Consiglio Nazionale delle Ricerche

ISTITUTO DI ELABORAZIONE DELLA INFORMAZIONE

PISA

Analisi di algoritmi per il calcolo della matrice esponenziale.

B. Codenotti C. Fassino

Nota interna B4 - 21

Maggio 1990

ANALISI DI ALGORITMI PER IL CALCOLO DELLA MATRICE ESPONENZIALE.

B. Codenotti [1] – C. Fassino [2]

Sommario. Si esaminano due algoritmi per il calcolo della matrice esponenziale: il metodo si troncamento della serie di Taylor e il metodo scaling and squaring. In particolare, si presentano alcune maggiorazioni dell'errore commesso nel calcolare l'esponenziale di una matrice normale, utilizzando gli algoritmi suddetti. Inoltre, si discutono alcuni risultati relativi all'esponenziale di matrici hermitiane definite positive.

[1] Istituto di Elaborazione dell'Informazione del CNR, Via S. Maria, 46, 56125-Pisa, Italia.

[2] Dipartimento di Informatica, Corso Italia, 40, 56100-Pisa, Italia.

1. Introduzione e notazioni.

La matrice esponenziale di una matrice A nxn è definita da

$$e^A = \sum_{i=0}^{\infty} \frac{A^i}{i!}.$$

Il calcolo di e^A, concettualmente semplice ma complesso dal punto di vista computazionale, è un problema frequente: ad esempio processi fisici, biologici ed economici possono coinvolgere la matrice esponenziale. In particolare, molti modelli matematici richiedono la risoluzione di sistemi di equazioni differenziali ordinarie lineari a coefficienti costanti del tipo

$$\begin{cases} \dot{x}(t) = Ax(t) \\ x(0) = x_0, \end{cases}$$

dove A è una matrice nxn e $x(t), x_0$ sono vettori di ordine n: tali sistemi, come noto, ammettono come unica soluzione la matrice e^{At}.

Per il calcolo della matrice esponenziale, sono stati presentati algoritmi che si basano su tecniche diverse. Alcuni sfruttano la definizione dell'esponenziale matriciale mediante serie [1], altri utilizzano metodi per la risoluzione di sistemi di equazioni differenziali [7]; esistono anche algoritmi che si basano su polinomi matriciali [6] o decomposizione di matrici [4]. Non si conoscono a fondo le caratteristiche numeriche di questi algoritmi: non esiste una formalizzazione dell'argomento che stabilisca a priori, data una qualsiasi classe di matrici, il metodo più efficiente per il calcolo dell'esponenziale o un'accurata maggiorazione dell'errore commesso.

Nel 1978 Moler e Van Loan hanno presentato un articolo ([5]) in cui vengono sottolineate le caratteristiche dei principali metodi per il calcolo della matrice esponenziale. Dalla loro analisi emerge che nessun algoritmo risulta essere totalmente soddisfacente. Inoltre, non è completa la comprensione dei fenomeni numerici; ad esempio, non si conosce a fondo il comportamento del condizionamento della matrice esponenziale, nè le caratteristiche di propagazione dell'errore dei vari algoritmi o le classi di matrici per le quali un certo metodo è efficiente.

Uno degli scopi principali di questo lavoro è di iniziare una trattazione sistematica del problema, per contribuire, in futuro, ad una formalizzazione dell'argomento simile a quella esistente per la risoluzione dei sistemi lineari. Più precisamente vengono esaminati due algoritmi che sfruttano la definizione della matrice esponenziale mediante serie di potenze matriciali; ossia

1) il metodo di troncamento della serie di Taylor;

2) il metodo scaling and squaring.

Il resto del lavoro è organizzato come segue. Nel secondo e terzo paragrafo vengono esaminate alcune caratteristiche della matrice esponenziale; più precisamente il secondo paragrafo tratta il condizionamento della matrice e^A ed il terzo studia il caso particolare in cui l'esponente è una matrice hermitiana e definita positiva. Il quarto e il quinto paragrafo contengono, rispettivamente, lo studio del comportamento del metodo di troncamento della serie di Taylor e del metodo scaling and squaring. Nel paragrafo sei è riportato il confronto tra i due algoritmi, di Taylor e scaling and squaring, e, infine, il paragrafo sette presenta alcuni esempi.

Nel seguito vengono utilizzate le seguenti notazioni.

$A, B, \ldots$ denotano le matrici;

a_{ij} denota l'elemento di posto (i,j) della matrice A;

$a_{ij}^{(k)}$ denota l'elemento di posto (i,j) della matrice A^k;

$\oplus$ è la versione floating point dell'operazione somma;

$\otimes$ è la versione floating point dell'operazione prodotto;

$A^{\textcircled{1}}$ è il valore calcolato in aritmetica floating point della matrice A^1;

$\| \bullet \|$ è la norma 1 matriciale, cioè

$$\|A\| = \max_{1 \le j \le n} \sum_{i=1}^{n} |a_{ij}|$$

$u = 2^{-t}$ denota la precisione di macchina;

$\overset{\bullet}{=}$ è l'eguaglianza a meno di termini dell' ordine di $O(u^2)$ o di ordine superiore;

$\overset{\bullet}{\le}$ è la maggiorazione a meno di termini dell' ordine di $O(u^2)$ o di ordine superiore;

$fl(\bullet)$ è il risultato floating point delle operazioni che compaiono come argomento;

$\alpha(A) = \max\{Re(\lambda) : \lambda \epsilon \Lambda(A)\}$;

$\log x$ è il logaritmo in base 2 di x;

$K(A)$ è il numero di condizionamento di A utilizzando la $\| \bullet \|_1$;

q^* è il numero complesso coniugato di q;

$\lceil x \rceil$ è il minimo intero maggiore o uguale ad x.

2. Sensibilità matrice esponenziale.

Dati una matrice A ad elementi complessi e un numero reale s non negativo lo studio del condizionamento del problema del calcolo della matrice esponenziale e^{As} viene affrontato a partire dalla funzione

$$\Phi(s) = \frac{\|e^{(A+E)s} - e^{As}\|}{\|e^{As}\|},$$

dove $\|\cdot\|$ è una norma matriciale.

In un lavoro di Van Loan [8], sono presentate alcune maggiorazioni della funzione Φ, espressa mediante la norma 2, ottenute sfruttando il seguente lemma.

Lemma 2.1

Sia $M(s)$ una funzione monotona non decrescente su $[0, +\infty)$ tale che $\|e^{As}\| \leq M(s)e^{\beta s}, \quad s \geq 0.$

Allora

$$\Phi(s) \leq \|E\| s M^2(s) e^{[(\beta - \alpha(A) + \|E\| M(s))s]}.$$

Come sottolineato in [8], le maggiorazioni ottenute non sempre sono stime accurate del condizionamento della matrice esponenziale, ma solo una limitazione superiore della funzione Φ. Tuttavia, nel caso in cui la matrice A è normale, tali maggiorazioni assumono il valore minimo; questo fatto suggerisce che il problema del calcolo dell'esponenziale matriciale sia ben condizionato se A è normale.

La conferma di questo risultato si ottiene introducendo il *numero di condizionamento esponenziale* $\nu(A, s)$.

Si definiscono

$$\nu_\delta(A, s) = \sup_{\|E\| \leq \delta \|A\|} \frac{\|e^{(A+E)s} - e^{As}\|}{\delta \|e^{As}\|}$$

e

$$\nu(A, s) = \lim_{\delta \to 0} \nu_\delta(A, s).$$

Geometricamente, si ha che

$$\delta \|e^{As}\| \nu_\delta(A, s)$$

è il raggio della più piccola sfera di centro e^{As} che contiene l'immagine, mediante la funzione esponenziale, dell'insieme

$$\{(A + E) : \|E\| \leq \delta \|A\|\}.$$

È ovvio che , se $\nu_\delta(A,s)$ è piccolo, a piccole perturbazioni della matrice A corrispondono piccole variazioni degli elementi della matrice esponenziale. Quindi $\nu_\delta(A,s)$ e, di conseguenza, $\nu(A,s)$ permettono di valutare il condizionamento della matrice esponenziale.

