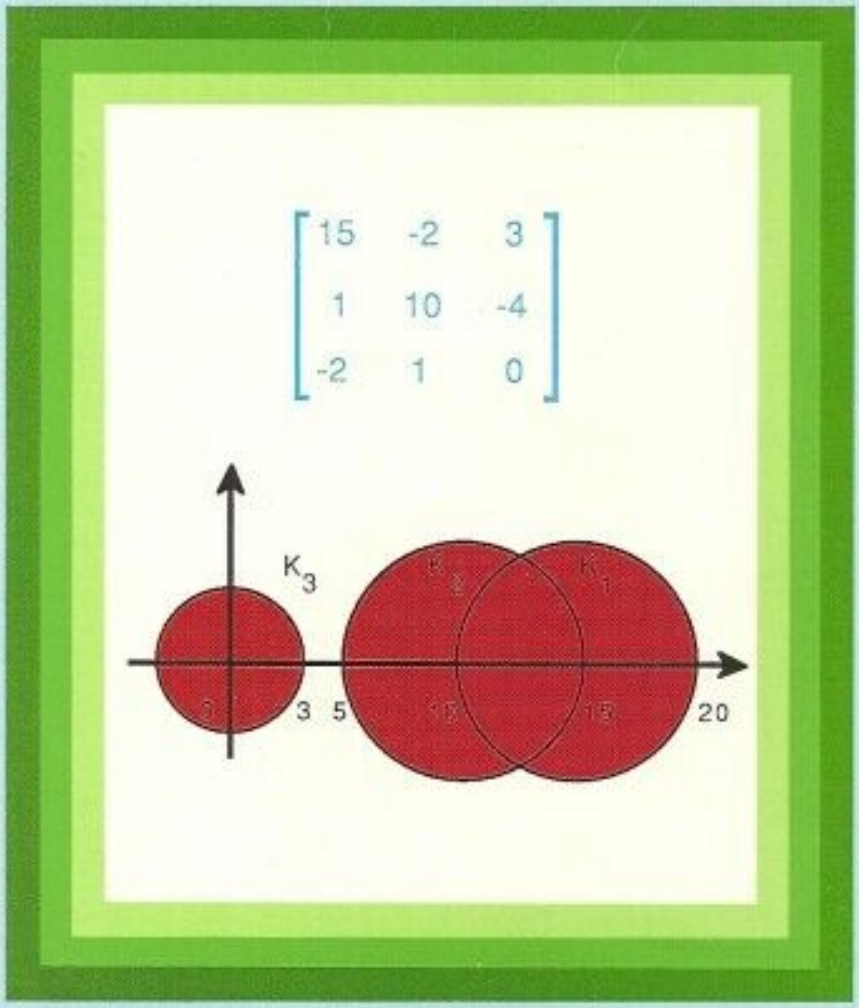
DARIO BINI MILVIO CAPOVANI
ORNELLA MENCHI

METODI NUMERICI PER L'ALGEBRA LINEARE

ZANICHELLI



Prima edizione: ottobre 1988
Ristampe:

6 5 4 3 2 1994 1995 1996

Il testo è stato composto dagli autori utilizzando un sistema MACINTOSH™ collegato ad una stampante LASER WRITER™, entrambi di produzione della Apple Computer Inc.

In copertina sono riportati, nel piano complesso, i cerchi di Gerschgorin K_i, $i = 1, 2, 3$, della matrice scritta nella parte superiore. Ciascun cerchio K_i ha per centro l'i-esimo elemento principale della matrice e per raggio la somma dei moduli degli elementi non principali della stessa riga. Per il teorema di Gerschgorin l'unione di tali cerchi contiene tutti gli autovalori della matrice.

Stampato a Bologna
dalla Tipostampa Bolognese, via Collemarini, 5/A,
per conto della Zanichelli Editore S. p. A.,
via Irnerio 34, 40126 Bologna.

INDICE

Capitolo 1 ELEMENTI DI ALGEBRA LINEARE

1. Matrici 1
2. Vettori 4
3. Matrici definite positive 10
4. Determinante 11
5. Matrice inversa 12
6. Sistemi lineari 13
7. Matrici a blocchi 16
8. Matrici riducibili 17

Esercizi proposti 21
Commento bibliografico 41
Bibliografia 43

Capitolo 2 AUTOVALORI E AUTOVETTORI

1. Definizioni 45
2. Proprietà degli autovalori 47
3. Proprietà degli autovettori 52
4. Trasformazioni per similitudine 55
5. Forme canoniche 59
6. Alcune proprietà delle matrici definite positive 73
7. Localizzazione degli autovalori 76
8. Predominanza diagonale 82

Esercizi proposti 83
Commento bibliografico 105
Bibliografia 107

Capitolo 3 NORME

1. Norme vettoriali 108
2. Norme matriciali 113
3. Alcune proprietà delle norme 118
4. Principali relazioni fra le norme matriciali 121

Esercizi proposti 122
Commento bibliografico 135
Bibliografia 135

Capitolo 4 METODI DIRETTI PER LA RISOLUZIONE DEI SISTEMI DI EQUAZIONI LINEARI

1. Analisi dell'errore 136
2. Sistemi lineari con matrice triangolare 141

3. Fattorizzazioni 143
4. Matrici elementari 149
5. Fattorizzazione mediante matrici elementari 154
6. Il metodo di Gauss per la fattorizzazione LU 157
7. Il metodo di Gauss per la risoluzione del sistema lineare . . . 159
8. Analisi dell'errore del metodo di Gauss 164
9. Massimo pivot 172
10. Implementazione del metodo di Gauss 178
11. Metodo di Gauss-Jordan 180
12. Metodo di Householder 182
13. Implementazione del metodo di Householder 185
14. Analisi dell'errore del metodo di Householder 187
15. Fattorizzazione QR con le matrici di Givens 191
16. Tecniche compatte 196
17. Metodo di Cholesky 198
18. Considerazioni sul costo computazionale 200
Esercizi proposti 203
Commento bibliografico 225
Bibliografia 228

Capitolo 5 METODI ITERATIVI PER LA RISOLUZIONE DEI SISTEMI DI EQUAZIONI LINEARI

1. Successioni di vettori e di matrici 231
2. Generalità sui metodi iterativi 235
3. Controllo della convergenza 238
4. Metodi iterativi di Jacobi e di Gauss-Seidel 242
5. Metodi di Jacobi e di Gauss-Seidel a blocchi 257
6. Metodi di rilassamento 261
7. Metodo del gradiente coniugato 272
Esercizi proposti 289
Commento bibliografico 313
Bibliografia 314

Capitolo 6 METODI PER IL CALCOLO DI AUTOVALORI E AUTOVETTORI

1. Teoremi di localizzazione 316
2. Teoremi di perturbazione 319
3. Caso delle matrici hermitiane 322
4. Introduzione ai metodi 331
5. Riduzione di una matrice hermitiana in forma tridiagonale: i metodi di Householder, di Givens e di Lanczos 333

