j, Y_j is an analytic function of the normal matrix $\frac{A}{2^{m-j}}$. Therefore from (2.19)

$$\|Y_j\| = \|\cos \frac{A}{2^{m-j}}\| = \sup_{\lambda\epsilon\sigma(\frac{A}{2^{m-j}})} |\cos \lambda| = \sup_{\mu\epsilon\sigma(A)} |\cos \frac{\mu}{2^{m-j}}| \le 1,$$

since μ is real. Assume that $\|U_0\| \le 1$. Recall that

$$U_{j+1} = 2U_j^{\ 2} - I = T_2(U_j).$$

For suitable choice of R, namely that no eigenvalue of A is a zero of the denominator of R, U_0 is an analytic function defined on the normal matrix $A/2^m$ and is thereby normal. From (2.18), since $\sigma(U_0) \subset [-1,1]$ (the eigenvalues of U_0 being real), by appropriate choice of m for the particular $R(z)$

$$\|U_1\| = \|T_2(U_0)\| = \sup_{\lambda\epsilon\sigma(U_0)} |T_2(\lambda)| \le 1.$$

Assume, inductively, that $\|U_j\| \le 1$. Then

$$U_{j+1} = T_2(U_j)$$

so that U_{j+1} is an analytic function of the matrix U_j which implies by Theorem 1 that

$$\|U_{j+1}\| = \|T_2(U_j)\| = \sup_{\lambda\epsilon\sigma(U_j)} |T_2(\lambda)| \le 1.$$

In conclusion we have for each $j = 0, \ldots, m$

$$\|U_j\| \leq 1. \qquad \blacksquare$$

Error Bound

Let $E_j = Y_j - U_j$, that is, E_j represents the error at the j^{th} step. Then

$$\begin{aligned} E_{j+1} &= (Y_{j+1} - U_{j+1}) \\ &= (2Y_j^2 - I) - (2U_j^2 - I) \\ &= 2(Y_j^2 - U_j^2) \\ &= 2(Y_j + U_j)(Y_j - U_j). \end{aligned}$$

Note that U_j and Y_j are both analytic functions of the matrix A so $Y_jU_j = U_jY_j$ [7]. Thus it is possible to factor $(Y_j^2 - U_j^2)$ as above. So

$$E_{j+1} = 2(Y_j + U_j)E_j.$$

Taking norms on both sides, we obtain

$$\begin{aligned} \|E_{j+1}\| &= \|2(Y_j + U_j)E_j\| \\ &\leq 2\|Y_j + U_j\| \; \|E_j\| \\ &\leq 4\|E_j\| \end{aligned} \tag{2.20}$$

since $\|Y_j\| \leq 1$ and $\|U_j\| \leq 1$. We recognize that this bound may be very pessimistic, but it appears to be the best we can do.

THEOREM 3: For every $j \geq 1$

(2.21) $$\|E_j\| \leq 2^{2j} \|E_0\|.$$

PROOF: We will show (2.21) by induction. For $j = 1$ we have

$$\|E_1\| \leq 4\|E_0\| = 2^2 \|E_0\|$$

which shows that the error bound (2.21) holds for $j = 1$.

Secondly let us assume that $\|E_j\| \leq 2^{2j}\|E_0\|$. By (2.20) we have

$$\begin{aligned} \|E_{j+1}\| &\leq 4\|E_j\| \\ &\leq 4\ 2^{2j}\|E_0\| \\ &= 2^2 2^{2j}\|E_0\| \\ &= 2^{2(j+1)}\|E_0\|. \end{aligned}$$

Consequently, for all j we have

$$\|E_j\| \leq 2^{2j}\|E_0\|. \qquad \blacksquare$$

Hence when $j = m$, $\|\cos A - U_m\| \leq 2^{2m}\|E_0\|$.

As stated, the error bound first obtained may look quite pessimistic, but suppose we have an approximation $r(z)$, $z \in \mathbb{R}$, such that

(i) $|r(z)| \leq 1 \qquad \forall\, |z| \leq \eta$ some constant (stability)

and

(ii) $|\cos z - r(z)| \leq C_\nu |z|^\nu$ for $|z| \leq \zeta$ some constant (accuracy) then this error bound may still prove to be useful in telling us how to choose the approximation and number of steps.

THEOREM 4: Suppose we have an approximation satisfying (i) and (ii) above. Then

$$\|\cos A - U_m\| \leq C_\nu 2^{m(2-\nu)}[\rho(A)]^\nu. \tag{2.22}$$

PROOF: By definition

$$\|E_0\| = \|\cos \frac{A}{2^m} - r(\frac{A}{2^m})\|$$

$$= \|\cos \frac{A}{2^m} - U_0\|.$$

Therefore if we desire

$$\|U_0\| = \|r(\frac{A}{2^m})\| = \sup_{\lambda\varepsilon\sigma(A)} |r(\frac{\lambda}{2^m})| \leq 1$$

we must make $|\frac{\rho(A)}{2^m}| \leq \eta$ to satisfy (i). Further we also require for (ii) $|\frac{\rho(A)}{2^m}| \leq \zeta$, ζ some constant determined by the scheme. Since

$$f(\frac{A}{2^m}) = \cos \frac{A}{2^m} - r(\frac{A}{2^m})$$

is an analytic function of $\frac{A}{2^m}$, from (2.1) and (ii)

$$\left\|\cos\frac{A}{2^m} - r\left(\frac{A}{2^m}\right)\right\| = \sup_{\lambda\varepsilon\sigma(A)} \left|\cos\frac{\lambda}{2^m} - r\left(\frac{\lambda}{2^m}\right)\right|$$

$$\leq C_\nu \left|\frac{\lambda}{2^m}\right|^\nu$$

$$\leq C_\nu \left|\frac{\rho(A)}{2^m}\right|^\nu$$

Thus

$$\|E_0\| = \left\|\cos\frac{A}{2^m} - r\left(\frac{A}{2^m}\right)\right\| \leq C_\nu \left|\frac{\rho(A)}{2^m}\right|^\nu .$$

Recalling that

$$\|\cos A - U_m\| \leq 2^{2m}\|E_0\|$$

we see that

$$\|\cos A - U_m\| \leq 2^{2m} C_\nu \left|\frac{\rho(A)}{2^m}\right|^\nu$$

$$= C_\nu \frac{2^{2m}}{2^{m\nu}} [\rho(A)]^\nu$$

$$= C_\nu 2^{m(2-\nu)} [\rho(A)]^\nu . \qquad \blacksquare$$

We remark that for A not normal, with real eigenvalues, this double-angle technique may still work, but the technique of proof will have to be altered.

With appropriate choices of m and ν the term (2.22) may be made small. For example, suppose we have

$$2^{\beta-1} \leq \rho(A) < 2^\beta \tag{2.23}$$

and we would like to make

$$\|\cos A - U_m\| \leq \frac{2^{2m}}{2^{\nu m}} 2^{\beta\nu} \tag{2.24}$$

$$= 2^{2m+(\beta-m)\nu} < \varepsilon$$

for some $\varepsilon > 0$. We want

$$2m + (\beta - m) < 0 \tag{2.25}$$

in order for (2.24) to hold for small ε. This requires

$$(2 - \nu)m + \beta\nu < 0 \tag{2.26}$$

or

$$m > \frac{\beta\nu}{\nu - 2}, \qquad \nu > 2.$$

Observe that if $\beta < 0$, this holds for any positive m. Suppose $\nu > 2$ is fixed and $\beta \geq 0$. We would expect that as ν increases, m may decrease. That is, as the accuracy increases, m may decrease. Clearly

$$g(\nu) \equiv \frac{\beta\nu}{\nu - 2}$$

decreases. Since

$$g'(\nu) = \frac{(\nu - 2)\ \beta - \beta\nu}{(\nu - 2)^2}$$

$$= \frac{-2\beta}{(\nu - 2)^2} < 0$$

if $\beta > 0$.