Inoltre, in [8], si ha che

$$\nu(A,s) \geq s\|A\|$$

e l'uguaglianza vale se e solo se A è normale.

Si può allora concludere che, se la matrice è normale, allora il problema del calcolo dell'esponenziale è ben condizionato.

3. Serie di Taylor.

Restringiamo il problema del calcolo delle matrice esponenziale e^{As} al caso in cui $s = 1$ e, per le conclusioni del paragrafo precedente, con A normale.

L'algoritmo considerato è il troncamento della serie di Taylor. Data la matrice A di ordine n ad elementi complessi, e^A è definita come

$$e^A = \sum_{i=0}^{+\infty} \frac{A^i}{i!}.$$

Troncando la serie di Taylor all'N-esimo termine, si approssima e^A con la matrice

$$T_N(A) = \sum_{i=0}^{N} \frac{A^i}{i!}.$$

3.1 Analisi dell'errore analitico

Sia $\varepsilon_T(N)$ la norma 1 dell'errore analitico assoluto commesso, cioè

$$\varepsilon_T(N) := \|e^A - T_N(A)\|_1.$$

È possibile maggiorare tale errore come segue ([5],[2]):

$$\varepsilon_T(N) \leq \frac{\|A\|^{N+1}}{(N+1)!} e^{\|A\|}, \tag{1}$$

$$\varepsilon_T(N) \leq \frac{\|A\|^{N+1}}{(N+1)!} \frac{1}{1 - \frac{\|A\|}{(N+2)}}, \quad \text{se} \quad N > \|A\| - 2. \tag{2}$$

La seconda maggiorazione è definitivamente migliore della prima, in quanto

$$e^{\|A\|} \geq \frac{1}{1 - \frac{\|A\|}{(N+2)}} \quad \text{se} \quad N \geq \frac{\|A\|}{1 - e^{-\|A\|}} - 2.$$

Le funzioni maggioranti tendono asintoticamente a zero, se $N \to +\infty$, sono definitivamente decrescenti in N e crescenti rispetto a $\|A\|$. Si noti che queste proprietà caratterizzano anche l'errore analitico.

Ci sono due diverse possibilità per la scelta dell'indice di troncamento N, ossia:

i) il valore N è tale che l'errore analitico commesso è trascurabile;

ii) il valore N è tale che l'errore analitico è dello stesso ordine dell'errore algoritmico.

Esaminiamo il primo caso rispetto alla $\|A\|$ e all'ordine di A.

Sia $\overline{N}$ un valore tale che, $\forall \quad N \geq \overline{N}$ l'errore analitico $\epsilon_T(N)$ è trascurabile. Tale indice ha valore solo teorico e permette di studiare il comportamento asintotico dell'errore algoritmico e quindi dell'errore totale del metodo esaminato. Vale il seguente teorema.

Teorema 3.1

Siano u la precisione di macchina, n la dimensione della matrice A e $M = \max_{1 \leq i,j \leq n} |a_{ij}|$.
Allora il valore $N(n, M, t) = (2nM + t)$ è tale che, $\forall N \geq N(n, M, t)$, $\epsilon_T(N) < u$, asintoticamente rispetto a n e a M.

Dimostrazione.

Si pone, per semplicità, $\overline{N} := N(n, M, t)$. Essendo $\|A\| \leq nM$, allora

$$\epsilon_T(\overline{N}) \leq \frac{(nM)^{\overline{N}}}{\overline{N}!} e^{nM}.$$

Poichè lo studio viene effettuato per $n, M \to +\infty$ allora $\overline{N}! \simeq \overline{N}^{\overline{N}}$, cioè

$$\frac{(nM)^{\overline{N}}}{\overline{N}!} e^{nM} \simeq \left(\frac{nM}{\overline{N}}\right)^{\overline{N}} e^{nM}.$$

Si vuole dimostrare che

$$\left(\frac{nM}{\overline{N}}\right)^{\overline{N}} e^{nM} < 2^{-t},$$

cioè

$$\overline{N} \log\left(\frac{nM}{\overline{N}}\right) < -t - nM \log(e)$$

$$(2nM + t) \log\left(\frac{nM}{\overline{N}}\right) < -t - nM \log(e)$$

$$t + nM \log(e) < (2nM + t) \log\left(\frac{2nM + t}{nM}\right).$$

Ma si ha che

$$(2nM + t) \log\left(\frac{2nM + t}{nM}\right) > (2nM + t) \log 2 = (2nM + t) > t + nM \log(e)$$

e quindi si conclude.

■

Il risultato espresso dal teorema 3.1 non è utilizzabile nella pratica in quanto fornisce un valore $\overline{N}$ troppo elevato.

È ora conveniente studiare separatamente i casi in cui $\|A\| \leq 1$ e $\|A\| > 1$.

Primo caso: $\|A\| \leq 1$.

Sotto queste ipotesi

$$\varepsilon_T(N) < \frac{1}{(N+1)!}\frac{(N+2)}{(N+1)}.$$

Imponendo la condizione

$$\varepsilon_T(\overline{N}) \leq \frac{1}{(\overline{N}+1)!}\quad\frac{(\overline{N}+2)}{(\overline{N}+1)} < u,$$

si ricava facilmente il valore $\overline{N}$.

Ad esempio, utilizzando un calcolatore IBM 3081K della serie 370, la cui precisione di macchina è pari a 2^{-20}, si ottiene $\overline{N} = 9$.

Secondo caso: $\|A\| > 1$.

In questo caso l'indice $\overline{N}$ dipende fortemente da $\|A\|$ ed è conveniente calcolarlo durante l'esecuzione dell'algoritmo sfruttando le maggiorazioni (1) o (2) dell'errore analitico.

3.2 Analisi dell'errore algoritmico.

Si vuole calcolare $T_N(A) = \sum_{i=0}^{N} \frac{A^i}{i!}$ mediante **l'algoritmo di troncamento della serie di Taylor** , organizzato come segue:

$T_N(A) := I$
for $i := 1$ to N
$\quad T_N(A) := T_N(A) \oplus \frac{A \otimes \mathbb{1}}{i!}$.

Si suppone che vengano eseguite esattamente le operazioni del tipo scalare per matrice.

Sia $\varepsilon_{ALG}(N)$ la norma 1 dell'errore algoritmico assoluto, cioè

$$\varepsilon_{ALG}(N) = \|fl(T_N(A)) - T_N(A)\|_1.$$

Nel seguito vengono pressentate due maggiorazioni dell'errore algoritmico assoluto. La prima, meno accurata, ha valore teorico e viene utilizzata per lo studio asintotico dell'errore; la seconda ha utilità pratica e permette di

valutare i risultati ottenuti mediante l'algoritmo di troncamento della serie di Taylor.

Il seguente teorema consente di ottenere la prima maggiorazione.

Teorema 3.2

Siano $u = 2^{-t}$ la precisione di macchina e t_1 un numero reale tale che $2^{-t_1} = 1.06 \quad 2^{-t}$. Allora l'errore algoritmico assoluto commesso nel calcolare $T_N(A)$ può essere maggiorato nel modo seguente

$$||fl(T_N(A)) - T_N(A)|| \leq$$

$$\leq 2^{-t}\left[N + \sum_{j=0}^{N-1}(N-j)\frac{||A||^{j+1}}{(j+1)!} + 1.06n||A||^2\sum_{j=0}^{N-2}\frac{||A||^j}{(j+2)!}(j+1)\right],$$

dove

$$2^{-t}\left[N + \sum_{j=0}^{N-1}(N-j)\frac{||A||^{j+1}}{(j+1)!}\right]$$

è la maggiorazione dell'errore dovuto alla somma di matrici e

$$2^{-t_1}n||A||^2\sum_{j=0}^{N-2}\frac{||A||^j}{(j+2)!}(j+1)$$

è la maggiorazione dell'errore dovuto al prodotto di matrici.

Dimostrazione.

Come riportato in Wilkinson [11], se A, B, E, F sono matrici tali che

$$A \oplus B = A + B + E,$$

$$A \otimes B = A * B + F,$$

allora si ha

$$||E|| \leq 2^{-t}(||A|| + ||B||),$$

$$||F|| \leq 2^{-t_1}(||A|| * ||B||).$$

La dimostrazione viene fatta nel caso N=4, ma è facilmente generalizzabile.