6. Calcolo degli autovalori di matrici tridiagonali hermitiane con la successione di Sturm 343
7. Riduzione di una matrice in forma di Hessenberg superiore . . 349
8. Metodo QR per il calcolo degli autovalori 353
9. Metodo di Jacobi 367
10. Metodo delle potenze 371
11. Varianti del metodo delle potenze 378
12. Metodo delle iterazioni ortogonali 386
13. Metodo di Lanczos 391
Esercizi proposti 400
Commento bibliografico 427
Bibliografia 429

Capitolo 7 IL PROBLEMA LINEARE DEI MINIMI QUADRATI

1. Le equazioni normali 432
2. Metodo QR per il calcolo della soluzione del problema dei minimi quadrati 438
3. Norme di matrici non quadrate 443
4. Decomposizione ai valori singolari di una matrice 444
5. Risoluzione del problema dei minimi quadrati con i valori singolari 454
6. Pseudoinversa di Moore-Penrose 456
7. Condizionamento del problema dei minimi quadrati 459
8. Teoremi di perturbazione per i valori singolari 462
9. Calcolo della forma normale di Schur di $A^H A$ 466
10. Calcolo della soluzione di minima norma con il metodo del gradiente coniugato 474
11. Il metodo di Lanczos per il calcolo dei valori e dei vettori singolari 479
Esercizi proposti 481
Commento bibliografico 496
Bibliografia 498

Bibliografia generale 501

Indice analitico 503

INTRODUZIONE

L'algebra lineare, parte essenziale del bagaglio culturale di base richiesto in molti campi della matematica e più in generale della scienza e della tecnica, fornisce strumenti fondamentali per risolvere gran parte dei problemi scientifici e tecnici. Ad esempio la risoluzione di sistemi di equazioni lineari, di problemi lineari di minimi quadrati e il calcolo di autovalori e autovettori di matrici sono problemi tipici dell'algebra lineare che si incontrano nella risoluzione numerica di equazioni differenziali, in ottimizzazione, nell'analisi di sistemi discreti, in statistica e in altri problemi di matematica applicata.

Nelle applicazioni, quanto più il fenomeno esaminato è complesso, o il modello per descriverlo è raffinato e aderente al problema reale, maggiore è la quantità di dati necessari per descrivere il problema originale, e quindi maggiore è il numero di variabili nel conseguente problema algebrico. L'introduzione e la vasta utilizzazione dei calcolatori nel campo scientifico, ha consentito di trattare problemi di sempre maggiori dimensioni ed allo stesso tempo ha portato ad individuare e sviluppare metodi di risoluzione computazionalmente più efficienti. Questo ha imposto un grande sviluppo degli studi sull'analisi e la sintesi di metodi numerici per risolvere i problemi dell'algebra lineare, sviluppo che ha potenziato il settore di ricerca noto come "numerical linear algebra".

Grandi progressi sono stati fatti in questo settore dopo la metà degli anni cinquanta ad opera di A.S Householder e di J.H. Wilkinson; i principali risultati sono stati raccolti in modo sistematico in questi fondamentali trattati: *The Theory of Matrices in Numerical Analysis* di A. S. Householder (1964), *The Algebraic Eigenvalue Problem* di J. H. Wilkinson (1965). Un trattato più recente sui metodi numerici per problemi di algebra lineare è *Matrix Computation* di G. H. Golub e C. F. Van Loan (1983). Gli ultimi sviluppi sono relativi principalmente a metodi che utilizzano proprietà specifiche del problema algebrico (tipo struttura o sparsità) o che tengono conto dei diversi sistemi di calcolo (calcolatori sequenziali, vettoriali, paralleli) prodotti dalle nuove tecnologie.

In *Metodi numerici per l'algebra lineare* sono esposti i principali algoritmi, con un'analisi delle proprietà teoriche e computazionali. Nei primi tre capitoli sono presentati aspetti e risultati della teoria delle matrici fondamentali per trattare in modo sintetico e sistematico i metodi di risoluzione dei problemi dell'algebra lineare. Nei capitoli successivi sono presentati i metodi numerici per risolvere sistemi di equazioni lineari, per calcolare autovalori e autovettori e per risolvere il problema lineare dei minimi quadrati, con l'analisi della stabilità numerica e del costo computazionale. Le valu-

tazioni teoriche dell'errore generato dall'uso di una aritmetica in virgola mobile, come pure le stime teoriche della convergenza e del costo computazionale vengono confrontate con i valori effettivamente ottenuti eseguendo il metodo su di un calcolatore. I risultati sono stati raccolti in tabelle e grafici che, riportando tempi di esecuzione, errori generati, e, nel caso dei metodi iterativi, anche il numero di iterazioni, mettono a confronto i diversi metodi di risoluzione. Ogni capitolo è completato da esempi che evidenziano gli aspetti algoritmici e numerici, e da numerosi esercizi in alcuni dei quali sono riportati risultati teorici di carattere più avanzato che, anche se non centrali per un testo con finalità anche didattiche, sono di grande importanza. Ogni capitolo contiene anche un inquadramento storico e un'ampia bibliografia commentata degli argomenti trattati.

La sperimentazione numerica è stata effettuata presso l'istituto CNUCE del C.N.R., su un calcolatore IBM 3081K della serie /370, che opera con una rappresentazione interna dei numeri in base 16, utilizzando in generale la precisione semplice (6 cifre esadecimali, corrispondenti a circa 7 cifre decimali), e talvolta la precisione doppia (14 cifre esadecimali, corrispondenti a circa 16 cifre decimali).

Il testo è rivolto principalmente agli studenti dei corsi di laurea in matematica, scienze dell'informazione, fisica e ingegneria, ed ai ricercatori che operano nel settore del calcolo scientifico.

Capitolo 1

ELEMENTI DI ALGEBRA LINEARE

1. Matrici

Siano $\mathbf{C}$ e $\mathbf{R}$ rispettivamente il campo dei numeri complessi e il campo dei numeri reali. Sia inoltre $\mathbf{i}$ l'unità immaginaria tale che $\mathbf{i}^2 = -1$. Con $\mathbf{C}^{m\times n}$ si indica l'insieme delle matrici ad elementi complessi con m righe ed n colonne; in alcuni casi si fa esplicito riferimento al sottoinsieme $\mathbf{R}^{m\times n}$ delle matrici ad elementi reali. Se $A \in \mathbf{C}^{n\times n}$, si dice che A è una matrice *quadrata di ordine n*.

Generalmente le matrici vengono indicate con lettera maiuscola, mentre i loro elementi sono indicati con lettera minuscola seguita dai due indici (indice di *riga* e indice di *colonna*): ad esempio a_{ij} è elemento della matrice A. Si usa scrivere

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}.$$

Gli elementi a_{ij} tali che $i = j$ vengono detti elementi *diagonali* o *principali* di A e formano la *diagonale principale* di A.