Now returning to (2.24) let us see how we need to pick m in order to insure $\|\cos A - U_m\| < \varepsilon$:

(2.27)
$$2m + (\beta - m)\nu < \log_2\varepsilon,$$

$$(2 - \nu)m < \log_2\varepsilon - \beta\nu,$$

$$m > \frac{\beta\nu - \log_2\varepsilon}{\nu - 2} .$$

For example, if $\beta = 5$ and $\varepsilon = 2^{-10}$

$$m > \frac{5\nu + 10}{\nu - 2}.$$

Under these conditions we have the following chart to illustrate what has been shown above

ν	3	4	5	6
m	25	15	12	10

To consider some specific approximations, let us examine first the Taylor method with matrix A being hermitian so that all its eigenvalues are real. By Taylor's theorem stated below, the cosine function may be put in the form

$$\cos x = \sum_{j=0}^{k} \frac{(-1)^j}{(2j)!} x^{2j} + R(x)$$

where

$$R(x) = \frac{f^{(k+2)}(c)}{(k + 2)!} x^{k+2}.$$

THEOREM (Taylor): Let I be an interval containing the number 0. If f has n continuous derivatives on I then for each x in I

$$f(x) = f(0) + f'(0)x + \frac{f''(0)}{2!} + \ldots + \frac{f^{(n-1)}(0)}{(n-1)!} x^{n-1} + R_n$$

where the remainder is given by the formula

$$R_n = \frac{1}{(n-1)!} \int_0^x f^{(n)}(t)(x - t)^{n-1} dt.$$

In general it is more convenient to write the remainder in Lagrangian form

$$R_n = \frac{f^{(n)}(c)}{n!} x^n$$

where c is some number between 0 and x.

Let us take x as a fixed real number and let

$$f(t) = \cos t.$$

Differentiating yields

$$f^{(2k)}(t) = (-1)^k \cos t$$

$$f^{(2k+1)}(t) = (-1)^k \sin t.$$

Clearly f has n continuous derivatives on any interval containing 0, for any n. The Taylor coefficients are

$$f^{(2k+1)}(0) = 0 \quad \text{and} \quad f^{(2k)}(0) = (-1)^k.$$

Therefore, letting $n = 2k + 2$,

$$\cos x = \sum_{j=0}^{k} \frac{(-1)^j}{(2j)!} x^{2j} + R_{2k+2}(x).$$

For all $x \in \mathbb{R}$, $|f^{(n)}(x)| \leq 1$, so Lagrange's form of the remainder satisfies

$$|R_{2k+2}| = \left|\frac{f^{(2k+2)}(c)}{(2k+2)!}\right| \; |x|^{2k+2} \leq \frac{|x|^{2k+2}}{(2k+2)!} \; .$$

Taking the approximation r(A) to be the truncated Taylor series

$$r(A) = \sum_{j=0}^{k} \frac{(-1)^j}{(2j)!} A^{2j},$$

we have for $x \in \mathbb{R}$

(2.28) (i) $|r(x)| \leq 1 \quad \forall \, |x| \leq \eta,$

and (ii) $|\cos x - r(x)| = |R(x)| \leq \dfrac{1}{(2k+2)!} \; |x|^{(2k+2)},$

$|x| \leq \delta$, some constant.

To show that there exists an η such that (i) holds, look at the expanded series r(x)

$$r(x) = 1 - \frac{x^2}{2} + \frac{x^4}{4!} - \ . \ . \ .$$

$$= 1 - \left(\frac{x^2}{2!} - \frac{x^4}{4!}\right) - \left(\frac{x^6}{6!} - \frac{x^8}{8!}\right) - \ . \ . \ .$$

For $x^2 \leq 12$, each of the terms in parentheses is positive so

$$r(x) \leq 1.$$

On the other hand

$$r(x) = \left(1 - \frac{x^2}{2!}\right) + \left(\frac{x^4}{4!} - \frac{x^6}{6!}\right) + \ . \ . \ .$$

implies

$$r(x) \geq 1 - \frac{x^2}{2!}$$

since each of the terms is parentheses is positive for $x^2 \leq 2$. Further, $r(x) \geq -1$ for $x^2 \leq 2$, so $|r(x)| \leq 1$ for $|x| \leq \eta = 2$.

Note that we must choose m to satisfy (from Theorem 4):

$$\left|\frac{\rho(A)}{2^m}\right| \leq \eta \quad \text{and} \quad \left|\frac{\rho(A)}{2^m}\right| \leq \delta.$$

There $\nu = 2k + 2$ while

$$C_\nu = \frac{1}{(2k+2)!}.$$

Thus

$$\|\cos A - U_m\| \leq \frac{1}{(2k+2)!} 2^{m(-2k)} [\rho(A)]^{2k+2}$$

where C_ν decreases as k increases or ν increases.

Next let us consider the Padé approximant

$$R_{mn}(A) = \frac{I + \alpha_1 A^2 + \ldots + \alpha_m A^{2m}}{I + \beta_1 A^2 + \ldots + \beta_m A^{2m}}.$$

For the Padé approximant we determine the coefficients $\{\alpha_i\}$ and $\{\beta_i\}$ in

$$R_{mn}(z) = \frac{\alpha_0 + \alpha_1 z^2 + \ldots + \alpha_m z^{2m}}{\beta_0 + \beta_1 z^2 + \ldots + \beta_m z^{2m}}$$

so that

$$E(z) = \left[\sum_{k=0}^{\infty} \frac{(-z)^k}{(2k)!}\right] Q_{mn}(z) - P_{mn}(z) = 0(z^\nu)$$

for some ν. Actually we are looking at the product

$$E(z) = \frac{1}{Q(z)} [Q(z) \cos z - P(z)] = 0(z^\nu).$$

Since the term $\frac{1}{Q(z)}$ is of order $0(1)$ for small z, it suffices to consider $[Q(z) \cos z - P(z)]$. Consider, for example, the case $m = n = 2$. Expanding the left side of $[Q(z) \cos z - P(z)]$ gives

$$E(z) = (1 - \alpha_0) + (\beta_0 - \alpha_1 - \frac{1}{2})z^2 + (\beta_2 - \frac{\beta_1}{2} - \alpha_2 - \frac{1}{24})z^4$$

$$+ (\frac{\beta_1}{24} - \frac{\beta_2}{2} - \frac{1}{720})z^6 + (\frac{\beta_2}{24} - \frac{\beta_1}{720} + \frac{1}{40320})z^8 + 0(z^{10}).$$

We can see then that with appropriate choices of $\{\alpha_i\}$ and $\{\beta_i\}$ that the truncation error may be $0(z^{10})$. However we may also choose $\{\alpha_i\}$ and $\{\beta_i\}$ such that ν may not be optimal but these methods may have features which make them attractive computationally even though some accuracy is given up.

Lastly, for the rational approximation of the form

$$R(A) = \frac{\sum_{j=0}^{\alpha} \phi_j(\beta)A^{2j}}{(I + \beta^2 A^2)^\alpha}$$

we can find α satisfying the conditions in Theroem 4. The results we needed are developed in [2] and listed below.

PROPOSITION 2.1: Let $\alpha \geq 1$ be an integer. Then there exists a set of polynomials $\{\phi_n^{(\alpha)}\}_{n \geq 0}$ of the real variable x, with real coefficients given by

$$\phi_n^{(\alpha)}(x) = \sum_{j=0}^{n} \frac{(-1)^j}{(2j)!} \binom{\alpha}{n-j} x^{2(n-j)}$$

such that for $x > 0$

$$(1 + x^2z^2)^\alpha \cos z = \sum_{n=0}^{\infty} \phi_n^{(\alpha)}(x)z^{2n}, \quad \text{all} \quad z \in C.$$

If we define the rational function

$$r_\alpha(x;z) = \left(\sum_{n=0}^{\alpha} \phi_n^{(\alpha)}(x) z^{2n}\right) / (1 + x^2 z^2)^\alpha \quad \text{for} \quad |\mathrm{Im} z| < x^{-1}$$

then there exists a constant $C = C(\alpha,x)$ such that

$$|r_\alpha(x;z) - \cos z| \leq C|z|^{2\alpha+2}, \quad \text{for} \quad |z| < x^{-1}.$$

Furthermore, there exists $x^{(\alpha)} > 0$ such that for all $x \geq x^{(\alpha)}$

$$|r_\alpha(x;\tau)| \leq 1, \quad \text{all} \quad \tau \geq 0. \qquad \blacksquare$$

For our purposes, then $\nu = 2\alpha + 2$.

CHAPTER III

DEVELOPMENT OF METHODS

Thus far we have examined approximation techniques in general. Now we would like to describe in more detail those methods used in this thesis to approximate the cosine of a matrix A. For each algorithm, we generate the matrix A in such a way to allow for comparison of our approximation to the actual cosine of A.

First, suppose we want to generate a normal matrix A. We generate an orthogonal matrix of the form

$$P = I - 2ww^T \tag{3.1}$$

where w is a vector

$$\|w\| = 1 = (w^T w)^{1/2}.$$

Recall that P orthogonal means

$$P^T = P \quad \text{or} \quad PP^T = I.$$

To verify that such a P is orthogonal simply note that

$$\begin{aligned} PP^T &= (I - 2ww^T)(I - 2ww^T)^T \\ &= (I - 2ww^T)^2 \\ &= I - 4ww^T + 4w(w^T w)w^T = I. \end{aligned}$$

Obtaining P in such a manner allows us to find a normal matrix A by the similarity transformation

$$P\Lambda P^T = A$$

where $\Lambda = (\lambda_1, \lambda_2, \ldots, \lambda_n)$ is a real diagonal matrix.