Si indicano con F_i gli errori dovuti ai prodotti matriciali e con E_i quelli dovuti alla somma di matrici.

$$I \oplus A \oplus \frac{A^{②}}{2!} \oplus \frac{A^{③}}{3!} \oplus \frac{A^{④}}{4!} =$$

$$I + A + E_0 + \frac{A^2 + F_0}{2!} + E_1 +$$

$$+\frac{(A^2+F_0)A+F_1}{3!}+E_2+\frac{(A^3+F_0A+F_1)A+F_2}{4!}+E_3,$$

dove

$$\|F_0\| \leq n2^{-t_1}\|A\|^2$$

$$\|F_1\| \leq n2^{-t_1}\|A^2+F_0\|\|A\| \stackrel{\bullet}{\leq} n2^{-t_1}\|A\|^3$$

$$\|F_2\| \leq n2^{-t_1}\|A^3+F_0A+F_1\|\|A\| \stackrel{\bullet}{\leq} n2^{-t_1}\|A\|^4$$

$$\|E_0\| \leq 2^{-t}(\|I\|+\|A\|) = 2^{-t}(1+\|A\|)$$

$$\|E_1\| \leq 2^{-t}\left(\|I+A+E_0\|+\left\|\frac{A^2+F_0}{2!}\right\|\right) \stackrel{\bullet}{\leq} 2^{-t}\left(1+\|A\|+\frac{\|A\|^2}{2!}\right)$$

$$\|E_2\| \leq 2^{-t}\left(\left\|I+A+E_0+\frac{A^2+F_0}{2!}+E_1\right\|+\left\|\frac{A^3+F_0A+F_1}{3!}\right\|\right) \stackrel{\bullet}{\leq}$$

$$\stackrel{\bullet}{\leq} 2^{-t}\left(1+\|A\|+\frac{\|A\|^2}{2!}+\frac{\|A\|^3}{3!}\right)$$

$$\|E_3\| \leq 2^{-t}\left(\left\|I+A+E_0+\frac{A^2+F_0}{2!}+E_1+\frac{A^3+F_0A+F_1}{3!}\right\|+\right.$$

$$\left.+\left\|\frac{A^4+F_0A^2+F_1A+F_2}{4!}\right\|\right) \stackrel{\bullet}{\leq}$$

$$\stackrel{\bullet}{\leq} 2^{-t}\left[1+\|A\|+\frac{\|A\|^2}{2!}+\frac{\|A\|^3}{3!}+\frac{\|A\|^4}{4!}\right]$$

Adesso è possibile maggiorare l'errore algoritmico.

$$\varepsilon_{ALG}(4) := \|fl(T_4(A))-T_4(A)\| =$$

$$= \left\|E_0+E_1+E_2+E_3+\frac{F_0}{2!}+\frac{F_0A}{3!}+\frac{F_1}{3!}+\frac{F_0A^2}{4!}+\frac{F_1A}{4!}+\frac{F_2}{4!}\right\| \leq$$

$$\leq 2^{-t}\left[\sum_{j=0}^{3}\sum_{i=0}^{j+1}\frac{\|A\|^i}{i!}+\|F_0\|\left(\frac{1}{2!}+\frac{\|A\|}{3!}+\frac{\|A\|^2}{4!}\right)+\right.$$

$$+\|F_1\|\left(\frac{1}{3!}+\frac{\|A\|}{4!}\right)+\|F_2\|\Big]$$

$$\leq 2^{-t}\left[4+\sum_{j=0}^{2}(4-j)\frac{\|A\|^{j+1}}{(j+1)!}+1.06n\left(\frac{\|A\|^2}{2!}+2\frac{\|A\|^3}{3!}+3\frac{\|A\|^4}{4!}\right)\right]=$$

essendo N=3

$$2^{-t}\left[N+\sum_{j=0}^{N-1}(N-j)\frac{\|A\|^{j+1}}{(j+1)!}+1.06n\|A\|^2\sum_{j=0}^{N-2}\frac{\|A\|^j}{(j+2)!}(j+1)\right].$$

■

Per poter valutare il comportamento asintotico dell'errore algoritmico, è opportuno distinguere i due casi seguenti
(1) $\|A\| \leq 1$,
(2) $\|A\| > 1$.

L'analisi viene effettuata per N= 2nM+t. Si denota con h la funzione maggiorante ottenuta a meno della costante 2^{-t}:

$$h(n)=N+\sum_{j=0}^{N-1}\|A\|^{j+1}\frac{(N-j)}{(j+1)!}+1.06n\sum_{j=0}^{N-2}\|A\|^{j+2}\frac{(j+1)}{(j+2)!}$$

Caso (1).
Si ha

$$h(n)\leq 2nM+t+\sum_{j=0}^{2nM+t-1}\frac{(2nM+t-j)}{(j+1)!}+1.06n\sum_{j=0}^{2nM+t-2}\frac{j+1}{(j+2)!}.$$

Si ottiene $h(n)=O(n)$. Infatti

$$h(n)\leq(2nM+t)+(2nM+t)\sum_{j=1}^{2nM+t}\frac{1}{j!}+1.06n\sum_{j=0}^{2nM+t-2}\frac{1}{j!}\leq$$

$$\leq(2nM+t)e+1.06ne=O(n),$$

viceversa

$$h(n)\geq 2nM+t=O(n).$$

Si osservi che le maggiorazioni degli errori dovuti alla somma e al prodotto di matrici sono entrambe di ordine O(n).

Caso (2).

Si ha che

$$h(n) = 2nM + t + \sum_{j=0}^{2nM+t-1} (2nM + t - j)\frac{||A||^{j+1}}{(j+1)!} +$$

$$+1.06n \sum_{j=0}^{2nM+t-2} \frac{||A||^{j+2}}{(j+2)!}(j+1),$$

da cui

i) $h(n) \leq O(n^2 M ||A||^2 e^{||A||})$;

ii) asintoticamente $h(n) \geq O(e^{||A||})$.

Queste due relazioni si dimostrano come segue.

$$i) \quad h(n) \leq (2nM + t)\left[\left(\sum_{j=0}^{2nM+t} \frac{||A||^j}{j!}\right) + 1.06n||A||^2 \sum_{j=0}^{2nM+t-2} \frac{||A||^j}{j!}\right] \leq$$

$$\leq (2nM + t + 1)[e^{||A||} + 1.06n||A||^2 e^{||A||}].$$

$$ii). \qquad h(n) \geq \sum_{j=1}^{2nM+t} \frac{||A||^j}{j!} + 1.06 \sum_{j=0}^{2nM+t-2} \frac{||A||^{j+2}}{(j+2)!}$$

Perciò si ha che

$$\lim_{n \to +\infty} h(n) \geq e^{||A||} - 1 + 1.06(e^{||A||} - 1 - ||A||).$$

Si può quindi concludere che se $||A|| \leq 1$ l'errore algoritmico è lineare rispetto alla dimensione e alla norma di A, mentre cresce in modo esponenziale se $||A|| > 1$.

È possibile ottenere una maggiorazione dell'errore algoritmico più accurata della precedente e quindi più utile per valutare, in pratica, l'errore commesso. Vale il seguente teorema.

Teorema 3.3

Siano $u = 2^{-t}$ la precisione di macchina e t_1 un numero reale tale che $2^{-t_1} = 1.06 \quad 2^{-t}$. L'errore algoritmico assoluto commesso nel calcolare $T_N(A)$ può essere maggiorato nel modo seguente:

$$\varepsilon_{ALG}(N) \leq 2^{-t} \max_{1\leq j\leq n} \left\{ \sum_{i=1}^{n} \left(\sum_{k=1}^{N} \frac{N-k+1}{k!} |a_{ij}^{(k)}| \right) + N + \right.$$

$$\left. +1.06 \sum_{i=1}^{n} \sum_{k=1}^{N-1} \left(\sum_{p=2}^{n} \frac{n-p+2}{(k+1)!} k |a_{ip}^{(k)} a_{pj}| + \frac{n}{(k+1)!} k |a_{i1}^{(k)} a_{1j}| \right) \right\}$$

dove $a_{ij}^{(k)}$ è l'elemento di posto (i,j) della matrice A^k.