Data una matrice $A \in \mathbf{C}^{m\times n}$, si definisce matrice *trasposta coniugata* di A la matrice $B \in \mathbf{C}^{n\times m}$ tale che

$$b_{ij} = \overline{a}_{ji},$$

dove $\overline{a}_{ji}$ è il coniugato del numero complesso a_{ji}, e si indica $B = A^H$. Se $A \in \mathbf{R}^{m\times n}$, la trasposta coniugata di A coincide con la matrice *trasposta* così definita

$$B = A^T, \qquad b_{ij} = a_{ji}.$$

Una matrice $A \in \mathbf{C}^{n\times n}$ è:

diagonale se $a_{ij} = 0$ per $i \neq j$;

scalare se è diagonale e $a_{ii} = \alpha \in \mathbf{C}$;

triangolare superiore (inferiore) se $a_{ij} = 0$ per $i > j$ (per $i < j$);

triangolare superiore (inferiore) in senso stretto se $a_{ij} = 0$ per $i \geq j$ (per $i \leq j$);

tridiagonale se $a_{ij} = 0$ per $|i - j| > 1$.

Le seguenti operazioni fra matrici:

addizione di matrici $(\mathbf{C}^{m\times n} \times \mathbf{C}^{m\times n} \to \mathbf{C}^{m\times n})$:

$$C = A + B, \qquad c_{ij} = a_{ij} + b_{ij},$$

moltiplicazione di un numero per una matrice $(\mathbf{C} \times \mathbf{C}^{m\times n} \to \mathbf{C}^{m\times n})$:

$$B = \alpha\, A, \qquad b_{ij} = \alpha\, a_{ij},$$

inducono su $\mathbf{C}^{m\times n}$ la struttura di *spazio vettoriale* su $\mathbf{C}$, in cui l'elemento *neutro* è la matrice con tutti gli elementi nulli, che viene indicata con $O_{m\times n}$ o semplicemente con O se dal contesto risulta chiaramente quali sono le dimensioni. Valgono le proprietà di associatività e commutatività per l'addizione e di distributività della moltiplicazione rispetto all'addizione.

Si definisce *prodotto righe per colonne* di due matrici $A \in \mathbf{C}^{m\times n}$ e $B \in \mathbf{C}^{n\times p}$ la matrice $C = A\,B \in \mathbf{C}^{m\times p}$, i cui elementi sono

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}$$

(si osservi che il numero di colonne di A è uguale al numero di righe di B).

La moltiplicazione fra matrici gode della proprietà associativa, di quella distributiva rispetto all'addizione, ma non di quella commutativa (si vedano gli esercizi 1.1 e 1.2). Vale inoltre la proprietà

$$(A\,B)^H = B^H A^H.$$

La matrice scalare di ordine n avente gli elementi principali uguali a 1 è detta matrice *identica* e viene indicata con I_n o semplicemente con I se dal contesto risulta chiaramente quale è l'ordine. Tale matrice verifica le relazioni

$$\left.\begin{array}{l} I_m A = A \\ A I_n = A \end{array}\right\} \qquad \text{per ogni matrice } \; A \in \mathbf{C}^{m\times n}.$$

Una matrice $A \in \mathbf{C}^{n\times n}$ si dice:

normale se $A^H A = A\,A^H$;

hermitiana se $A^H = A$;

unitaria se $A^H A = A\,A^H = I$.

Sia $A \in \mathbf{R}^{n\times n}$; se A è hermitiana, allora risulta $A^T = A$ e A è detta *simmetrica*; se A è unitaria, allora risulta $A^T A = A\,A^T = I$ e A è detta *ortogonale*.

1.1 Esempio. La matrice

$$G = \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix}, \qquad \phi \in \mathbf{R},$$

è unitaria. Infatti

$$G\,G^H = G^H G = \begin{bmatrix} \sin^2\phi + \cos^2\phi & 0 \\ 0 & \sin^2\phi + \cos^2\phi \end{bmatrix} = I.$$

Inoltre, poiché è reale, G è anche ortogonale. ∎

Particolarmente importanti fra le matrici ortogonali sono le *matrici di permutazione*, cioè matrici ottenute permutando le righe della matrice identica I. Le matrici di permutazione in ogni riga e ogni colonna hanno un solo elemento diverso da zero e uguale a 1.

Un sottoinsieme di $\mathbf{C}^{n\times n}$ si dice *chiuso rispetto all'operazione di moltiplicazione*, se date due matrici A e B appartenenti al sottoinsieme, anche il prodotto $A\,B$ appartiene al sottoinsieme. I seguenti sottoinsiemi di $\mathbf{C}^{n\times n}$ sono chiusi rispetto all'operazione di moltiplicazione:

- matrici triangolari superiori (inferiori),
- matrici triangolari superiori (inferiori) in senso stretto,
- matrici unitarie.

Data una matrice $A \in \mathbf{C}^{m\times n}$, una matrice $B \in \mathbf{C}^{k\times h}$, $0 \le k < m$, $0 \le h < n$, è detta *sottomatrice* di A se è ottenuta da A eliminando $m-k$ righe e $n-h$ colonne. Data una matrice $A \in \mathbf{C}^{n\times n}$, una sottomatrice quadrata B di ordine $k \le n$ di A è detta *principale* se gli elementi principali di B sono anche elementi principali di A (cioè le righe e le colonne di A che concorrono alla costruzione di B hanno medesimo indice). Una sottomatrice B principale di ordine k di A è detta *principale di testa* se è formata dagli elementi $a_{ij},\ i,\ j = 1,\ldots,k$.

1.2 Esempio. Si consideri la matrice $A \in \mathbf{R}^{3\times 3}$:

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$$

La matrice

$$\begin{bmatrix} 1 & 3 \\ 4 & 6 \end{bmatrix}$$

è sottomatrice di ordine 2 di A, la matrice

$$\begin{bmatrix} 1 & 3 \\ 7 & 9 \end{bmatrix}$$

è sottomatrice principale di ordine 2 di A, la matrice

$$\begin{bmatrix} 1 & 2 \\ 4 & 5 \end{bmatrix}$$

è sottomatrice principale di testa di ordine 2 di A. ∎

2. Vettori

Se $A \in \mathbf{C}^{m\times 1}(A \in \mathbf{C}^{1\times m})$, la matrice si riduce ad una sola colonna (riga) e viene detta *vettore colonna (riga) ad m elementi* o *componenti*.

Comunemente con il termine *vettore* si intende un vettore colonna e lo spazio vettoriale $\mathbf{C}^{m\times 1}$ dei vettori ad m componenti viene indicato con $\mathbf{C}^m$. Un vettore è generalmente indicato con lettera minuscola in grassetto e le singole componenti sono indicate con lettera minuscola seguita da un indice: ad esempio x_i è l'i-esima componente del vettore $\mathbf{x}$. Si usa scrivere

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} \quad \text{o anche} \quad \mathbf{x} = [x_1,\, x_2,\, \ldots,\, x_m]^T.$$

Si indica con $\mathbf{0}$ il vettore di componenti nulle. Se $\mathbf{x} \in \mathbf{C}^m$, allora $\mathbf{x}^H \in \mathbf{C}^{1\times m}$ è il vettore riga le cui componenti sono le coniugate di quelle di $\mathbf{x}$.