In fact, we see that

$$A^T = P\Lambda^T P^T = P\Lambda P^T = A,$$

so not only is A normal, but it is also symmetric.

For any similarity transformation $A = P\Lambda P^{-1}$ we have

$$\phi(A) = P\phi(\Lambda)P^{-1} \tag{3.2}$$

where ϕ is an analytic function defined on the spectrum of A [7]. To compute $\phi(\Lambda)$ we only have to take the functional values of the diagonal elements so (3.2) is easily computed. The matrix $P \cos \Lambda P^{-1}$ is the one compared to our approximation.

There are also some examples included in which A is generated by a similarity transformation but the matrix A is not normal. We use a matrix S and its known inverse to generate

$$A = S\Lambda S^{-1}.$$

The specific S and S^{-1} are:

$$S = \begin{bmatrix} N & N-1 & N-2 & \cdots & 2 & 1 \\ N-1 & N-1 & N-2 & \cdots & 2 & 1 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 2 & 2 & 2 & \cdots & 2 & 1 \\ 1 & 1 & 1 & \cdots & 1 & 1 \end{bmatrix}$$

$$S^{-1} = \begin{bmatrix} 1 & -1 & & & \\ -1 & 2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & -1 \\ & & & -1 & 2 \end{bmatrix}$$

where N is the order of the matrix [3].

In each case the diagonal matrix Λ is read in along with the vector w or the value of N as the case may be.

With our calculations we compare the approximation to the actual cosine in the infinity norm which is defined below:

$$\|A\|_\infty = \max_{\|x\|_\infty=1} \|Ax\|_\infty = \max_{1 \le i \le N} \sum_{j=1}^{N} |a_{ij}|$$

where N is the order of the matrix A [6].

Method I

For the first approximation technique we take a given N-by-N matrix A, let m be an appropriate power of 2 to form the matrix A/m. Then the matrix cosine A/m is calculated by a Taylor approximation. By repeated use of the double angle formula

$$\cos 2A = \cos^2 A - I$$

we find our approximation to cos A.

An important step is the calculation of the matrix $\cos A/m$ by a Taylor approximation. There are two ideas which parallel methods used in the computation of $e^{A/m}$ by a Taylor series, that may be applied to the calculation of cosine A/m. One way is to fix the order of the Taylor approximating polynomial. Fix a positive integer q and approximate cosine A/m by

$$I - \frac{(\frac{A}{m})^2}{2!} + \frac{(\frac{A}{m})^4}{4!} - \cdots + \frac{(\frac{A}{m})^{2q}}{(2q)!} . \tag{3.3}$$

In the second way we take

$$T_k(\frac{A}{m}) = \sum_{\ell=0}^{k} \frac{(\frac{A}{m})^{2\ell}}{(2\ell)!} \tag{3.4}$$

and denote by $f\ell[T_k(\frac{A}{m})]$ the matrix obtained after computing $T_k(\frac{A}{m})$ in floating point arithmetic on the computer. Performing the approximation (3.4) for increasing integral values of k until

$$\|f\ell[T_k(\frac{A}{m})] - f\ell[T_{k+1}(\frac{A}{m})]\|_\infty < \varepsilon$$

for some small ε, insures that the order of the Taylor approximation yields accurate results [6]. However, to obtain this accuracy, the number of operations may be increased greatly since the order of the approximating polynomial may be large. In this way, a valid comparison could not be made with the other methods.

Let us mention some programming aspects of this method. Given the N-by-N matrix A we form A/m with $m = 2^j$ for some given integer j. The matrix cosine A/m is formed using the Taylor series with predetermined value of q, as in (3.3). Then the approximation cosine A is found by repeated use of the double angle formula. It is compared with the actual cosine in the infinity-norm and accepted as our approximation if close enough. Otherwise, the value of j is increased and the program is repeated.

One measure of the efficiency of such a computation would be to count the number of arithmetic operations necessary for the calculation. We are concerned normally only with the number of multiplications and divisions used. Let N be the order of the matrix, j be the power of 2 being used and p the order of the approximating polynomial.

To form the matrix $\frac{A}{m}$ requires N^2 multiplications. A Taylor approximation of order p, using nested multiplication, requires $(N^3 + N^2)p$ multiplications. Then the double angle formula requires jN^3 operations. So the total operation count for Method I is approximately

$$(p + j)N^3 + pN^2.$$

Note the important role of p.

Method II

In the second method we employ rational approximations to the cosine of a given matrix A. As mentioned earlier, we shall use the Padé approach to determine the coefficients $\{\alpha_i\}$ and $\{\beta_i\}$ in the approximation

$$R_{mn}(z) = \frac{\alpha_0 + \alpha_1 z^2 + \ldots + \alpha_m z^{2m}}{\beta_0 + \beta_1 z^2 + \ldots + \beta_n z^{2n}} \equiv \frac{P_{mn}(z)}{Q_{mn}(z)} \tag{3.3}$$

so that the error is

$$(3.4) \qquad E \equiv \left[\sum_{k=0}^{\infty} \frac{(-1)^k}{(2k)!} z^{2k}\right] Q_{mn}(z) - P_{mn}(z) = O(z^{\nu})$$

for some ν. Here we shall consider only the cases $m,n \leq 2$. We then formally replace the scalar z by the matrix A to obtain the approximation $R_{mn}(A) = Q_{mn}^{-1}(A)P_{mn}(A)$.

Expanding the left side of (3.3) yields, as before

$$(3.5) \qquad E = (\beta_0 - \alpha_0) + (\beta_1 - \alpha_1 - \frac{1}{2})z^2 + (\beta_2 - \frac{\beta_1}{2} - \alpha_2 + \frac{1}{24})z^4$$
$$+ (\frac{\beta_1}{24} - \frac{\beta_2}{2} - \frac{1}{720})z^6 + (\frac{\beta_2}{24} - \frac{\beta_1}{720} + \frac{1}{40320})z^8 + O(z^{10}).$$

We begin by normalizing so that $\beta_0 = 1$. Then leaving β_1 and β_2 arbitrary we specify that

$$\alpha_0 = 1 \qquad \alpha_1 = \beta_1 - 1/2.$$

This dictates that ν is at least 2 in (3.4).

Specifically, for the first method we have

$$(3.6) \qquad \beta_1 = \frac{1}{12}, \quad \beta_2 = 0, \quad \alpha_1 = -\frac{5}{12}, \quad \alpha_2 = 0.$$

With these choices we also eliminate the z^4 term in (3.4). This is the optimal (1,1) Padé approximation, also known as the Stormer-Numerov method.

Proceeding to the general (2,2) approximation we specify that

$$\alpha_2 = \beta_2 - \frac{\beta_1}{2} + \frac{1}{24}$$

so that $\nu \geq 6$ for all β_1 and β_2. Our second method requires that

$$(3.7) \qquad \beta_1 = \frac{1}{4}, \quad \beta_2 = 0, \quad \alpha_1 = -\frac{1}{4}, \quad \alpha_2 = -\frac{1}{12}$$

The stability behavior of certain rational approximants of the Padé type have been analyzed in [8].

A brief description of the program used for Method II follows. We generate the N-by-N matrix A and for a given positive integer j we form the matrix A/m, $m = 2^j$. Afterwards we form the numerator P and the denominator Q of the approximation R. Then using Crout reduction with pivoting [9], we compute an LU decomposition of Q. To find R we solve the system

$$LUr_j = p_j$$

where

$$R = [r_1, r_2, \ldots, r_N]$$

by solving in respective order the system

$$Lz_j = p_j$$

$$Ur_j = z_j.$$

We thereby produce the approximation R column by column.

As in Method I, the double angle formula is applied j times to the matrix R to obtain our approximation to cosine A. We compare this approximation to the actual cosine A in the infinity norm, and if our approximation is not close enough we increase the value of j and repeat the program.

As a measure of efficiency we offer an operation count of Method II. In the first case (3.6), the calculation of P and Q require

$$2N^2 + N^3$$

operations. For Crout reduction, that is to obtain an LU decomposition of Q, it takes approximately $N^3/3$ multiplications. To backsolve requires N^2 operations for each column so we have in addition $2N^3$ multiplications. Applying the double angle formula to R takes jN^3 operations. Thus, for a given j, the approximate operation count is

$$2N^2 + (\frac{10}{3} + j)N^3.$$

In second case (3.7), the calculation of P and Q requires $2N^3 + 3N^2$ multiplications. Therefore the operation count is approximately

$$3N^2 + (\frac{13}{3} + j)N^3$$

when $\{\alpha_i\}$ and $\{\beta_i\}$ satisfy (3.7).