Dimostrazione.
Sono necessarie alcune premesse.

Premessa 1. Studio dell'errore commesso per calcolare il seguente prodotto matriciale:

$$(A \otimes A) \otimes A,$$

dove $A \otimes A = A^2 + F_0$.

Si ha che

$$(A^2 + F_0) \otimes A = A^3 + F_0 A + F_1$$

dove $\|F_1\| \leq n 2^{-t_1} \|A^2 + F_0\| \|A\|$.

Poichè, nella precedente maggiorazione F_0 è trascurabile, F_1 può essere considerato come l'errore dovuto al calcolo di $A^2 \otimes A$.

Premessa 2. Valutazione degli errori commessi nel calcolo di

i) $$A^2 \otimes A$$

e di

ii) $$A \otimes A^2$$

i)
$$|fl(a_{ij}^{(3)}) - a_{ij}^{(3)}| =$$

$$\left| fl \left(\sum_{p=1}^{n} a_{ip}^{(2)} a_{pj} \right) - a_{ij}^{(3)} \right| =$$

$$\left| a_{ij}^{(3)} + \sum_{p=1}^{n} \varepsilon_p a_{ip}^{(2)} a_{pj} - a_{ij}^{(3)} \right| =$$

$$\left| \sum_{p=1}^{n} \varepsilon_p \sum_{q=1}^{n} a_{iq} a_{qp} a_{pj} \right|,$$

con $|\varepsilon_p| \leq 2^{-t_1}(n-p+2)$ se $p \geq 2$ e $|\varepsilon_1| \leq 2^{-t_1} n$.

ii)

$$|fl(a_{ij}^{(3)}) - a_{ij}^{(3)}| =$$

$$\left| fl\left(\sum_{p=1}^{n} a_{ip} a_{pj}^{(2)}\right) - a_{ij}^{(3)} \right| =$$

$$\left| a_{ij}^{(3)} + \sum_{p=1}^{n} \delta_p a_{ip} a_{pj}^{(2)} - a_{ij}^{(3)} \right| =$$

$$\left| \sum_{p=1}^{n} \delta_p \sum_{q=1}^{n} a_{ip} a_{pq} a_{qj} \right|,$$

con $|\delta_q| \leq 2^{-t_1}(n-q+2)$ se $q \geq 2$ e $|\delta_1| \leq 2^{-t_1} n$.

Perciò gli errori commessi hanno analogo comportamento e per lo studio dell'errore può essere esaminato l'uno o l'altro caso indifferentemente.

Dopo queste premesse è possibile dimostrare il teorema; la dimostrazione viene fatta, per semplicità, nel caso N=3, ma è facilmente generalizzabile.

Consideriamo dapprima l'errore commesso calcolando il prodotto matriciale e la variazione, dovuta a tale errore, dei singoli elementi delle matrici.

$$|fl(a_{ij}^{(2)}) - a_{ij}^{(2)}| =$$

$$\left| a_{ij}^{(2)} + \sum_{p=1}^{n} \varepsilon_p a_{ip} a_{pj} - a_{ij}^{(2)} \right| \leq$$

$$\leq \sum_{p=1}^{n} |\varepsilon_p| |a_{ip} a_{pj}| \leq$$

$$\leq 2^{-t_1} \left(n|a_{i1} a_{1j}| + \sum_{p=1}^{n} (n-p+2)|a_{ip} a_{pj}| \right),$$

poichè $|\varepsilon_p| \leq 2^{-t_1}(n-p+2)$, $p \geq 2$ e $|\varepsilon_1| \leq 2^{-t_1} n$.

$$|fl(a_{ij}^{(3)}) - a_{ij}^{(3)}| =$$

$$\left| fl\left(\sum_{p=1}^{n} a_{ip} fl(a_{pj}^{(2)}) \right) - a_{ij}^{(3)} \right| =$$

$$\left| \sum_{p=1}^{n} a_{ip} fl(a_{pj}^{(2)})) + f_{ij} - a_{ij}^{(3)} \right|$$

dove f_{ij} è l'elemento (i,j) della matrice F tale che

$$A \otimes (A \otimes A) = A(A \otimes A) + F.$$

Ma per le premesse precedenti vale che

$$(A \otimes A) \otimes A \simeq (A \otimes A)A + F,$$

ed F può essere maggiorata come l'errore dovuto al calcolo di $A^2 \otimes A$.
Perciò si ottiene

$$f_{ij} = \sum_{p=1}^{n} \delta_p a_{ip}^{(2)} a_{pj},$$

con $|\delta_q| \leq 2^{-t_1}(n-q+2)$ se $q \geq 2$ e $|\delta_1| \leq 2^{-t_1} n$.
Ritornando alla relazione precedente si ha che

$$|fl(a_{ij}^{(3)}) - a_{ij}^{(3)}| =$$

$$\left| f_{ij} + \sum_{p=1}^{n} a_{ip} a_{pj}^{(2)} + \sum_{p,q=1}^{n} \gamma_q a_{ip} a_{pq} a_{qj} - a_{ij}^{(3)} \right| =$$

$$= \left| f_{ij} + \sum_{p,q=1}^{n} \gamma_q a_{ip} a_{pq} a_{qj} \right| =$$

$$= \left| \sum_{p=1}^{n} \delta_p a_{ip}^{(2)} a_{pj} + \sum_{q=1}^{n} \gamma_q a_{iq}^{(2)} a_{qj} \right| =$$

$$= \left| \sum_{p=1}^{n} (\delta_p + \gamma_p) a_{ip}^{(2)} a_{pj} \right| \leq$$

$$\leq 2^{-t_1} 2 \left(\sum_{p=2}^{n} (n-p+2) |a_{ip}^{(2)} a_{pj}| + n |a_{i1}^{(2)} a_{1j}| \right).$$

Perciò, in generale , si ha che

$$\left| fl\left(\frac{a_{ij}^{(k)}}{k!} \right) - \frac{a_{ij}^{(k)}}{k!} \right| \leq$$

$$\leq 2^{-t_1}\frac{k-1}{k!}\left[\sum_{p=2}^{n}(n-p+2)|a_{ip}^{(k-1)}a_{pj}|+n|a_{i1}^{(k-1)}a_{1j}|\right].$$

Consideriamo l'elemento (i,j) della matrice errore algoritmico totale commesso per calcolare $T_N(A)$, $N=3$.

$$e_{ij}=\left|fl\left(\delta_{ij}+a_{ij}+\frac{a_{ij}^{(2)}}{2!}+\frac{a_{ij}^{(3)}}{3!}\right)-\left(\delta_{ij}+a_{ij}+\frac{a_{ij}^{(2)}}{2!}+\frac{a_{ij}^{(3)}}{3!}\right)\right|=$$

$$=\left|\left(\left((\delta_{ij}+a_{ij})(1+\gamma_{ij}^{(1)})\delta_{ij}+fl\left(\frac{a_{ij}^{(2)}}{2!}\right)\right)(1+\gamma_{ij}^{(2)})+\right.$$

$$\left.+fl\left(\frac{a_{ij}^{(3)}}{3!}\right)\right)(1+\gamma_{ij}^{(3)})-\left(\delta_{ij}+a_{ij}+\frac{a_{ij}^{(2)}}{2!}+\frac{a_{ij}^{(3)}}{3!}\right)\Bigg|\doteq$$

$$\doteq\Bigg|(\delta_{ij}+a_{ij})+\gamma_{ij}^{(1)}\delta_{ij}(\delta_{ij}+a_{ij})+fl\left(\frac{a_{ij}^{(2)}}{2!}\right)+$$

$$+\gamma_{ij}^{(2)}(\delta_{ij}+a_{ij})+fl\left(\frac{a_{ij}^{(2)}}{2!}\right)\gamma_{ij}^{(2)}+fl\left(\frac{a_{ij}^{(3)}}{3!}\right)+$$