Casi particolari del prodotto righe per colonne di matrici:

prodotto di una matrice per un vettore $(\mathbf{C}^{m\times n} \times \mathbf{C}^n \to \mathbf{C}^m)$:

$$\mathbf{y} = A\mathbf{x}, \qquad y_i = \sum_{j=1}^{n} a_{ij}\, x_j, \qquad i = 1, \ldots, m;$$

prodotto interno fra vettori $(\mathbf{C}^m \times \mathbf{C}^m \to \mathbf{C})$:

$$\alpha = \mathbf{x}^H\mathbf{y}, \qquad \alpha = \sum_{i=1}^{n} \overline{x}_i\, y_i;$$

prodotto esterno fra vettori $(\mathbf{C}^m \times \mathbf{C}^{1\times n} \to \mathbf{C}^{m\times n})$:

$$A = \mathbf{x}\,\mathbf{y}^H, \qquad a_{ij} = x_i\overline{y}_j, \qquad i = 1, \ldots, m, \quad j = 1, \ldots, n.$$

Il vettore $\dfrac{1}{\alpha}\,\mathbf{x},\ \alpha \neq 0, \alpha \in \mathbf{C}$, viene talvolta indicato con $\dfrac{\mathbf{x}}{\alpha}$.

1.3 Esempio. Dati i vettori

$$\mathbf{x} = [1,\, \mathbf{i},\, -\mathbf{i}]^T \qquad \text{e} \qquad \mathbf{y} = [\mathbf{i},\, 1,\, \mathbf{i}]^T,$$

risulta

$$\mathbf{x}^H\mathbf{y} = -1,$$

$$\mathbf{x}\,\mathbf{y}^H = \begin{bmatrix} \text{-}\mathbf{i} & 1 & \text{-}\mathbf{i} \\ 1 & \mathbf{i} & 1 \\ \text{-}1 & \text{-}\mathbf{i} & \text{-}1 \end{bmatrix}. \qquad \blacksquare$$

Il prodotto interno fra vettori definisce un *prodotto scalare* su $\mathbf{C}^n$ e gode delle seguenti proprietà (si veda l'esercizio 1.26):

1. $\mathbf{x}^H\mathbf{x}$ è reale e non negativo, ed è nullo se e solo se $\mathbf{x} = \mathbf{0}$;
2. $\overline{\mathbf{x}^H\mathbf{y}} = \mathbf{y}^H\mathbf{x}$;
3. $\mathbf{x}^H(\alpha\mathbf{y}) = \alpha\,\mathbf{x}^H\mathbf{y} \qquad$ per $\alpha \in \mathbf{C}$;
4. $\mathbf{x}^H(\mathbf{y}+\mathbf{z}) = \mathbf{x}^H\mathbf{y} + \mathbf{x}^H\mathbf{z} \qquad$ per $\mathbf{z} \in \mathbf{C}^n$.

La quantità $\sqrt{\mathbf{x}^H\mathbf{x}}$ è la *lunghezza euclidea* del vettore $\mathbf{x}$, per cui il vettore $\dfrac{\mathbf{x}}{\sqrt{\mathbf{x}^H\mathbf{x}}}$ ha lunghezza 1. In $\mathbf{R}^n$, se $\mathbf{x}$ ha lunghezza 1, il prodotto $\mathbf{x}^H\mathbf{y}$ dà la *proiezione* di $\mathbf{y}$ sulla semiretta su cui giace il vettore $\mathbf{x}$. Poiché vale la disuguaglianza di *Cauchy-Schwarz*

$$|\mathbf{x}^H\mathbf{y}|^2 \le (\mathbf{x}^H\mathbf{x})\,(\mathbf{y}^H\mathbf{y}), \tag{1}$$

(si veda l'esercizio 1.30) è possibile definire l'*angolo* θ formato da due vettori $\mathbf{x}, \mathbf{y} \in \mathbf{R}^n$:

$$\theta = \arccos \frac{\mathbf{x}^H\mathbf{y}}{\sqrt{(\mathbf{x}^H\mathbf{x})\,(\mathbf{y}^H\mathbf{y})}}.$$

È facile verificare che in $\mathbf{R}^2$ e in $\mathbf{R}^3$ questa definizione corrisponde al concetto geometrico di angolo, come si può vedere nel caso di $\mathbf{R}^2$ nella figura 1.1.

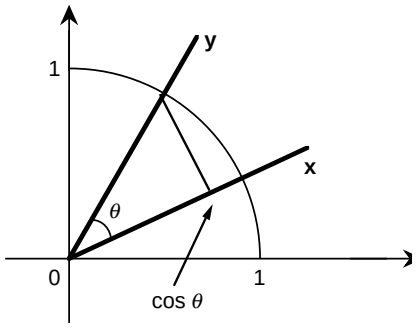


Fig.1.1 - Angolo fra due vettori.

Se $\mathbf{x}^H\mathbf{y} = 0$, i due vettori $\mathbf{x}$ e $\mathbf{y}$ sono detti *ortogonali*.

1.4 Definizione. I vettori $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbf{C}^m, n \leq m$, si dicono *linearmente indipendenti* se dalla condizione

$$\sum_{i=1}^{n} \alpha_i \, \mathbf{x}_i = \mathbf{0}, \quad \alpha_i \in \mathbf{C},$$

segue che

$$\alpha_i = 0, \quad i = 1, \ldots, n. \qquad \blacksquare$$

n vettori, che non sono linearmente indipendenti, si dicono *linearmente dipendenti*; in tal caso se $\alpha_k \neq 0$, si ha

$$\mathbf{x}_k = \sum_{\substack{i=1 \\ i \neq k}}^{n} \beta_i \, \mathbf{x}_i, \quad \text{dove} \quad \beta_i = -\frac{\alpha_i}{\alpha_k}, \quad i = 1, \ldots, n, \;\; i \neq k.$$

1.5 Definizione. Sia S un sottospazio di $\mathbf{C}^n$. k vettori $\mathbf{x}_1, \ldots, \mathbf{x}_k \in S$ costituiscono una *base* di S se ogni vettore $\mathbf{v} \in S$ può essere espresso, in modo unico, come combinazione lineare dei vettori della base

$$\mathbf{v} = \sum_{i=1}^{k} \alpha_i \, \mathbf{x}_i.$$

Si dice anche che S è *generato* dalla base $\mathbf{x}_1, \ldots, \mathbf{x}_k$. $\blacksquare$

Una base di $\mathbf{C}^n$ particolarmente importante è la cosiddetta *base canonica*, formata dai vettori

$$\mathbf{e}_i = [0, \ldots, 0, \underset{\substack{\uparrow \\ i}}{1}, 0, \ldots, 0]^T, \quad i = 1, \ldots, n,$$

che sono le colonne della matrice identica di ordine n.

I k vettori $\mathbf{x}_1, \ldots, \mathbf{x}_k$ di una base sono linearmente indipendenti; inoltre tutte le basi di un sottospazio hanno lo stesso numero di elementi, e tale numero, indicato con $\dim S$, è detto *dimensione* del sottospazio. Lo spazio $\mathbf{C}^n$, come spazio vettoriale sul campo $\mathbf{C}$, ha dimensione n, e ogni insieme di n vettori linearmente indipendenti di $\mathbf{C}^n$ costituisce una base di $\mathbf{C}^n$.