Method III

The third approximation scheme also uses a form of rational approximation. In general, the idea is to approximate cosine A by

(3.8)
$$R(A) = \frac{\sum_{j=0}^{\alpha} \phi_j(\beta)A^{2j}}{(I + \beta A^2)^\alpha} \equiv \frac{P}{Q}$$

for some α.

Specifically we choose $\alpha = 2$ and try the following schemes

(3.9)
$$P = I + (2\beta - \frac{1}{2})A^2 + (\beta^2 - \beta + \frac{1}{24})A^4$$

(3.10)
$$Q = (I + \beta A^2)^2$$

with

$$\beta = 1/4 + \sqrt{1/24} \tag{3.11}$$

or

$$\beta = (5 + \sqrt{15})/60. \tag{3.12}$$

This choice of parameter is discussed in [2]. The value of β in (3.11) gives an unconditionally stable fourth order approximation, while that in (3.12) gives a conditionally stable sixth order approximation.

An outline for Method III is given by the following. We form the matrix A/m for some appropriate m, as a power of 2 and then form the numerator P as in (3.9) replacing A with A/m. Next we form the matrix $I + \beta(A/m)^2$. Again using Crout reduction with pivoting we find an LU decomposition for $[I + \beta(\frac{A}{m})^2]$. Then we have the system

$$[I + \beta(\frac{A}{m})^2][I + \beta(\frac{A}{m})^2]R = P$$

to solve. We do this by solving each of the following linear systems

$$[I + \beta(\frac{A}{m})^2]Y = P$$

and

$$[I + \beta(\frac{A}{m})^2]R = Y$$

as in Method II. After obtaining $R(\frac{A}{m})$, repeated use of the double angle formula gives us our approximation to the cosine of A.

Once more let us use the number of multiplications as a measure of efficiency. To form P and Q requires approximately $2N^3 + 3N^2$ operations. The LU decomposition takes $\frac{N^3}{3}$ multiplications. Solving the linear systems of equations requires $4N^3$ operations. Lastly, use of the double angle formula takes jN^3 multiplications for a given j

where $m = 2^j$. Therefore the operation count is approximately

$$3N^2 + (\frac{13}{3} + j)N^3$$

for Method III.

Parlett's Method

Lastly, we use a method developed by B. N. Parlett for the computation of functions of triangular matrices [7] in calculating an approximation for the cosine matrix. In our case we will assume that we can find that Schur decomposition of a matrix , i.e.,

$$B = PTP^*, \qquad P^* = P^{-1}$$

and then employ Parlett's method to find $\cos T$. We find $\cos B$ using the definition $\cos B = P \cos T P^*$.

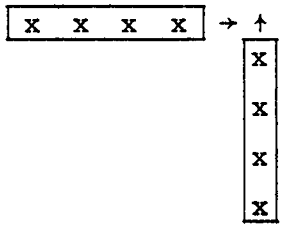
Let ϕ represent an analytic function on the spectrum of B. A simple relation exists among the elements of $\phi(T)$ which allows us to build up $\phi(T)$, one super-diagonal after another, going away from the main diagonal. A comparable relation exists for block upper triangular matrices. In our discussion we do not deal with matrices having multiple eigenvalues, but the confluent case is included by Parlett.

There are two important lemmas used in the development of Parlett's algorithm.

LEMMA 4 (Parlett): Let $\phi(z)$ be an analytic function which is regular inside and on some contour Γ which contains T's spectrum where T is an upper triangular matrix. Then $f_{rs} = \phi(T)_{rs}$ is completely determined by ϕ and those elements t_{ij} of T satisfying $r \leq i \leq j \leq s$. [7]

The matrix $F = \phi(T)$ is well defined and also upper triangular. Any element f_{rs}, $r < s$ of F is related to the elements in its row and column which are closer to the main diagonal and which have already been computed. The following diagram illustrates this relationship [7]:

row r



column S

The result is a consequence of the following lemma:

LEMMA 5 (Parlett): Any square matrix B commutes with $\phi(B)$ for any scalar function ϕ for which $\phi(B)$ is well defined. [7].

From these lemmas Parlett proves this theorem which gives us a formula for computing $F = \phi(T)$:

THEOREM (Parlett): Let $F = \phi(T)$ be an analytic function of an upper triangular matrix T. For $r < s$,

$$(t_{rr} - t_{ss})f_{rs} = (f_{rr} - f_{rs})t_{rs} + \sum_{k=0}^{s-r-1} (f_{r,r+k}t_{r+k,s} - t_{r,s-k}f_{s-k,s}). \tag{3.13}$$

The main diagonal elements are found by

$$f_{rr} = \phi(t_{rr}).$$

For $r > s$

$$F_{rs} = 0. \quad [7] \quad \blacksquare$$

The idea of the proof given by Parlett is that F is upper triangular by lemma 4 and

$$FT - TF = 0$$

by lemma 5. If we write out the (r,s) element of the left hand side and rearrange terms, then (3.13) is obtained.

If T is complex then we have an alternative so that real functions of real matrices with complex eigenvalues may be evaluated in real arithmetic. If the Schur decomposition of B yields a complex matrix T we have a modification of T called the real Schur form $\hat{T}$ which is block upper triangular, and the diagonal blocks are either 1×1 or 2×2. Each complex conjugate pair of eigenvalues λ and $\overline{\lambda}$ in S has a corresponding real 2×2 diagonal block in $\hat{T}$ whose eigenvalues are λ and $\overline{\lambda}$. It may be convenient to standardize the real Schur form by requiring the 2×2 diagonal blocks to have the form

$$\begin{bmatrix} \rho & \beta \\ \gamma & \rho \end{bmatrix}, \quad \gamma > 0, \quad \beta > 0, \quad -\beta\gamma = \mu^2$$

where $\gamma = \rho \pm i\mu$, $\mu > 0$. Generally, we cannot arrange that $\gamma = -\beta = \mu$.

When $\hat{T}$ is block upper triangular we have the following corollary to compute cosine of $\hat{T}$.

COROLLARY: Let $\hat{T} = (\hat{T}_{ij})$ be block upper triangular. Then

$F = \phi(\hat{T})$ will have the same block structure and for $r < s$,

$$(3.4) \qquad \hat{T}_{rr}F_{rs} - F_{rs}\hat{T}_{ss} = \sum_{k=0}^{s-r-1} (F_{r,r+k}\hat{T}_{r+k,s} - \hat{T}_{r,s-k}F_{s-k,s}).$$

PROOF: By lemma 5 we know that

$$TF = FT. \tag{3.15}$$

Writing out the (r,s) block of (3.15) we get

$$T_{rr}F_{rs} + T_{r,r+1}F_{r+1,s} + \cdots T_{r,r+(s-1)}F_{r+(s-1),s} + T_{rs}F_{ss}$$

$$= F_{rr}T_{rs} + F_{r,r+1}T_{r+1,s} + \cdots + F_{r,r+(s-1)}T_{r+(s-1),s} + F_{rs}T_{ss}$$

$$\Rightarrow T_{rr}F_{rs} - F_{rs}T_{ss} = F_{rr}T_{rs} + F_{r,r+1}T_{r+1,s} + \cdots + F_{r,r+s-1}T_{r+s-1,s}$$

$$- T_{r,r+1}F_{r+1,s} - \cdots - T_{r,r+s-1}F_{r+s-1,s}$$

$$- T_{rs}F_{ss}$$

$$\Rightarrow T_{rr}F_{rs} - F_{rs}T_{ss} = \sum_{k=0}^{s-r-1} (F_{r,r+k}T_{r+k,s} - T_{r,s-k}F_{s-k,s}). \qquad \blacksquare$$

Now let us consider in particular the cases:

(3.16) (1) T_{rr} and T_{ss} both 1×1, so F_{rs} is 1×1

(3.17) (2) T_{rr} being 2×2, T_{ss} being 1×1 so that F_{rs} is 2×1

(3.18) (3) T_{rr} and T_{ss} both 2×2 so F_{rs} is 2×2.