$$+\gamma_{ij}^{(3)}(\delta_{ij}+a_{ij})+fl\left(\frac{a_{ij}^{(2)}}{2!}\right)\gamma_{ij}^{(3)}+fl\left(\frac{a_{ij}^{(3)}}{3!}\right)\gamma_{ij}^{(3)}+$$

$$-\left(\delta_{ij}+a_{ij}+\frac{a_{ij}^{(2)}}{2!}+\frac{a_{ij}^{(3)}}{3!}\right)\Bigg|\doteq$$

$$\doteq\left|fl(\frac{a_{ij}^{(2)}}{2!})-\frac{a_{ij}^{(2)}}{2!}\right|+\left|fl(\frac{a_{ij}^{(3)}}{3!})-\frac{a_{ij}^{(3)}}{3!}\right|+$$

$$+\left|(\delta_{ij}+a_{ij})(\gamma_{ij}^{(1)}\delta_{ij}+\gamma_{ij}^{(2)}+\gamma_{ij}^{(3)})\right|+$$

$$+\left|\frac{a_{ij}^{(2)}}{2!}(\gamma_{ij}^{(2)}+\gamma_{ij}^{(3)})\right|+\left|\frac{a_{ij}^{(3)}}{3!}\gamma_{ij}^{(3)}\right|\leq$$

poichè $|\gamma_{ij}^{(h)}|\leq 2^{-t}$,

$$\leq 2^{-t}\left(1.06\sum_{k=2}^{3}\frac{k-1}{k!}\left[\sum_{p=2}^{n}(n-p+2)|a_{ip}^{(k-1)}a_{pj}|+n|a_{i1}^{(k-1)}a_{1j}|\right]+\right.$$

$$+3|\delta_{ij}+a_{ij}|+2\left|\frac{a_{ij}^{(2)}}{2!}\right|+\left|\frac{a_{ij}^{(3)}}{3!}\right|\Bigg)$$

ed essendo $N=3$ si ottiene

$$\leq 2^{-t}\Bigg(1.06\sum_{k=1}^{N-1}\frac{k}{(k+1)!}\Bigg[\sum_{p=2}^{n}(n-p+2)|a_{ip}^{(k)}a_{pj}|+n|a_{i1}^{(k)}a_{1j}|\Bigg]+$$

$$+N+\sum_{k=1}^{N}\frac{N-k+1}{k!}|a_{ij}^{(k)}|\Bigg).$$

Poichè $\|E\|=\max_{1\leq j\leq n}\sum_{i=1}^{n}|e_{ij}|$ si conclude. ∎

4. Calcolo dell'esponenziale di una matrice hermitiana e definita positiva.

Esaminiamo il comportamento della serie di potenze matriciale, che definisce la matrice esponenziale, nel caso di matrici hermitiane definite positive. Le osservazioni che seguono giustificano il fatto che il metodo di Taylor è stabile per tale classe di matrici.

Sia A una matrice nxn hermitiana e definita positiva. Denotiamo con S_k le somme parziali della serie esponenziale, cioè

$$S_k = \sum_{i=0}^{k} \frac{A^i}{i!}.$$

Vale il seguente risultato.

Teorema 4.1

Sia A una matrice nxn hermitiana e definita positiva; allora si ha che:

(1) gli elementi principali delle matrici A^k, e quindi delle matrici S_k, sono positivi.

(2) Definitivamente ogni elemento di posto (i,j), $i \neq j$, della matrice A^k conserva il segno, se la matrice è reale. Se la matrice è complessa, tale proprietà vale per la parte reale e per la parte immaginaria di ogni elemento.

(3) Ogni elemento di posto (i,j), $i \neq j$, della matrice A^k cambia segno al più due volte, se la matrice è reale. Se la matrice è complessa, tale proprietà vale per la parte reale e per la parte immaginaria di ogni elemento.

Dimostrazione.

(1) È ovvio che, essendo A una matrice hermitiana definita positiva, allora anche A^k è hermitiana e definita positiva. Infatti gli autovalori di A^k sono le potenze k-esime degli autovalori di A. È quindi sufficiente dimostrare che gli elementi principali di matrici hermitiane e definite positive sono positivi. Siano B una matrice hermitiana e definita positiva ed e_i l'i-esimo vettore della base canonica. Essendo e_i non nullo si ha che

$$b_{ii} = e_i^T B e_i > 0.$$

(2) Siano $d_1, ..., d_n$ gli autovalori di A (reali e positivi per ipotesi) tali che

$$d_1 = ... = d_r > d_{r+1} \geq ... \geq d_n$$

Caso reale.

Poichè per ipotesi $A = QDQ^T$, Q ortogonale e D diagonale, si ha

$$a_{ij}^{(k)} = \sum_{p=1}^{n} d_p^k q_{ip} q_{jp}$$

da cui si ottiene

$$a_{ij}^{(k)} = d_1^k \Big(\sum_{p=1}^{r} q_{ip} q_{jp} + \sum_{p=r+1}^{n} \left(\frac{d_p}{d_1} \right)^k q_{ip} q_{jp} \Big).$$

Per semplicità si denota con

$$\alpha := \sum_{p=1}^{r} q_{ip} q_{jp}$$

Primo caso: $r = n$.

$$a_{ij}^{(k)} = d_1^k \Big(\sum_{p=1}^{n} q_{ip} q_{jp} \Big) = d_1^k \delta_{ij},$$

e si conclude perchè $d_1 > 0$.

Secondo caso: $r < n$.
Sia $\alpha \neq 0$.

i) $d_1 \geq 1$.

$$a_{ij}^{(k)} = d_1^k \Big(\alpha + \sum_{p=r+1}^{n} \left(\frac{d_p}{d_1} \right)^k q_{ip} q_{jp} \Big).$$

Poichè $\left(\frac{d_p}{d_1} \right) < 1$ si ha che, per $k \to +\infty$,

$$a_{ij}^{(k)} \to \alpha \quad \text{se} \quad d_1 = 1,$$

$$a_{ij}^{(k)} \to segno(\alpha)\infty \quad \text{se} \quad d_1 > 1.$$

Perciò definitivamente $a_{ij}^{(k)}$ ha lo stesso segno di α.

ii) $d_1 < 1$.
Poichè definitivamente

$$\alpha > \sum_{p=r+1}^{n} \left(\frac{d_p}{d_1} \right)^k q_{ip} q_{jp},$$

si ha che

$$a_{ij}^{(k)} \to 0^+ \quad \text{se} \quad \alpha > 0$$

$$a_{ij}^{(k)} \to 0^- \quad \text{se} \quad \alpha < 0$$

si conclude come nel punto (i).
Se $\alpha = 0$, allora

$$a_{ij}^{(k)} = \sum_{p=r+1}^{n} d_p^k q_{ip} q_{jp}.$$

Si conclude con un ragionamento analogo al precedente utilizzando d_{r+1} al posto di d_1.

Caso complesso.

Poichè per ipotesi $A = QDQ^H$, Q unitaria e D diagonale, si ha

$$a_{hj}^{(k)} = \sum_{p=1}^{n} d_p^k q_{hp} q_{jp}^*.$$

Denotiamo, per semplicità, con

$$\alpha_p := Re(q_{hp} q_{jp}^*)$$

$$\beta_p := Im(q_{hp} q_{jp}^*)$$

sottointendendo gli indici h e j. Allora si ha che

$$a_{hj}^{(k)} = \sum_{p=1}^{n} d_p^k (\alpha_p + i\beta_p),$$

cioè

$$Re(a_{hj}^{(k)}) = \sum_{p=1}^{n} d_p^k \alpha_p$$

$$Im(a_{hj}^{(k)}) = \sum_{p=1}^{n} d_p^k \beta_p.$$

Si conclude con dimostrazione analoga al caso reale.

(3) Sono necessari due fatti preliminari.

(i) Tutte le successioni del tipo $\alpha(1 - a^n)$ con $\alpha > 0, 0 < a < 1$, sono monotone crescenti e giacciono su funzioni con concavità verso il basso.
Infatti per la crescenza si ha:
$\alpha(1 - a^{n+1}) - \alpha(1 - a^n) = \alpha(a^n)(1 - a) > 0$ per ipotesi.