Siano S e T due sottospazi di $\mathbf{C}^n$. Allora la somma

$$S + T = \{\mathbf{s} + \mathbf{t}, \; \mathbf{s} \in S, \; \mathbf{t} \in T\}$$

e l'intersezione $S \cap T$ sono ancora sottospazi. Per le loro dimensioni vale la seguente relazione

$$\dim(S+T) = \dim S + \dim T - \dim(S \cap T), \tag{2}$$

da cui segue che

$$\max\{\dim S, \dim T\} \leq \dim(S+T) \leq \min\{\dim S + \dim T, n\}, \tag{3}$$

$$\max\{0, \dim S + \dim T - n\} \leq \dim(S \cap T) \leq \min\{\dim S, \dim T\}. \tag{4}$$

Se $S \cap T = \{\,\mathbf{0}\,\}$, il sottospazio $X = S + T$ è detto *somma diretta* di S e T e viene di solito indicato con $S \oplus T$. In tal caso

$$\dim X = \dim S + \dim T,$$

e gli elementi $\mathbf{x}$ di X possono essere espressi univocamente con la somma

$$\mathbf{x} = \mathbf{s} + \mathbf{t}, \quad \mathbf{s} \in S,\ \mathbf{t} \in T.$$

1.6 Definizione. Sia S un sottospazio di $\mathbf{C}^n$. Il sottospazio

$$S^{\perp} = \{\mathbf{u} \in \mathbf{C}^n \ :\ \mathbf{u}^H\mathbf{v} = 0 \text{ per ogni } \mathbf{v} \in S\}$$

è detto *sottospazio ortogonale* ad S. Valgono le seguenti relazioni

$$S \cap S^{\perp} = \{\mathbf{0}\},$$

$$S \oplus S^{\perp} = \mathbf{C}^n,$$

$$\dim S^{\perp} = n - \dim S.$$

Quindi ogni vettore $\mathbf{x} \in \mathbf{C}^n$ può essere espresso univocamente come somma

$$\mathbf{x} = \mathbf{s} + \mathbf{t},\ \mathbf{s} \in S,\ \mathbf{t} \in S^{\perp}. \tag{5}$$

Il vettore $\mathbf{s}$ è detto *proiezione ortogonale* di $\mathbf{x}$ su S. ∎

1.7 Esempio. In $\mathbf{R}^3$ sia S il sottospazio generato dal vettore

$$\mathbf{x}_1 = [0,\ 0,\ 1]^T.$$

S è quindi costituito da tutti i vettori le cui due prime componenti sono nulle, e la sua dimensione è 1. Lo spazio $S^{\perp}$, costituito dai vettori la cui terza componente è nulla, è quindi generato dai vettori

$$\mathbf{x}_2 = [1,\ 0,\ 0]^T \qquad \text{e} \qquad \mathbf{x}_3 = [0,\ 1,\ 0]^T$$

e la sua dimensione è 2. La figura 1.2 fornisce per questo caso l'interpretazione geometrica della relazione (5). ∎

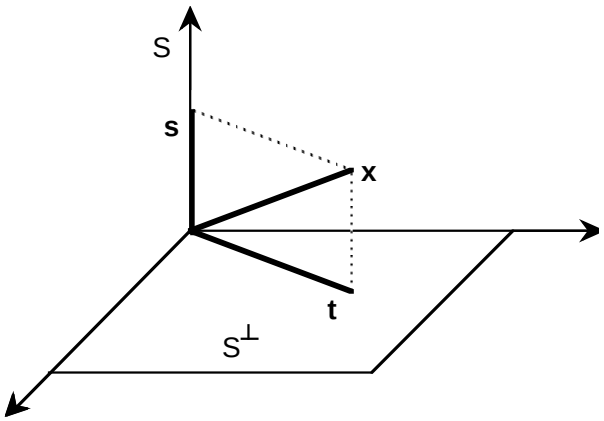


Fig. 1.2 - Proiezione ortogonale di $\mathbf{x}$ su S.

1.8 Definizione. n vettori non nulli $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbf{C}^m$ si dicono *ortogonali* se $\mathbf{x}_i^H \mathbf{x}_j = 0$ per $i \neq j$; si dicono *ortonormali* se sono ortogonali ed inoltre $\mathbf{x}_i^H \mathbf{x}_i = 1$, cioè se hanno lunghezza 1 o, come si dice, se sono *normalizzati*. In questo caso si usa anche la notazione

$$\mathbf{x}_i^H \mathbf{x}_j = \delta_{ij},$$

dove

$$\delta_{ij} = \begin{cases} 1 & \text{se } i = j, \\ 0 & \text{se } i \neq j, \end{cases}$$

è il *delta di Kronecker*. ∎

Si osservi che n vettori ortogonali sono anche linearmente indipendenti.

1.9 Esempio. I vettori

$$\mathbf{x} = [1,\, \mathbf{i},\, -\mathbf{i}]^T \qquad \text{e} \qquad \mathbf{y} = [\mathbf{i},\, 1,\, \mathbf{i}]^T,$$

dell'esempio 1.3 sono linearmente indipendenti, ma non ortogonali: infatti $\mathbf{x}^H \mathbf{y} = -1 \neq 0$. I vettori

$$\mathbf{u} = \frac{1}{\sqrt{3}}\, \mathbf{x} \qquad \text{e} \qquad \mathbf{v} = \frac{1}{\sqrt{8}}\, [-2\mathbf{i},\, -1 - \mathbf{i},\, 1 - \mathbf{i}]^T$$

sono ortonormali: infatti $\mathbf{u}^H \mathbf{u} = 1$, $\mathbf{v}^H \mathbf{v} = 1$ e $\mathbf{u}^H \mathbf{v} = 0$. Il vettore

$$\mathbf{z} = \mathbf{i}\mathbf{x} + \mathbf{y} = [2\mathbf{i},\, 0,\, \mathbf{i} + 1]^T$$

è combinazione lineare di $\mathbf{x}$ e $\mathbf{y}$: quindi i vettori $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$ sono linearmente dipendenti. ∎

Fra le diverse basi di $\mathbf{C}^n$ sono particolarmente importanti le *basi ortonormali*, cioè quelle in cui i vettori $\mathbf{x}_1, \ldots, \mathbf{x}_n$ sono ortonormali.

Se del sottospazio S di $\mathbf{C}^n$ è nota una base $\mathbf{x}_1, \ldots, \mathbf{x}_k$, è possibile costruire una base ortonormale $\mathbf{y}_1, \ldots, \mathbf{y}_k$ con il metodo di *ortogonalizzazione di Gram-Schmidt* basato sul seguente teorema.