The solutions for each of these cases are

CASE 1: $F_{rs} = [\sum_{k=0}^{s-r-1} (F_{r,r+k}T_{r+k,s} - T_{r,s-k}F_{s-k,s})](t_{rr} - t_{ss})^{-1}$

CASE 2: $F_{rs} = [\sum_{k=0}^{s-r-1} (F_{r,r+k}T_{r+k,s} - T_{r,s-k}F_{s-k,s})]\,(T_{rr} - t_{ss}I)^{-1}$

CASE 3: Let

$$T_{rr} = \begin{bmatrix} \rho & \beta \\ \gamma & \rho \end{bmatrix}, \quad T_{ss} = \begin{bmatrix} \mu & \alpha \\ \phi & \mu \end{bmatrix}, \quad F_{rs} = \begin{bmatrix} f_{11} & f_{12} \\ f_{21} & f_{22} \end{bmatrix}$$

We must solve the system

$$T_{rr}F_{rs} - F_{rs}T_{ss} = C = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix}$$

where

$$C = \sum_{k=0}^{s-r-1} (F_{r,r+k}T_{r+k,s} - T_{r,s-k}F_{s-k,s}).$$

We have the linear system

$$(\rho - \mu)f_{11} - \phi f_{12} + \beta f_{21} = c_{11}$$

$$-\alpha f_{11} + (\rho - \mu)f_{12} + \beta f_{22} = c_{12}$$

$$\gamma f_{11} + (\rho - \mu)f_{21} - \phi f_{22} = c_{21}$$

$$\gamma f_{12} - \alpha f_{21} + (\rho - \mu)f_{22} = c_{22}.$$

Using the elimination method we obtain

$$f_{22} = \{\alpha c_{21}[-(\rho - \mu)^2 - \gamma\beta + \alpha\phi] + (\rho - \mu)[\alpha\phi - (\rho - \mu)^2 + \gamma\beta]c_{22}$$

$$+ [(\rho-\mu)^2 - \gamma\beta + \alpha\phi]\gamma c_{12} + 2\alpha(\rho-\mu)\gamma c_{11}\}/\{(\rho-\mu)^2[2\gamma\beta + 2\alpha\phi - (\rho-\mu)^2]$$

$$- \gamma^2\beta^2 - \alpha^2\phi^2 + 2\gamma\beta\alpha\phi\}$$

$$f_{21} = \{(\rho-\mu)c_{21} - \gamma c_{11} - \phi c_{22} + 2\phi(\rho-\mu)f_{22}\} \,/\, [(\rho-\mu)^2 - \gamma\beta + \alpha\phi]$$

$$f_{12} = \{\alpha c_{21} + \gamma c_{12} - \alpha(\rho-\mu)f_{21} + (\gamma\beta + \gamma\phi)f_{22}\} \;/\; \gamma(\rho-\mu)$$

$$f_{11} = (c_{11} + \phi f_{12} - \beta f_{21}) \;/\; (\rho-\mu)f_{11}.$$

To utilize Parlett's method we must compute the real Schur form of a N-by-N matrix A. To obtain the real Schur form requires about $10N^3 + 30N^2$ multiplications using well known routines in the EISPACK library. Then the building up of the cosine matrix takes only $\frac{N^3}{3}$ operations. So the total operation count is about

$$30N^2 + 10N^3 + \frac{N^3}{3}$$

for Parlett's method.

CHAPTER IV

NUMERICAL RESULTS

Each of the methods discussed in Chapter III has been applied to calculate the cosine of various matrices. Data tables have been constructed to show the numerical results. The test matrices will be identified by the order N of the matrix and its eigenvalues. Recall that normality played a significant role in our discussion, but we have included normal as well as non-normal matrices in our tests. In fact the normal matrices are symmetric.

Methods I, II, and III involve a scaling factor m which is a power of 2, e.g. $m = 2^j$ for some positive integer j. The value of j will be important. For method I the number 2p denotes the order of the matrix polynomial used in the Taylor approximation.

DIFF represents the difference in the matrix infinity norm between the actual cosine matrix and the cosine matrix computed by other methods. When DIFF became less than 10^{-10} in double precision or DIFF began to increase the program was ended. Blanks in the table indicate that one of these conditions has occurred.

We wish to examine the results and make some comments on the data. Any observations or conclusions are based solely on the test matrices used. We have considered matrices of order $N = 2, 3, 4, 10$. Note that no matrices having confluent eigenvalues have been included.

Let us look at Tables 1 - 10. Each table contains five values of DIFF, for certain values of j, corresponding to the methods as below:

DIFF 1--METHOD I

DIFF 2--METHOD II $\beta_1 = 1/12$

DIFF 3--METHOD II $\beta_1 = 1/4$

DIFF 4--METHOD III $\beta = 1/4 + \sqrt{1/24}$

DIFF 5--METHOD III $\beta = (5 + \sqrt{15})/60$

The aim here was to demonstrate the effectiveness of the double angle formula with various choices of approximation schemes and strict comparison was not intended.

First of all, for j selected sufficiently large each of the methods provide a reasonably accurate approximation. Secondly, if the number of steps, j, is adequate, we obtain increasingly better results as the order of the Taylor approximation increases. However, this increased accuracy is at the expense of higher cost. Thirdly, for the problems run, the choice of parameter $\beta_1 = 1/12$ in Method II and $\beta = (5 + \sqrt{15})/60$ in Method III afford greater accuracy at the same expense than the other choices given. Lastly, the Taylor methods compare favorably with any of those tested when we have large enough p. However, a fair comparison of these approaches would also involve consideration of cost as well as accuracy.

Observing Table 4 we see that as expected no method gives a reasonable approximation until $\|A\| = \rho(A) \leq 1$. That is, convergence does not begin until $\left|\frac{\rho(A)}{2^j}\right| \leq 1$. Once this condition is satisfied, with $\beta = (5 + \sqrt{15})/60$ in Method III we obtain a better initial

TABLE 1

COMPARISON OF THREE METHODS WITH
N=2 AND POSITIVE EIGENVALUES

N=2 EIGENVALUES 1,2

j	P=2 DIFF 1	β_1=1/12 DIFF 2	β_1=1/4 DIFF 3	$1/4+\sqrt{1/24}$ DIFF 4	$(5+\sqrt{15})/60$ DIFF 5
1	.3500(-2)	.4707(-2)	.1773(-1)	.7381(-1)	.9793(-3)
2	.1943(-3)	.2835(-3)	.1182(-2)	.7362(-2)	.2245(-5)
4	.7324(-6)	.1097(-5)	.4741(-5)	.3461(-4)	.4321(-8)
6	.2855(-8)	.4281(-8)	.1856(-7)	.1369(-6)	.1012(-11)
8	.1121(-10)	.1574(-10)	.6789(-10)	.5336(-9)	
10	.6241(-10)	.1418(-9)	.1078(-9)	.2003(-9)	

j	P=3 DIFF 1	P=6 DIFF 1	P=8 DIFF 1
1	.6304(-4)	.2939(-10)	.4996(-15)
2	.8715(-6)	.6328(-14)	
4	.2051(-9)		
6	.1258(-12)		

TABLE 2

COMPARISON OF THREE METHODS WITH
N=2 AND NEGATIVE EIGENVALUES

N=2 EIGENVALUES -1,-2

j	P=2 DIFF 1	1/12 DIFF 2	1/4 DIFF 3	$1/4+\sqrt{1/24}$ DIFF 4	$(5=\sqrt{15})/60$ DIFF 5
1	.3500(-2)	.4701(-2)	.1773(-1)	.7381(-1)	.9793(-3)
2	.1943(-3)	.2835(-3)	.1182(-2)	.7362(-2)	.1702(-4)
4	.7324(-6)	.1097(-5)	.4741(-5)	.3461(-4)	.4321(-8)
6	.2855(-8)	.4281(-8)	.1855(-7)	.1369(-6)	.1011(-11)
8	.1121(-10)	.1574(-10)	.6789(-10)	.5336(-9)	.1062(-10)
10	.6241(-10)	.1418(-9)	.1078(-9)	.2003(-9)	
j	**P=3 DIFF 1**	**P=6 DIFF 1**	**P=8 DIFF 1**		
1	.6304(-4)	.2939(-10)	.4996(-15)		
2	.8715(-6)	.6328(-14)			
4	.2051(-9)				
6	.1258(-12)				