Concavità verso il basso:
si conclude poichè la funzione $f(x) = \alpha(1 - a^x)$ con $x > 0$ e $\alpha > 0$, $0 < a < 1$, ha derivata seconda $f(x) = -\alpha a^x (loga)^2 < 0$.

(ii) I termini della successione

$$\left(\sum_{i=1}^{k} \alpha_i (1 - a_i^n) \right)_{n \geq 1} \qquad 0 < a_i < 1$$

presentano al più due cambiamenti di segno. Infatti, siano

$$\alpha_i^+ = \begin{cases} \alpha_i & \text{se} \quad \alpha_i \geq 0 \\ 0 & \text{altrimenti,} \end{cases}$$

$$\alpha_i^- = \begin{cases} -\alpha_i & \text{se} \quad \alpha_i < 0 \\ 0 & \text{altrimenti} \end{cases}$$

e

$$s_n := \sum_{i=1}^{k} \alpha_i^+ (1 - a_i^n),$$

$$v_n := \sum_{i=1}^{k} \alpha_i^- (1 - a_i^n).$$

Le successioni s_n e v_n sono monotone crescenti e giacciono su funzioni con concavità verso il basso, perchè somma di successioni con tali proprietà. Ma due funzioni crescenti, positive, con concavità verso il basso si intersecano al più due volte. Da ciò si ha che i termini

$$\sum_{i=1}^{k} \alpha_i (1 - a_i^n) = s_n - v_n$$

cambiano segno al più due volte e si conclude.

Utilizzando questi preliminari è possibile concludere la dimostrazione.
Caso reale.

$$a_{ij}^{(k)} = d_1^k \left(\sum_{p=1}^{r} q_{ip} q_{jp} + \sum_{p=r+1}^{n} \left(\frac{d_p}{d_1} \right)^k q_{ip} q_{jp} \right).$$

Essendo Q una matrice ortogonale vale che

$$\sum_{p=1}^{n} q_{ip} q_{jp} = 0,$$

cioè

$$\sum_{p=1}^{r} q_{ip}q_{jp} = -\sum_{p=r+1}^{n} q_{ip}q_{jp}.$$

Perciò $a_{ij}^{(k)}$ può essere riscritta nella forma

$$a_{ij}^{(k)} = d_1^k \left(\sum_{p=r+1}^{n} \left(\left(\frac{d_p}{d_1} \right)^k - 1 \right) q_{ip}q_{jp}. \right.$$

È sufficiente studiare il segno di

$$\sum_{p=r+1}^{n} \left(1 - \left(\frac{d_p}{d_1} \right)^k \right) q_{ip}q_{jp}.$$

Ponendo

$$\alpha_{p-r} := q_{ip}q_{jp} \qquad \text{e} \qquad a_{p-r} := \left(\frac{d_p}{d_1} \right) \quad 0 < a_{p-r} < 1,$$

la sommatoria precedente può essere riscritta come

$$t_k = \sum_{p=1}^{n-r} \alpha_p \left(1 - a_p^k \right).$$

Utilizzando le premesse si conclude che questi elementi possono, al variare di k, cambiare segno al più due volte.

Il caso complesso si dimostra in modo analogo, operando separatamente sulla parte reale e su quella immaginaria.

■

5. Metodo scaling and squaring.

Il metodo scaling and squaring per il calcolo della matrice esponenziale e^A si basa sulla seguente proprietà.

Dati una matrice A di ordine n e un numero $m \geq 0$, si ha che

$$e^A = \left(e^{\frac{A}{m}}\right)^m.$$

Poichè dall'analisi dell'errore algoritmico del metodo di troncamento della serie di Taylor per il calcolo della matrice e^A risulta che si possono ottenere risultati numericamente più attendibili se $\|A\| < 1$, è possibile calcolare la matrice esponenziale mediante l'algoritmo scaling and squaring, come riportato in [5]. Tale algoritmo è organizzato come segue.

Algoritmo scaling and squaring.

1. Scegliere l'intero m tale che

$$m = \min\left\{n \epsilon N : n = 2^p, n \geq \|A\|\right\}.$$

2. Calcolare

$$T_N\left(\frac{A}{m}\right) = \sum_{i=0}^{N} \left(\frac{A}{m}\right)^i \frac{1}{i!}$$

mediante l'algoritmo di Taylor.

3. Calcolare la potenza

$$\left(T_N\left(\frac{A}{m}\right)\right)^m.$$

■

Si noti che l'algoritmo di Taylor è usato in modo conveniente poichè $\left\|\frac{A}{m}\right\| < 1$.

5.1 Studio dell'errore.

Un'utile premessa è lo studio dell'errore inerente dovuto al calcolo delle potenze di una matrice.

Teorema 5.1

Siano

B una matrice di ordine n,
E la matrice degli errori da cui è affetta B tale che $EB = BE$
E' la matrice degli errori relativi da cui è affetta B e
m un intero.

Allora l'errore commesso nel calcolare $(B+E)^m$ è tale che

$$\frac{\|B^m - (B+E)^m\|}{\|B^m\|} \leq muK(B).$$

Dimostrazione.

Trascuriamo $E^k B^{m-k}$, $k > 1$, rispetto a EB^{m-1}, poichè $E = E'B$ e possiamo supporre che la matrice E' degli errori relativi sia, in norma, maggiorata da u.

Sotto queste ipotesi si ha che

$$(B+E)^m \stackrel{\bullet}{=} B^m + B^{m-1}E + BEB^{m-2} + \ldots + EB^{m-1} =$$

per la commutatività di B ed E

$$= B^m + mB^{m-1}E.$$

È quindi possibile valutare l'errore relativo commesso.

$$\frac{\|B^m - (B+E)^m\|}{\|B^m\|} \stackrel{\bullet}{\leq} \frac{m\|B^{m-1}\|\|E\|}{\|B^m\|} =$$

$$= \frac{m\|B^{m-1}BB^{-1}\|\|E\|}{\|B^m\|} \leq m\|B^{-1}\|\|E\|.$$

Sostituendo alla matrice E la matrice E' degli errori relativi si ottiene

$$\frac{\|B^m - (B+E)^m\|}{\|B^m\|} \leq m\|E'\|K(B) \leq muK(B).$$

■

È ragionevole supporre che:

1) l'errore inerente abbia, nella maggior parte dei casi, comportamento analogo a quello descritto nel teorema precedente;
2) l'errore algoritmico sia dello stesso ordine dell'errore inerente, cioè

$$\frac{\|A^m - fl(A^m)\|}{\|A^m\|} \leq c_1 muK(A).$$

Vale il seguente teorema.

Teorema 5.2

Siano A una matrice di ordine n e m un intero. Allora

$$\frac{\|A^m - fl(A^m)\|}{\|A^m\|} \leq nmuK(A).$$

Dimostrazione.

Per le considerazioni fatte precedentemente è sufficiente dimostrare che $c_1 = n$. Poichè c_1 è indipendente da m è sufficiente calcolarlo per $m = 2$. Utilizzando le maggiorazioni di Wilkinson [11], si ottiene

$$\frac{\|fl(A^2) - A^2\|}{\|A^2\|} \leq n2^{-t_1}\frac{\|A\|^2}{\|A^2\|} \simeq nuK(A).$$

■

È possibile utilizzando le conclusioni precedenti maggiorare l'errore totale relativo commesso dall'algoritmo scaling and squaring.

Teorema 5.3

Siano A una matrice di ordine n e m un intero. Per semplicità denotiamo con E_{ALG} l'errore algoritmico commesso per calcolare $T_N(\frac{A}{m})$,

E_{TR} l'errore analitico commesso per calcolare $T_N(\frac{A}{m})$,

E^{rel} l'errore relativo e con

T_N la matrice $T_N\left(\frac{A}{m}\right)$.