1.10 Teorema. *Se $\mathbf{x}_1, \ldots, \mathbf{x}_k \in \mathbf{C}^n$, $k \leq n$, sono k vettori linearmente indipendenti, i vettori $\mathbf{y}_1, \ldots, \mathbf{y}_k$, così costruiti*

$$\mathbf{t}_1 = \mathbf{x}_1 \qquad \mathbf{y}_1 = \mathbf{t}_1 / \sqrt{\mathbf{t}_1^H \mathbf{t}_1},$$
$$\mathbf{t}_i = \mathbf{x}_i - \sum_{j=1}^{i-1} (\mathbf{y}_j^H \mathbf{x}_i)\, \mathbf{y}_j, \qquad \mathbf{y}_i = \mathbf{t}_i / \sqrt{\mathbf{t}_i^H \mathbf{t}_i}, \qquad i = 2, \ldots, k,$$

sono ortonormali.

Dim. I vettori $\mathbf{y}_i$ sono normalizzati. Per dimostrare l'ortogonalità si procede per induzione su k. Per $k = 2$, poiché

$$\mathbf{t}_2^H \mathbf{y}_1 = \mathbf{x}_2^H \mathbf{y}_1 - (\mathbf{x}_2^H \mathbf{y}_1)\, \mathbf{y}_1^H \mathbf{y}_1 = 0,$$

ne segue che $\mathbf{y}_2^H \mathbf{y}_1 = 0$. Per $k > 2$, supponendo che i vettori $\mathbf{y}_1, \ldots, \mathbf{y}_{k-1}$ siano ortonormali, si dimostra che $\mathbf{t}_k$ è ortogonale a $\mathbf{y}_1, \ldots, \mathbf{y}_{k-1}$. Infatti, poiché

$$\mathbf{y}_j^H \mathbf{y}_i = 0 \quad \text{per} \quad j,\ i \leq k-1,\ i \neq j,$$

risulta:

$$\begin{aligned} \mathbf{t}_k^H \mathbf{y}_i &= \mathbf{x}_k^H \mathbf{y}_i - \sum_{j=1}^{k-1} (\mathbf{x}_k^H \mathbf{y}_j)\, \mathbf{y}_j^H \mathbf{y}_i \\ &= \mathbf{x}_k^H \mathbf{y}_i - (\mathbf{x}_k^H \mathbf{y}_i)\, \mathbf{y}_i^H \mathbf{y}_i = 0. \end{aligned}$$

∎

1.11 Esempio. I vettori di $\mathbf{C}^n$

$$\mathbf{x}_i = [\ \underbrace{1, \ldots, 1}_{i \text{ componenti}}\ , 0, \ldots, 0]^T, \quad i = 1, \ldots, n,$$

costituiscono una base non ortonormale di $\mathbf{C}^n$. Applicando il metodo di Gram-Schmidt ai vettori $\mathbf{x}_i$, si ottengono i vettori $\mathbf{e}_i$, $i = 1, \ldots, n$, della base canonica di $\mathbf{C}^n$. I vettori di $\mathbf{C}^n$

$$\mathbf{x}_1 = \mathbf{e}_1 + \mathbf{e}_2, \quad \mathbf{x}_2 = \mathbf{e}_2 + \mathbf{e}_3, \quad \ldots, \quad \mathbf{x}_{n-1} = \mathbf{e}_{n-1} + \mathbf{e}_n, \quad \mathbf{x}_n = \mathbf{e}_n + \mathbf{e}_1$$

sono linearmente indipendenti. Applicando il metodo di Gram-Schmidt si ottengono i vettori

$$\begin{aligned}
\mathbf{y}_1 &= \frac{1}{\sqrt{2}}\,[1,\ 1,\ 0,\ \ldots,\ 0]^T,\\
\mathbf{y}_2 &= \frac{1}{\sqrt{2}\sqrt{3}}\,[1,\ -1,\ -2,\ 0,\ \ldots,\ 0]^T,\\
\mathbf{y}_3 &= \frac{1}{\sqrt{3}\sqrt{4}}\,[1,\ -1,\ 1,\ 3,\ 0,\ \ldots,\ 0]^T,\\
&\vdots\\
\mathbf{y}_{n-1} &= \frac{1}{\sqrt{n-1}\sqrt{n}}\,[1,\ -1,\ \ldots,\ (-1)^n,\ (-1)^n(n-1)]^T,\\
\mathbf{y}_n &= \frac{1}{\sqrt{n}}\,[1,\ -1,\ 1,\ \ldots,\ (-1)^n,\ (-1)^{n+1}]^T,
\end{aligned}$$

che costituiscono una base ortonormale di $\mathbf{C}^n$. ∎

3. Matrici definite positive

Se $A \in \mathbf{C}^{n\times n}$ è una matrice hermitiana, cioè $A = A^H$, e $\mathbf{x} \in \mathbf{C}^n$, il numero

$$\alpha = \mathbf{x}^H A\mathbf{x}$$

è reale. Infatti, poiché A è hermitiana, si ha:

$$\overline{\alpha} = \overline{\mathbf{x}^H A\mathbf{x}} = (\mathbf{x}^H A\mathbf{x})^H = \mathbf{x}^H A^H\mathbf{x} = \mathbf{x}^H A\mathbf{x} = \alpha.$$

1.12 Definizione. Sia $A \in \mathbf{C}^{n\times n}$ una matrice hermitiana. Se per qualsiasi $\mathbf{x} \in \mathbf{C}^n,\ \mathbf{x} \neq \mathbf{0}$, il numero reale $\alpha = \mathbf{x}^H A\mathbf{x}$ mantiene lo stesso segno, si dice che la matrice A è *definita in segno*, e in particolare:

se $\mathbf{x}^H A\mathbf{x} > 0$ $\quad$ A è *definita positiva*,
se $\mathbf{x}^H A\mathbf{x} \geq 0$ $\quad$ A è *semidefinita positiva*,
se $\mathbf{x}^H A\mathbf{x} \leq 0$ $\quad$ A è *semidefinita negativa*,
se $\mathbf{x}^H A\mathbf{x} < 0$ $\quad$ A è *definita negativa*. ∎

1.13 Esempio. La matrice hermitiana

$$A = \begin{bmatrix} 3 & \mathbf{i} \\ \mathbf{-i} & 3 \end{bmatrix}$$

è definita positiva. Infatti per ogni $\mathbf{x} = [x_1\ ,\ x_2]^T \neq \mathbf{0}$ risulta:

$$\mathbf{x}^H A\mathbf{x} = |x_1 - \mathbf{i}x_2|^2 + 2|x_2 - \mathbf{i}x_1|^2 > 0.$$ ∎

1.14 Teorema. *Se una matrice $A \in \mathbf{C}^{n\times n}$ è definita positiva, anche tutte le sue sottomatrici principali sono definite positive.*

Dim. Sia B una sottomatrice principale di A ottenuta eliminando $(n-i)$ righe e le corrispondenti $(n-i)$ colonne. Per ogni vettore $\mathbf{x} \in \mathbf{C}^i$, $\mathbf{x} \neq 0$, si consideri il vettore $\mathbf{y} \in \mathbf{C}^n$ che ha nulli gli elementi con indici uguali a quelli delle colonne soppresse e i rimanenti elementi uguali ai corrispondenti elementi di $\mathbf{x}$. Allora, poiché A è definita positiva, si ha

$$\mathbf{x}^H B\mathbf{x} = \mathbf{y}^H A\mathbf{y} > 0. \qquad \blacksquare$$

Poiché le sottomatrici principali di ordine 1 sono formate da un solo elemento principale, ne segue che in una matrice definita positiva tutti gli elementi principali, oltre a essere reali perché la matrice è hermitiana, sono positivi.