TABLE 3

COMPARISON OF THREE METHODS WITH
N=2 AND MIXED EIGENVALUES

N=2 EIGENVALUES -1,1

j		1/12	1/4	$1/4+\sqrt{1/24}$	$(5+\sqrt{15})/60$
	DIFF 1	DIFF 2	DIFF 3	DIFF 4	DIFF 5
1	.7584(-4)	.1107(-3)	.1316(-3)	.2869(-2)	.6624(-5)
2	.4608(-5)	.6865(-5)	.2165(-5)	.2079(-3)	.1068(-6)
4	.1783(-7)	.2675(-7)	.1159(-6)	.8531(-6)	.2627(-10)
6	.6955(-10)	.1044(-9)	.4521(-9)	.3342(-8)	.3027(-12)
8	.1426(-11)	.1632(-11)	.7758(-11)	.1695(-10)	
10				.9635(-10)	
12					

j	P=3	P=6	P=8
	DIFF 1	DIFF 1	DIFF 1
1	.3391(-6)	.2591(-14)	.3009(-15)
2	.5145(-8)		
4	.1257(-11)		
6			

TABLE 4

COMPARISON OF THREE METHODS WITH
N=2 AND WIDESPREAD EIGENVALUES

N=2 EIGENVALUES -100,100

j	P=2 DIFF 1	1/12 DIFF 2	1/4 DIFF 3	$1/4+\sqrt{1/24}$ DIFF 4	$(5+\sqrt{15})/60$ DIFF 5
1	.1343(12)	.4757(2)	.1388(7)	.12749	.2763(2)
2	.5197(18)	.4372(4)	.1497(11)	.2089(+1)	.1535(4)
4	.9403(31)	.1811(14)	.1332(23)	.1748(1)	.6840(12)
6	.1060(1)	.2545	.1301(1)	.7562	.1366
8	.1680(-2)	.2461(-2)	.1028(-1)	.7724(-1)	.9206(-4)
10	.6406(-5)	.9598(-5)	.4153(-4)	.3047(-3)	.2303(-7)
12	.2498(-7)	.3748(-7)	.1624(-6)	.1199(-5)	.1639(-11)
14	.6923(-10)	.8211(-10)	.3843(-9)	.4824(-8)	

j	P=3 DIFF 1	P=6 DIFF 2	P=8 DIFF 1
1	.9195(15)	.4679(24)	.8821(28)
4	.5528(30)	.1425(11)	.2312(2)
6	.2648(-1)	.1903(-6)	.1555(-10)
8	.4570(-5)	.4218(-13)	.4354(-13)(j=7)
10	.1091(-8)		
12	.1640(-11)		

TABLE 5

COMPARISON OF THREE METHODS WITH
N=3 AND POSITIVE EIGENVALUES

N=3 EIGENVALUES 1,2,3

j	P=2 DIFF 1	1/12 DIFF 2	1/4 DIFF 3	$1/4+\sqrt{1/24}$ DIFF 4	$(5+\sqrt{15})/60$ DIFF 5
1	.5587(-2)	.5392(-2)	.1837(-1)	.9609(-1)	.2164(-2)
2	.2424(-3)	.3401(-3)	.1341(-2)	.8181(-2)	.4021(-4)
4	.8752(-6)	.1308(-5)	.5639(-5)	.4071(-4)	.1071(-7)
6	.3403(-8)	.5104(-8)	.2211(-7)	.1630(-6)	.2539(-11)
8	.1369(-10)	.2109(-10)	.8656(-10)	.6389(-9)	
10	.5257(-10)	.1413(-9)	.1602(-9)	.1993(-9)	
12				.1721(-8)	

j	P=3 DIFF 1	P=6 DIFF 1	P=8 DIFF 1
1	.1970(-3)	.5156(-5)	.7669(-13)
2	.2285(-5)	.1475(-7)	
4	.5116(-9)	.2064(-12)	
6	.1727(-12)		

TABLE 6

COMPARISON OF THREE METHODS WITH
N=3 AND NEGATIVE EIGENVALUES

N=3 EIGENVALUES -1,-2,-3

j	P=2 DIFF 1	1/12 DIFF 2	1/4 DIFF 3	$1/4+\sqrt{1/24}$ DIFF 4	$(5+\sqrt{15})/60$ DIFF 5
1	.5487(-2)	.5392(-2)	.1837(-1)	.9609(-1)	.2064(-2)
2	.2424(-3)	.3401(-3)	.1341(-2)	.8181(-2)	.4021(-4)
4	.8752(-6)	.1308(-5)	.5639(-5)	.4071(-4)	.1071(-7)
6	.3403(-8)	.5104(-8)	.2211(-7)	.1630(-6)	.2539(-11)
8	.1369(-10)	.2109(-10)	.8656(-10)	.6389(-9)	.1421(-10)
10	.5257(-10)	.1413(-9)	.1602(-10)	.1993(-9)	
12					

j	P=3 DIFF 1	P=6 DIFF 1	P=8 DIFF 1
1	.1970(-3)	.1104(-8)	.7769(-13)
2	.2285(-5)	.1984(-12)	
4	.5116(-9)		
6	.1727(-12)		
8			

TABLE 7

COMPARISON OF THREE METHODS WITH
N=3 AND MIXED EIGENVALUES

N=3 EIGENVALUES 10,-5,10

j	P=2 DIFF 1	1/12 DIFF 2	1/4 DIFF 3	$1/4+\sqrt{1/24}$ DIFF 4	$(5+\sqrt{15})/60$ DIFF 5
1	.4961(3)	.2153(2)	.1460(3)	.1607(1)	.1528(2)
2	.4137	.3329(1)	.2755(2)	.1415(1)	.1547(1)
4	.1421(-2)	.2052(-2)	.8447(-2)	.4473(-1)	.1871(-3)
6	.5260(-5)	.7869(-5)	.3397(-4)	.2463(-3)	.4848(-7)
8	.2048(-7)	.3071(-7)	.1330(-6)	.9812(-6)	.1048(-10)
10	.8021(-10)	.1165(-9)	.5168(-9)	.3838(-8)	.1337(-9)
12	.2651(-9)	.1015(-8)	.7895(-9)	.2123(-9)	

j	P=3 DIFF 1	P=6 DIFF 1	P=8 DIFF 1
1	.1201(3)	.9363(-1)	.7444(-3)
2	.1854	.1776(-4)	.9538(-8)
4	.1002(-4)	.2774(-12)	.4918(-13)(j=3)
6	.2308(-8)		
8	.8183(-12)		

TABLE 8

COMPARISON OF THREE METHODS WITH
N=4 AND NEGATIVE EIGENVALUES

N=4 EIGENVALUES -1,-2,-3,-4

j	P=2 DIFF 1	1/12 DIFF 2	1/4 DIFF 3	$1/4+\sqrt{1/24}$ DIFF 4	$(5+\sqrt{15})/60$ DIFF 5
1	.1241	.1536	.5425	.3322	.77802(-1)
2	.4897(-2)	.6643(-2)	.2536(-1)	.9618(-1)	.1373(-2)
4	.1658(-4)	.2470(-4)	.1060(-3)	.7478(-3)	.3840(-6)
6	.6419(-7)	.9624(-7)	.5375(-9)	.3069(-5)	.9566(-10)
8	.2502(-9)	.3731(-9)	.1629(-8)	.1203(-7)	.4710(-11)
10	.4019(-10)	.3088(-9)	.3117(-10)	.9661(-10)	(j=7)
12		.3376(-8)	.1048(-8)	.1235(-8)	

j	P=3 DIFF 1	P=6 DIFF 1	P=8 DIFF 1
1	.1019(-1)	.3077(-6)	.6745(-10)
2	.8824(-4)	.4110(-10)	.3719(-14)
4	.1851(-7)	.8893(-14)(j=3)	
6	.4531(-11)		

TABLE 9

COMPARISON OF THREE METHODS WITH
N=4 AND MIXED EIGENVALUES

N=4 EIGENVALUES -1,1,-2,2

j	P=2 DIFF 1	1/12 DIFF 2	1/4 DIFF 3	$1/4+\sqrt{1/24}$ DIFF 4	$(5+\sqrt{15})/60$ DIFF 5
1	.2952(-2)	.3972(-2)	.1496(-1)	.6247(-1)	.8237(-3)
2	.1639(-3)	.2393(-3)	.9978(-3)	.6218(-2)	.1432(-6)
4	.6181(-6)	.9256(-6)	.4001(-5)	.2921(-4)	.3634(-8)
6	.2409(-8)	.3613(-8)	.1566(-7)	.1155(-6)	.2057(-11)
8	.8497(-11)	.1301(-10)	.6330(-10)	.4571(-9)	.2903(-10)
10	.3883(-10)	.1026(-10)	.3858(-10)	.4575(-10)	
12		.2077(-8)		.1293(-8)	

j	P=3 DIFF 1	P=6 DIFF 1	P=8 DIFF 1
1	.5301(-4)	.2469(-10)	.4441(-15)
2	.7330(-6)	.4552(-14)	
4	.1725(-9)		
6	.9853(-13)		

TABLE 10

COMPARISON OF THREE METHODS WITH
N=10 AND NEGATIVE EIGENVALUES

N=10 -1,-2,-3,-4,-5,-6,-7,-8,-9,-10

j	P=2 DIFF 1	1/12 DIFF 2	1/4 DIFF 3	$1/4+\sqrt{1/24}$ DIFF 4	$(5+\sqrt{15})/60$ DIFF 5
1	.1017(4)	.2562(2)	.1771(3)	.2592(1)	.1829(2)
2	.2881(1)	.4142(1)	.3421(2)	.29516(1)	.1925(1)
4	.8957(-2)	.2583(-2)	.1064(-1)	.5648(-1)	.2341(-3)
6	.3341(-4)	.9906(-5)	.4276(-4)	.3101(-3)	.6063(-7)
8	.1302(-6)	.3866(-7)	.1675(-6)	.1235(-8)	.1780(-10)
10	.1234(-7)	.1397(-9)	.6504(-9)	.4843(-8)	.2899(-9)
12	.1316(-6)	.2831(-8)	.2741(-8)		

j	P=3 DIFF 1	P=6 DIFF 1	P=8 DIFF 1
1	.2562(3)	.3224	.2368(-2)
2	.7218	.5844(-4)	.2939(-7)
4	.5267(-4)	.5441(-11)	.1075(-11)(j=3)
6	.1224(-7)		
8	.7213(-9)		

approximation. Observe that we are looking for the lowest value of j for which we have a reasonable approximation.