Allora

$$\frac{\|e^A - fl\,(fl\,(T_N)^m)\,\|}{\|e^A\|} \leq umK(e^{\frac{A}{m}})\left(\frac{\beta}{\|e^A\|} + n\right),$$

dove

$$\beta = \max_{1\leq j\leq n}\left\{\sum_{i=1}^{n}\left(\sum_{k=1}^{N}\frac{N-k+1}{k!}|a_{ij}^{(k)}|\right) + N + \right.$$

$$\left. +1.06\sum_{i=1}^{n}\sum_{k=1}^{N-1}\left(\sum_{p=2}^{n}\frac{n-p+2}{(k+1)!}k|a_{ip}^{(k)}a_{pj}| + \frac{n}{(k+1)!}k|a_{i1}^{(k)}a_{1j}|\right)\right\}$$

e

$$\|E_{ALG}^{rel}\| \leq \frac{\beta}{\|e^A\|}.$$

Dimostrazione.

$$\frac{\|e^A - fl\,(fl\,(T_N)^m)\,\|}{\|e^A\|} =$$

$$= \frac{\|e^A - fl\,((T_N + E_{ALG})^m)\,\|}{\|e^A\|} =$$

$$\frac{\|e^A - (e^{\frac{A}{m}} - E_{TR} + E_{ALG})^m + fl((T_N + E_{ALG})^m) - (T_N + E_{ALG})^m\|}{\|e^A\|} \leq$$

$$\leq \frac{\|e^A - (e^{\frac{A}{m}} - E_{TR} + E_{ALG})^m\| +}{\|e^A\|} \leq$$

$$\frac{\| + fl((T_N + E_{ALG})^m) - (T_N + E_{ALG})^m\|}{\|e^A\|} \leq$$

$$mK(e^{\frac{A}{m}})\|E_{TR}^{rel} + E_{ALG}^{rel}\| + mnuK(T_N + E_{ALG})$$

$$\simeq mK(e^{\frac{A}{m}})\left(\|E_{TR}^{rel} + E_{ALG}^{rel}\| + nu\right)$$

$$\simeq mK(e^{\frac{A}{m}})\left(\|E_{ALG}^{rel}\| + nu\right) \leq$$

$$umK(e^{\frac{A}{m}})\left(\frac{\beta}{\|e^A\|} + n\right).$$

■

Osservazioni.

1. La maggiorazione precedente è limitata in quanto, essendo $\frac{\|A\|}{m} < 1$, $\|E_{ALG}^{rel}\|$ non può raggiungere valori elevati, come riportato nel paragrafo 4, ed, inoltre, la matrice $e^{\frac{A}{m}}$ è ben condizionata. Infatti

   $$K(e^{\frac{A}{m}}) = \|e^{\frac{A}{m}}\|\|(e^{\frac{A}{m}})^{-1}\| = \|e^{\frac{A}{m}}\|\|e^{\frac{-A}{m}}\| \leq e^{2\frac{\|A\|}{m}} \leq e^2.$$

   Sia D una matrice diagonale i cui elementi sono d e $-d$; si ottiene $K(e^D) = e^{2d}$. Se $d = 1 - \frac{\epsilon}{2}$ allora $K(e^D) = e^{2-\epsilon}$. Da questa osservazione si vede che e^2 è una maggiorazione accurata del numero di condizionamento dell'esponenziale di una qualsiasi matrice con norma minore di 1.
2. I risultati esposti in questo paragrafo non sfruttano il fatto che $m = 2^p$. Tale scelta è utile per rendere più efficiente il calcolo di A^m.

6. Confronto tra il metodo di Taylor e il metodo scaling and squaring per matrici con norma maggiore di 1.

Sia A una matrice n x n tale che $||A|| > 1$.
Il confronto tra il metodo di Taylor e il metodo scaling and squaring si basa, ovviamente, sulla valutazione della stabilità e della complessità.

6.1 Complessità.

Il metodo scaling and squaring richiede un numero di prodotti matriciali pari alla somma di quelli sufficienti a calcolare l'esponenziale di una matrice con norma minore di 1 mediante il metodo di Taylor e di quelli per calcolare l'm-esima potenza di una matrice, dove m è la minima potenza di 2 che maggiora $||A||$.

Utilizzando un calcolatore IBM 3081K della serie 370, si ottiene un numero di prodotti matriciali pari a $\lceil 9 + \log_2 ||A|| \rceil$.

D'altro canto, sperimentalmente si vede che il numero di prodotti matriciali calcolati con il metodo di Taylor è assai più elevato e sembra crescere linearmente con $||A||$. Questa ipotesi è anche supportata dal fatto che il valore $\overline{N} = 2nM + t$, per cui l'errore analitico con cui $T_{\overline{N}}(A)$ approssima e^A è trascurabile, dipende linearmente da nM, quantità legata a $||A||$.

6.2 Stabilità numerica.

Nel caso in cui sia possibile calcolare, senza gravi errori la matrice esponenziale utilizzando il metodo di troncamento della serie di Taylor (ad esempio se la matrice in input è hermitiana e definita positiva), allora il risultato ottenuto è affetto da un errore che non si discosta in modo sensibile da quello commesso con il metodo scaling and squaring. Questo secondo algoritmo, però, garantisce risultati attendibili anche quando il metodo di Taylor fallisce.

In conclusione, se la norma delle matrice è maggiore di 1, il metodo scaling and squaring garantisce, rispetto al metodo di Taylor, risultati numericamente più attendibili, ottenibili, inoltre, con un numero minore di prodotti matriciali.

7. Esempi.

Gli esempi riportati in questo paragrafo sono stati calcolati utilizzando un calcolatore IBM 3081K della serie 370,la cui precisione di macchina è 2^{-20}.

Data una matrice A nxn si ha che:

la norma utilizzata è la norma 1, cioè

$$\|A\|_1 = \max_{1 \le j \le n} \sum_{i=1}^{n} |a_{ij}|;$$

con il termine errore relativo si indica la quantità

$$\frac{\|e^A - fl(T_N(A))\|}{\|e^A\|};$$

con il termine valore esatto di una matrice si intende che, per ogni elemento di tale matrice, il valore esatto viene troncato, ma le cifre riportate sono esatte.

Esempio 1.

In questo esempio viene calcolata l'esponenziale di matrici hermitiane definite positive e definite negative, con norma minore di 1, utilizzando l'algoritmo di troncamento della serie di Taylor. Poichè la norma delle matrici considerate è minore di 1, i risultati ottenuti non sono affetti da errori elevati.

Sia

$$A = \begin{pmatrix} .552 & -.256 \\ -.256 & .168 \end{pmatrix};$$

Tale matrice è:

1) hermitiana (rientra perciò nella classe di matrici esaminate);
2) ammette la seguente decomposizione:

$$A = \begin{pmatrix} \frac{1}{\sqrt{5}} & \frac{-2}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}} \end{pmatrix} \begin{pmatrix} \frac{1}{25} & 0 \\ 0 & \frac{17}{25} \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \\ \frac{-2}{\sqrt{5}} & \frac{1}{\sqrt{5}} \end{pmatrix};$$

dove le matrici

$$\begin{pmatrix} \frac{1}{\sqrt{5}} & \frac{-2}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}} \end{pmatrix} \quad \text{e} \quad \begin{pmatrix} \frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \\ \frac{-2}{\sqrt{5}} & \frac{1}{\sqrt{5}} \end{pmatrix};$$

sono unitarie;

3) è definita positiva.

La matrice esponenziale

$$e^A = \begin{pmatrix} 1.787264 & -.3732267 \\ -.3732267 & 1.227424 \end{pmatrix}$$

può essere approssimata se si effettuano nove passi del metodo di Taylor, ottenendo

$$fl(T_9(A)) = \begin{pmatrix} 1.78726 & -.3732263 \\ -.3732263 & 1.227421 \end{pmatrix}.$$

L'errore relativo commesso è $.2097 \quad 10^{-5}$.

In questo caso alla matrice può essere applicato il teorema 4.1; più precisamente, gli elementi che occupano la medesima posizione nelle matrici A^k, $k = 1 \ldots 9$, hanno lo stesso segno: non si verificano perciò fenomeni di cancellazione nel calcolo delle somme parziali.