4. Determinante

1.15 Definizione. Sia $A \in \mathbf{C}^{n\times n}$. Si definisce *determinante* di A il numero

$$\det A = \sum_{\pi\in P} \operatorname{sgn}(\pi)\, a_{1,\pi_1}\, a_{2,\pi_2} \dots a_{n,\pi_n}$$

dove P è l'insieme degli $n!$ vettori $\pi = [\pi_1, \pi_2, \dots, \pi_n]^T$, ottenuti permutando in tutti i modi possibili le componenti del vettore $[1, 2, \dots, n]^T$; il fattore $\operatorname{sgn}(\pi)$ vale $+1$ o -1 a seconda che sia pari o dispari il numero degli scambi necessari per portare il vettore $[1, 2, \dots, n]^T$ nel vettore π. $\blacksquare$

Il determinante di una matrice può essere più semplicemente espresso utilizzando la *regola di Laplace*. Indicata con A_{ij} la sottomatrice quadrata di ordine $n-1$ ottenuta dalla matrice A eliminando la i-esima riga e la j-esima colonna, per un qualunque indice di riga i si ha:

$$\det A = \begin{cases} a_{11} \quad \text{se } n = 1, \\ \displaystyle\sum_{j=1}^{n} (-1)^{i+j} a_{ij} \det A_{ij} \quad \text{se } n > 1. \end{cases} \tag{6}$$

Siano A, $B \in \mathbf{C}^{n\times n}$, $\alpha \in \mathbf{C}$; valgono le seguenti proprietà :

$\det A = \prod_{i=1}^{n} a_{ii}$ se A è diagonale o triangolare;

$\det I = 1$;

$\det A^T = \det A$

$\det A^H = \overline{\det A}$;

$\det(A\,B) = \det A \det B$ (regola di *Binet*);

$\det B = \alpha \det A$, se B è ottenuta da A moltiplicando per α una riga (o una colonna);

$\det(\alpha A) = \alpha^n \det A$;

$\det B = -\det A$, se B è ottenuta da A scambiando fra loro due righe (o colonne);

$\det B = \det A$, se B è ottenuta da A aggiungendo ad una riga (o colonna) un'altra riga (o colonna) moltiplicata per un numero;

$\det A = 0$, se due o più righe (o colonne) di A sono linearmente dipendenti.

Poiché $\det A = \det A^T$, la regola di Laplace per il calcolo del determinante di A può essere applicata sommando nella (6) rispetto all'indice di riga i.

5. Matrice inversa

1.16 Definizioni. Sia $A \in \mathbf{C}^{n\times n}$, si definiscono:

matrice inversa di A una matrice $B \in \mathbf{C}^{n\times n}$ tale che

$$AB = BA = I,$$

matrice aggiunta di A la matrice $\mathrm{adj}A \in \mathbf{C}^{n\times n}$, il cui elemento (i,j)-esimo è dato da

$$(-1)^{i+j} \det A_{ji},$$

dove A_{ji} è la sottomatrice ottenuta da A cancellando la j-esima riga e la i-esima colonna. ∎

Una matrice A per cui non esiste la matrice inversa è detta *singolare*. Poiché vale la relazione (si veda l'esercizio 1.48)

$$A\,\mathrm{adj}A = (\det A)I,$$

ne segue che A è non singolare se e solo se $\det A \neq 0$, quindi se A è non singolare, la matrice inversa, che viene indicata con A^{-1}, è unica ed è

$$A^{-1} = \frac{1}{\det A}\,\mathrm{adj}A.$$

Valgono le seguenti proprietà:

$(A^H)^{-1} = (A^{-1})^H$ (si indica anche con A^{-H});

$A^{-1} = A^H$ se A è unitaria, cioè tale che $A^H A = AA^H = I$;

$\det A^{-1} = 1/\det A$;

$(A\,B)^{-1} = B^{-1}A^{-1}$.

I seguenti sottoinsiemi di $\mathbf{C}^{n\times n}$ sono *chiusi rispetto all'operazione di inversione*, cioè se A è una matrice non singolare appartenente al sottoinsieme, anche A^{-1} appartiene al sottoinsieme:

- matrici hermitiane,
- matrici unitarie,
- matrici normali,
- matrici definite positive (negative),
- matrici triangolari superiori (inferiori),
- matrici diagonali.

6. Sistemi lineari

Sia $A \in \mathbf{C}^{m\times n}$ e si considerino i seguenti sottospazi

$$S(A) = \{\mathbf{y} \in \mathbf{C}^m \ : \ \mathbf{y} = A\mathbf{x},\ \mathbf{x} \in \mathbf{C}^n\},$$

detto *immagine* di A e

$$N(A) = \{\mathbf{x} \in \mathbf{C}^n \ : \ A\mathbf{x} = \mathbf{0}\},$$

detto *nucleo* di A. Si può dimostrare che

$$S(A)^\perp = N(A^H),$$

e quindi

$$\dim S(A) + \dim N(A^H) = m.$$

Il numero $\dim S(A)$ viene detto *rango* di A ed è uguale al numero delle righe (e delle colonne, si veda l'esercizio 1.35) linearmente indipendenti di A. Poiché il rango di A e il rango di A^H sono uguali, risulta

$$\dim S(A) + \dim N(A) = n. \tag{7}$$

Più in generale, se T è un sottospazio di $\mathbf{C}^n$, posto

$$S_T(A) = \{\mathbf{y} \in \mathbf{C}^m \ : \ \mathbf{y} = A\mathbf{x},\ \mathbf{x} \in T\},$$

$$N_T(A) = \{\mathbf{x} \in T \ : \ A\mathbf{x} = \mathbf{0}\} = N(A) \cap T,$$

si ha

$$\dim S_T(A) + \dim N_T(A) = \dim T.$$

1.17 Esempio. Sia

$$A = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & \text{-}1 & 1 & \text{-}1 \\ 1 & 0 & 1 & 0 \end{bmatrix}.$$