As N increases the initial approximation becomes worse, and a higher value of j is required to obtain a more accurate approximation, and therefore more operations are required. Note that the effect of round-off error becomes significant as j increases. We will see this demonstrated in Table 11. With larger N we still get more accurate results in Method I by increasing P.

Returning to the problem of round-off error, we see that even though desired accuracy may not have been obtained, DIFF begins to grow as j continues to increase. Table 11 contains both DIFF and $\|R(A/2^j) - \cos (A/2^j)\|$, where $R(A/2^j)$ is the rational approximation before applying the double angle method, for certain values of j. Notice that although $\|R(A/2^j - \cos (A/2^j)\|$ improves as j gets larger, the value of DIFF becomes worse. We therefore conclude that the problem lies in repeated use of the double angle method required by using larger j. Round-off error takes over at some point and DIFF starts to increase. This provides another reason for wanting to find a reasonably accurate approximation with j as small as possible.

Now let us concentrate on the data involving the Taylor approximation, Method I. We would like to compare the results for different values of j and p. Being more specific, we ask if it would be more practical to use a higher order polynomial and small j or a larger j and a smaller order polynomial. Recall that the key to the operation count for this method is $(j + p)$. Tables 12 and 13 show DIFF for

TABLE 11

ILLUSTRATION OF ROUND-OFF ERROR
IN DOUBLE ANGLE METHOD

N=4 EIGENVALUES -1,-2,-3,-4

$\beta = \frac{1}{12}$	j	$\|R(\frac{A}{2^j}) - \cos(\frac{A}{2^j})\|$	DIFF 2
	1	.8385(-1)	.1536
	2	.1841(-2)	.6643(-2)
	4	.5046(-6)	.2470(-4)
	6	.1241(-9)	.9625(-7)
	8	.3005(-13)	.3731(-9)
	10	.4565(-15)	.3088(-9)
	12	.3608(-15)	.3376(-8)
	14	.2634(-15)	.3848(-7)
	16	.2363(-15)	.7783(-6)
	18	.2360(-15)	.1282(-4)
	20	.2360(-15)	.2054(-3)

TABLE 12

RESULTS OF METHOD I FOR VALUES
OF j+p WITH N=2

N=2 EIGENVALUES 1,2

j+p	j	p	DIFF 1
4	2	2	.1943(-3)
4	1	3	.6304(-4)
5	1	4	.7035(-6)
5	3	2	.4570(-7)
5	2	3	.8715(-6)
7	1	6	.2939(-10)
7	4	3	.2051(-9)
7	5	2	.4569(-7)
9	1	8	.4996(-15)
9	7	2	.1779(-9)
9	6	3	.1258(-12)

TABLE 13

RESULTS OF METHOD I FOR VALUES
OF j+p WITH N=4

N=4 EIGENVALUES -1,-2,-3,-4

j+p	j	p	DIFF 1
4	2	2	.4897(-2)
4	1	3	.1019(-1)
5	1	4	.4557(-3)
5	3	2	.2728(-3)
5	2	3	.8824(-4)
7	1	6	.3077(-6)
7	4	3	.1851(-7)
7	5	2	.1029(-5)
9	1	8	.6745(-11)
9	7	2	.4010(-8)
9	6	3	.4531(-11)

matrices of order $N = 2$ and 4 using different values of j and p to obtain a given value of $j + p$.

Based on the data we have, we cannot determine which case provides the greater accuracy. However, considering the operation count it would be less expensive to choose a small p and larger j rather than taking a large p and a small j, until round-off becomes a factor. These conclusions are based only on our tests.

Let us focus our attention on comparison of Parlett's method to the other methods. For an upper triangular or block upper triangular matrix, Parlett's method offers an accurate means to calculate the cosine. In order to compare Parlett's method with the other methods, we have computed the cosine of an upper triangular matrix T using Parlett's method, then multiplied $\phi(T)$ by P and P* to obtain the matrix

$$P\phi(T)P^*$$

where $P = I - 2ww^T$. We compared this to $\cos(PTP^*)$ computed by Methods I and III. Our purpose here is not to see whether Parlett's method works, but rather to see how well our double angle approach agrees with that obtained by his method. DIFF represents the difference in the infinity norm between the approximation R and the cosine found by using Parlett's method. In particular, the upper triangular matrices of order $N = 2, 3, 4$ and 10 are

$$\begin{bmatrix} 1 & 1 \\ 0 & 2 \end{bmatrix}, \quad \begin{bmatrix} 5 & 3 & -1 \\ 0 & -50 & -2 \\ 0 & 0 & 100 \end{bmatrix}, \quad \begin{bmatrix} -4 & 1 & 5 & 10 \\ 0 & -3 & 1 & -6 \\ 0 & 0 & -2 & 3 \\ 0 & 0 & 0 & -1 \end{bmatrix}$$

$$\begin{bmatrix} 1 & -1 & 0 & 1 & -1 & 0 & 1 & -1 & 1 & 0 \\ 0 & 2 & 0 & 1 & -1 & 0 & 5 & 0 & 0 & -1 \\ 0 & 0 & 3 & -1 & -2 & 0 & 3 & 0 & 0 & 1 \\ 0 & 0 & 0 & 4 & -1 & 0 & 5 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & -5 & 0 & 1 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 6 & 3 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 7 & -1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 8 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 9 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 10 \end{bmatrix}.$$

Observing Table 14, for small j and p not too large in Method I, Method III affords the greatest agreement with the Parlett computation. Further, as j gets large, all methods give about the same agreement and in particular the order of Taylor approximation in Method I is not important.

Recall that the error bounds developed in Chapter II relied upon the matrix A being normal, with real eigenvalues. As stated earlier, this does not imply that these error bounds are incorrect for other matrices, but another technique of proof is needed to verify this. In Tables 15-18 we have included examples of non-normal matrices.

TABLE 14

COMPARISON OF PARLETT'S METHOD TO
METHOD I AND METHOD III

	j	P=2 DIFF 1	P=3 DIFF 1	P=6 DIFF 1	P=8 DIFF 1	$1/4+\sqrt{1/24}$ DIFF 3	$(5+\sqrt{15})/60$ DIFF 3
N=2	2	.2496(-3)	.1129(-5)	.5736(-8)	.5736(-8)	.9449(-2)	.2200(-6)
	4	.9364(-6)	.5590(-8)	.5736(-8)	.5736(-8)	.4445(-4)	.4510(-7)
	6	.7798(-8)	.5735(-8)	.5736(-8)	.5736(-8)	.1714(-6)	.4445(-7)
	8	.5741(-8)	.5734(-8)	.5734(-8)	.5734(-8)	.6110(-8)	.4221(-7)
	10	.5720(-8)	.5720(-8)	.5735(-8)	.5720(-8)	.5733(-8)	.4220(-7)
N=3	2	.6105(18)	.1024(24)	.1002(34)	.3607(37)	.5999	.1864(4)
	4	.1105(32)	.6494(30)	.1054	.2716(2)	.2421(1)	.8036(12)
	6	.1244(1)	.3109(-1)	.2205(-6)	.5814(-8)	.8860	.1062
	8	.1968(-2)	.5367(-5)	.5815(-8)	.5815(-8)	.1113(1)	.1114(1)
	10	.7501(-5)	.5995(-8)	.5797(-8)	.5797(-8)	.3569(-3)	.3008(-7)
N=4	2	.1670(-1)	.3060(-3)	.7405(-7)	.7392(-7)	.3263	.4772(-2)
	4	.5636(-4)	.7404(-7)	.7392(-7)	.7392(-7)	.2538(-2)	.1323(-5)
	6	.2779(-6)	.7390(-7)	.7391(-7)	.7391(-7)	.1044(-4)	.7358(-7)
	8	.7383(-7)	.7388(-7)	.7388(-7)	.7388(-7)	.1121(-6)	.7383(-7)
	10	.7382(-7)	.7380(-7)	.7380(-7)	.7380(-7)	.7348(-7)	.7195(-7)
N=10	2	.1244(1)	.2537	.2310(-4)	.1290(-6)	.3899(1)	
	4	.2805(-2)	.1575(-4)	.1318(-6)	.1318(-6)	.1049	
	6	.1058(-4)	.1335(-6)	.1318(-6)	.1318(-6)	.4994(-3)	
	8	.1276(-6)	.1318(-6)	.1318(-6)	.1318(-6)	.1980(-5)	
	10	.1317(-6)	.1318(-6)	.1318(-6)	.1318(-6)	.1277(-6)	