Ad esempio si ha che:

$$\text{valore esatto di } \frac{A^5}{5!} = \begin{pmatrix} .9692888 & -.484644 \\ -.484644 & .2423228 \end{pmatrix} 10^{-3}$$

$$\text{valore di } fl\left(\frac{A^5}{5!}\right) = \begin{pmatrix} .969285 & -.484643 \\ -.484643 & -.242322 \end{pmatrix} 10^{-3}$$

$$\text{valore esatto di } \sum_{i=0}^{5} \frac{A^i}{i!} = \begin{pmatrix} 1.787143 & -.373166 \\ -.373166 & 1.227394 \end{pmatrix}$$

$$\text{valore di } fl(T_5(A)) = \begin{pmatrix} 1.787140 & -.3731657 \\ -.3631657 & 1.227392 \end{pmatrix}.$$

Cambiando segno alla matrice A si ottiene la matrice normale definita negativa

$$A' = \begin{pmatrix} -.552 & .256 \\ .256 & -.168 \end{pmatrix},$$

la cui matrice esponenziale è

$$e^{A'} = \begin{pmatrix} .5974515 & .1816690 \\ .1816690 & .8699549 \end{pmatrix}.$$

Dopo nove passi dell'algoritmo di Taylor, si ottiene la matrice

$$fl(T_9(A')) = \begin{pmatrix} .5974515 & 1816690 \\ 1816690 & .8699549 \end{pmatrix},$$

che approssima l'esponenziale con un errore relativo uguale a $.3217 \quad 10^{-6}$.

$$\text{valore calcolato di } \frac{B^{15}}{15!} = \begin{pmatrix} 1751141 & -875570.2 \\ -875571 & 437785 \end{pmatrix}$$

$$\text{valore esatto di } \sum_{i=0}^{15} \frac{B^i}{i!} = \begin{pmatrix} 7177957 & -3588977 \\ -3588977 & 1794491 \end{pmatrix}$$

$$\text{valore calcolato di } \sum_{i=0}^{15} \frac{B^i}{i!} = \begin{pmatrix} 7177916 & -3588951 \\ -3588957 & 1794478 \end{pmatrix}.$$

$$k = 30$$

$$\text{valore esatto di } \frac{B^{30}}{30!} = \begin{pmatrix} 24711.41 & -12355.71 \\ -12355.71 & 6177.853 \end{pmatrix}$$

$$\text{valore calcolato di } \frac{B^{30}}{30!} = \begin{pmatrix} 24710.98 & -12355.54 \\ -12355.49 & 6177.77 \end{pmatrix}$$

$$\text{valore esatto di } \sum_{i=0}^{30} \frac{B^i}{i!} = \begin{pmatrix} 19295970 & -9647985 \\ -9647985 & 4823995 \end{pmatrix}$$

$$\text{valore calcolato di } \sum_{i=0}^{30} \frac{B^i}{i!} = \begin{pmatrix} 19295740 & -9647899 \\ -9647899 & 4823949 \end{pmatrix}.$$

$$k = 45$$

$$\text{valore esatto di } \frac{B^{45}}{45!} = \begin{pmatrix} .1568482 & -.0784241 \\ -.0784241 & .03921205 \end{pmatrix}$$

$$\text{valore calcolato di } \frac{B^{45}}{45!} = \begin{pmatrix} .1568442 & -.07842231 \\ -.07842207 & .03921117 \end{pmatrix}$$

$$\text{valore esatto di } \sum_{i=0}^{45} \frac{B^i}{i!} = \begin{pmatrix} 1932396 & -9661979 \\ -9661979 & 4830992 \end{pmatrix}$$

$$\text{valore calcolato di } \sum_{i=0}^{45} \frac{B^i}{i!} = \begin{pmatrix} 19323630 & -9661887 \\ -9661887 & 4830939 \end{pmatrix}$$

Cambiando segno alla matrice B si ottiene la matrice

$$B' = \begin{pmatrix} -13.8 & 6.4 \\ 6.4 & -4.2 \end{pmatrix}$$

$$\text{valore esatto di } \frac{B'^{44}}{44!} = \begin{pmatrix} .4151864 & -.2075932 \\ -.2075932 & .1037966 \end{pmatrix}$$

$$\text{valore calcolato di } \frac{B'^{44}}{44!} = \begin{pmatrix} .415176 & -.2075887 \\ -.2075880 & .1037943 \end{pmatrix}$$

$$\text{valore esatto di } \sum_{i=0}^{44} \frac{B'^i}{i!} = \begin{pmatrix} .1879193 & .08998006 \\ .08998006 & .3228894 \end{pmatrix}$$

$$\text{valore calcolato di } \sum_{i=0}^{44} \frac{B'^i}{i!} = \begin{pmatrix} -.150905 & .2277483 \\ -.1107576 & .4610824 \end{pmatrix}.$$

$$k = 45$$

$$\text{valore esatto di } \frac{B'^{45}}{45!} = \begin{pmatrix} -.1568482 & .0784241 \\ .0784241 & .03921205 \end{pmatrix}$$

$$\text{valore calcolato di } \frac{B'^{45}}{45!} = \begin{pmatrix} -.1568442 & .07842231 \\ .07842207 & -.03921117 \end{pmatrix}$$

$$\text{valore esatto di } \sum_{i=0}^{45} \frac{B'^i}{i!} = \begin{pmatrix} .3107109 & .1684042 \\ .1684042 & .2836773 \end{pmatrix}$$

$$\text{valore calcolato di } \sum_{i=0}^{45} \frac{B'^i}{i!} = \begin{pmatrix} -.3077492 & .3061706 \\ -.3233558 & .4218712 \end{pmatrix}$$

Calcoliamo, adesso, l'esponenziale della matrice B' mediante il metodo scaling and squaring: nel caso presentato tale algoritmo risulta stabile, contrariamente a quello di Taylor.

Dividendo la matrice B' per la costante $m = 32$, si ottiene la matrice C,

$$C = \begin{pmatrix} -.43125 & .2 \\ .2 & -.13125 \end{pmatrix}$$

con norma minore di 1, a cui si può applicare il metodo di Taylor. Il risultato parziale $fl(T_9(C))$ non è affetto da un elevato errore,infatti

$$\text{valore esatto di } e^C = \begin{pmatrix} .664142 & .1525454 \\ .1525454 & .8929605 \end{pmatrix}$$

$$\text{valore calcolato di } fl(T_9(C)) = \begin{pmatrix} .664142 & .1525453 \\ .1525453 & .8929604 \end{pmatrix}.$$

Bibliografia.

[1] T. A. Bickart, Matrix exponential: approximation by truncated power series.
Proc. IEEE, 56 (1968).

[2] A. Frederik Fath, Evaluation of a Matrix Polinomial
IEEE Transactions on Automatic Control, April 1968.

[3] Gene H. Golub, C. Van Loan, Matrix computations.
The Jonhs Hopkins University Press.
Baltimore, Maryland, 1983

[4] Gene H. Golub, J. H. Wilkinson, Ill-conditioned eigensystems and the computation of the Jordan canonical form.
Siam J. Numer. Anal., Vol 18, 1976.

[5] C. Moler, C. Van Loan, Nineteen dubious ways to compute the exponential of a matrix.
Siam Review, vol 20, n. 4, October 1978.

[6] K. R. Rao, N. Ahmed, Evaluation of transition matrices.
IEEE Transaction Automatic Control, vol. AC-14, 1969.

[7] L. F. Shampine, M. K. Gordon, Computer solution of ordinary differential equations - The initial value problem.
W. H. Freeman and Co., san Francisco, 1975.

[8] C. Van Loan, The sensitivity of the matrix exponential.
Siam J. Numer. Anal., Vol 14, N. 6,December 1977.

[9] C. Van Loan, A note on the evaluation of matrix polynomials.
IEEE Transaction Automatic Control, vol. AC-24, no. 2, April 1979.

[10] R. C. Ward, Numerical computation of the matrix exponential with accuracy estimate.
SIAM J. Numer. Anal.,14 (1977), pp. 600-610.

[11] J. H. Wilkinson, Rounding errors in algebraic processes.
Prentice Hall, Englewood Cliffs, N. J., 1963.