TABLE 15

RESULTS OF THREE METHODS ON NON-NORMAL MATRIX OF ORDER N=2

N=2 EIGENVALUES 3,2

j	P=2 DIFF 1	1/12 DIFF 2	1/4 DIFF 3	$1/4+\sqrt{1/24}$ DIFF 4	$(5+\sqrt{15})/60$ DIFF 5
1	.1019(-1)	.5955(-2)	.2528(-1)	.1388	.4911(-2)
2	.3231(-3)	.4185(-3)	.1445(-2)	.8335(-2)	.1017(-3)
4	.1068(-5)	.1590(-5)	.6813(-5)	.4794(-4)	.2751(-7)
6	.4133(-8)	.6194(-8)	.2682(-7)	.1974(-6)	.5297(-10)
8	.2571(-10)	.3087(-10)	.8555(-10)	.7990(-9)	

j	P=3 DIFF 1	P=6 DIFF 1	P=8 DIFF 1
1	.5396(-3)	.3681(-8)	.2602(-12)
2	.5974(-5)	.6652(-12)	
4	.1318(-8)		
6	.2709(-11)		

TABLE 16

RESULTS OF THREE METHODS ON NON-NORMAL MATRIX OF ORDER N=3

N=3 EIGENVALUES -5,10,20

j	P=2 DIFF 1	1/12 DIFF 2	1/4 DIFF 3	$1/4+\sqrt{1/24}$ DIFF 4	$(5+\sqrt{15})/60$ DIFF 5
1	.8102(6)	.1027(3)	.6273(4)	.6396(1)	.7267(2)
4	.2428	.3235	.1197(1)	.2059(1)	.9085(-1)
6	.7713(-3)	.1145(-2)	.4892(-2)	.3368(-1)	.2688(-4)
8	.2974(-5)	.4458(-5)	.1930(-4)	.1419(-3)	.6830(-8)
10	.1218(-7)	.1819(-7)	.7531(-7)	.5570(-6)	.1962(-8)
11	.2845(-8)	.1570(-8)	.4869(-8)	.3301(-7)	.1187(-7)

j	P=3 DIFF 1	P=6 DIFF 1	P=8 DIFF 1
1	.6257(7)	.3752(7)	.8894(5)
4	.4226	.1197(-7)	.4696(-12)
6	.1300(-5)	.4374(-11)	
8	.3158(-9)		

TABLE 17

RESULTS OF THREE METHODS ON NON-NORMAL MATRIX OF ORDER N=4

N=4 EIGENVALUES -1,-2,-3,-4

j	P=2 DIFF 1	1/12 DIFF 2	1/4 DIFF 3	$1/4+\sqrt{1/24}$ DIFF 4	$(5+\sqrt{15})/60$ DIFF 5
1	.4087	.5092	.1754(1)	.1680(1)	.2457
2	.1646(-1)	.2246(-1)	.8632(-1)	.3516	.4350(-2)
4	.5626(-4)	.8385(-4)	.3601(-3)	.2549(-2)	.1214(-5)
6	.2180(-6)	.3268(-6)	.1415(-5)	.1042(-4)	.3029(-9)
8	.8996(-9)	.1272(-8)	.5510(-8)	.4094(-7)	.1859(-9)
10	.1361(-8)	.1302(-8)	.2294(-8)	.2193(-8)	.4049(-8)

j	P=3 DIFF 1	P=6 DIFF 1	P=8 DIFF 1
1	.3715(-1)	.9305(-6)	.2029(-9)
2	.2782(-3)	.1246(-9)	.4274(-13)
4	.5851(-7)	.2038(-12)*3	
6	.2096(-10)		

TABLE 18

RESULTS OF THREE METHODS ON NON-NORMAL
MATRIX OF ORDER N=10

N=10 EIGENVALUES -1,1,-2,2,-3,3,-4,4,-5,5

j	P=2 DIFF 1	1/12 DIFF 2	1/4 DIFF 3	$1/4+\sqrt{1/24}$ DIFF 4	$(5+\sqrt{15})/60$ DIFF 5
1	.4556(1)	.5435(1)	.1835(2)	.8810(1)	.3239(1)
2	.1292	.1657	.5932	.2091(1)	.4522(-1)
4	.4068(-3)	.6043(-3)	.2585(-2)	.1790(-1)	.1329(-4)
6	.1569(-5)	.2353(-5)	.1019(-4)	.7490(-4)	.3427(-8)
8	.6422(-8)	.1063(-7)	.4028(-7)	.2944(-6)	.2916(-8)
10	.2368(-7)	.3972(-7)	.3815(-7)	.3884(-7)	.5382(-7)

j	P=3 DIFF 1	P=6 DIFF 1	P=8 DIFF 1
1	.6235	.6797(-4)	.3610(-7)
2	.3337(-2)	.5405(-8)	.5154(-12)
4	.6470(-6)	.1313(-10)	
6	.2900(-9)		

The matrices are generated as explained in Chapter III and compared to the actual cosine. Test matrices of order $N = 2, 3, 4$, and 10 are considered.

Based on this data we obtain reasonably accurate results for non-normal matrices. Again Method I with a high value of p proves to be more accurate. As N increases a larger value of j is required for convergence. Method III gives a better initial approximation than Method II. Keeping in mind that we are looking for the lowest value of j for which we have a reasonable approximation, Method III seems to be the more efficient.

Although we have not proven any results for matrices which are not normal with real eigenvalues, it seems that the same observations may be made in Tables 15-18 as in Tables 1-10. These comments are based entirely on the matrices that were used; perhaps further investigation computationally and theoretically might address these questions.

We might also try to compare the different methods by fixing a desired accuracy and then trying to determine which scheme obtains this order with least work. Here, the role of the rational type approximations might become more important for problems with widely separated eigenvalues. Our investigation demonstrated the potential of the various double angle methods for accurate computation of the cosine and has opened the way for further study.

REFERENCES

REFERENCES

1. Apostol, Tom, "Explicit Formulas for Solutions of the Second-Order Matrix Differential Equation Y" = AY," American Mathematical Monthly, Volume 82, No. 2, Feb. 1975, pp. 159-162.

2. Baker, G. A., Dougalis, V. A., Serbin, S. M., "An Approximation Theorem for Second-Order Evolution Equations," to appear in Numerische Mathematik.

3. Gregory, Robert T., Karney, David L., A Collection of Matrices for Testing Computational Algorithms, John Wiley & Sons Inc., New York, 1969.

4. Lancaster, Peter, Theory of Matrices, Academic Press, Inc., New York, 1969.

5. Moler, C. B. and Van Loan, C. F., "Nineteen Dubious Ways to Compute the Exponential of a Matrix," SIAM Review, Volume 30, No. 4, Oct. 1978, pp. 801-836.

6. Ortega, James M., Numerical Analysis, A Second Course, Academic Press, Inc., New York, 1972.

7. Parlett, B. N., "Computations of Functions of Triangular Matrices," Memo No., ERL-M481, University of California, Berkeley, 1974.

8. Serbin, S. M., "Rational Approximations of Trigonometric Matrices with Application to Second-Order Systems of Differential Equations," Applied Mathematics and Computation, Volume 5, (1979), pp. 75-92.

9. Stewart, G. W., Introduction to Matrix Computations, Academic Press, Inc., New York, 1973.

10. Ward, R. C., "Numerical Computation of the Matrix Exponential with Accuracy Estimate," SIAM Journal of Numerical Analysis 14, (1977), pp. 600-619.

VITA

Sybil Ann Blalock is the daughter of Mr. and Mrs. William A. Blalock. She was born in Sevierville, Tennessee, on January 14, 1954. She attended elementary school in that city and graduated from Sevier County High School in June 1971. The following September she entered Mercer University, Macon, Georgia, and in June 1975, she received a Bachelor of Science degree in Mathematics and Physical Education.

In September 1975, she entered the Graduate School at The University of Tennessee, Knoxville. She was a teaching assistant in the Department of Mathematics during the time of her graduate studies. In December 1979, she received the Master of Science degree with a major in Mathematics.