Il sottospazio $S(A)$ è quello generato dai vettori

$$\mathbf{y}_1 = [1,\, 1,\, 1]^T \quad \text{e} \quad \mathbf{y}_2 = [1,\, -1,\, 0]^T,$$

e quindi

$$\text{rango di } A = \dim S(A) = 2.$$

Il nucleo di A è il sottospazio generato dai vettori

$$\mathbf{x}_1 = [1,\, 0,\, -1,\, 0]^T \quad \text{e} \quad \mathbf{x}_2 = [0,\, 1,\, 0,\, -1]^T,$$

e quindi

$$\dim N(A) = 2.$$

Il sottospazio $S(A^T)$ è quello generato dai vettori

$$\mathbf{x}_3 = [1,\, 1,\, 1,\, 1]^T \quad \text{e} \quad \mathbf{x}_4 = [1,\, -1,\, 1,\, -1]^T,$$

e infatti

$$\text{rango di } A^T = \dim S(A^T) = \dim S(A) = 2.$$

Il nucleo di A^T è il sottospazio generato dal vettore

$$\mathbf{y}_3 = [1,\, 1,\, -2]^T$$

e infatti

$$\dim N(A^T) = 1. \qquad \blacksquare$$

1.18 Esempio. Se $\mathbf{x},\, \mathbf{y} \in \mathbf{C}^n, \mathbf{x},\, \mathbf{y} \neq \mathbf{0}$, la matrice (detta *diade*)

$$A = \mathbf{x}\,\mathbf{y}^H$$

ha rango 1. Infatti le colonne di A sono i vettori

$$\overline{y}_1\,\mathbf{x},\, \overline{y}_2\,\mathbf{x},\, \ldots,\, \overline{y}_n\,\mathbf{x},$$

che sono a due a due linearmente dipendenti. ∎

Se $m = n$, A è non singolare se e solo se rango di $A = n$, e dalla (7) segue che A è non singolare se e solo

$$\dim N(A) = 0,$$

cioè il nucleo di A è costituito dal solo vettore nullo.

Se rango di $A = r = \min\{m, n\}$, allora la matrice A si dice di *rango massimo*. In tal caso la matrice $A^H A \in \mathbf{C}^{n\times n}$ ha rango r e se $r = n$, la matrice $A^H A$ è non singolare. Viceversa se $A^H A$ è non singolare, allora $m \geq n$ e il rango di A è massimo.

1.19 Definizione. Siano $A \in \mathbf{C}^{m\times n}$, $\mathbf{b} \in \mathbf{C}^m$; si definisce *sistema lineare di m equazioni in n incognite* il sistema

$$A\mathbf{x} = \mathbf{b}, \tag{8}$$

dove $\mathbf{x} \in \mathbf{C}^n$ è il vettore delle incognite, A è la *matrice del sistema* e $\mathbf{b}$ è il *vettore dei termini noti*. Il sistema si dice *consistente* se ha almeno una soluzione. ∎

1.20 Teorema. *Le seguenti condizioni sono equivalenti*:

a) *il sistema (8) è consistente*,

b) $\mathbf{b} \in S(A)$,

c) *la matrice A e la matrice $[A|\mathbf{b}]$, ottenuta aggiungendo il vettore $\mathbf{b}$ alle colonne di A, hanno lo stesso rango.* ∎

Se il sistema (8) è consistente e $\mathbf{x}$ è una sua soluzione, allora ogni soluzione di (8) può essere espressa come $\mathbf{x} + \mathbf{y}$, dove $\mathbf{y}$ è tale che $A\mathbf{y} = \mathbf{0}$, cioè $\mathbf{y} \in N(A)$. Perciò la soluzione è unica se e solo se $\dim N(A) = 0$.

Vi sono vari casi possibili:

1. Se $n = m$, e la matrice A è non singolare, allora $S(A) = \mathbf{C}^n$ e $N(A) = \{\mathbf{0}\}$. Quindi il sistema è consistente, la soluzione è unica ed è data da

$$\mathbf{x} = A^{-1}\mathbf{b},$$

e mediante la *regola di Cramer* può essere così espressa

$$x_i = \frac{\det A_i}{\det A}, \quad i = 1, \ldots, n,$$

dove A_i è la matrice ottenuta da A sostituendo alla i-esima colonna il vettore $\mathbf{b}$. Se $\mathbf{b} = \mathbf{0}$ (sistema *omogeneo*), il sistema ha la sola soluzione nulla.

Se invece la matrice A è singolare, il sistema può non essere consistente. Il sistema è comunque consistente se è omogeneo, perché aggiungendo il vettore $\mathbf{b}$ nullo alla matrice A si ottiene una matrice con lo stesso rango.

2. Se $m < n$, cioè vi sono più incognite che equazioni, il sistema, se è consistente, ha infinite soluzioni in quanto $\dim S(A) \leq m$ e quindi $\dim N(A) \geq n - m > 0$.

3. Se $n < m$, cioè vi sono più equazioni che incognite, il sistema può essere consistente solo se vi sono almeno $m - n$ equazioni che sono combinazioni lineari delle altre.

7. Matrici a blocchi

Spesso è conveniente descrivere una matrice in termini di sue sottomatrici anziché in termini dei suoi elementi. Ad esempio la matrice

$$A = \begin{bmatrix} 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \end{bmatrix}$$

può essere così descritta in modo più compatto

$$A = \begin{bmatrix} I_2 & E \\ E^T & I_3 \end{bmatrix},$$

dove $E \in \mathbf{R}^{2\times 3}$ è la matrice

$$E = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}.$$

Si dice allora che la matrice A è *partizionata a blocchi* o anche che A è una matrice 2×2 a blocchi. In generale una matrice $p\times q$ a blocchi è una matrice della forma

$$A = \begin{bmatrix} A_{11} & A_{12} & \dots & A_{1q} \\ A_{21} & A_{22} & \dots & A_{2q} \\ \vdots & \vdots & & \vdots \\ A_{p1} & A_{p2} & \dots & A_{pq} \end{bmatrix},$$

dove $A_{ij} \in \mathbf{C}^{m_i\times n_j}$, e m_i, n_j sono interi positivi, per $i = 1,\dots,p$, $j = 1,\dots,q$, e quindi $A \in \mathbf{C}^{m\times n}$, con

$$m = \sum_{i=1}^{p} m_i, \quad n = \sum_{j=1}^{q} n_j.$$

Un caso frequente è quello in cui alcuni blocchi sono vettori riga o colonna, come ad esempio per la matrice $A \in \mathbf{C}^{n\times n}$

$$A = \begin{bmatrix} \alpha & \mathbf{v}^H \\ \mathbf{u} & B \end{bmatrix}$$

dove $\alpha \in \mathbf{C}$, $\mathbf{u}$, $\mathbf{v} \in \mathbf{C}^{n-1}$, $B \in \mathbf{C}^{(n-1)\times(n-1)}$.

Alle matrici a blocchi si possono estendere molte delle definizioni date nei precedenti paragrafi. Ad esempio la matrice

$$A = \begin{bmatrix} A_{11} & O & O \\ A_{21} & A_{22} & O \\ A_{31} & A_{32} & A_{33} \end{bmatrix},$$

è detta *triangolare inferiore a blocchi*. L'operazione di moltiplicazione fra due matrici A e B a blocchi può essere descritta in termini di prodotti righe per colonne di blocchi. Ad esempio, se

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \quad B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}, \quad C = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix},$$

sono matrici 2×2 a blocchi tali che $C = AB$, allora vale

$$C_{ij} = A_{i1}B_{1j} + A_{i2}B_{2j}, \quad i, j = 1, 2.$$

Tale proprietà vale nel caso generale di matrici a blocchi, purché il numero dei blocchi e le loro dimensioni siano compatibili.

8. Matrici riducibili

1.21 Definizione. Una matrice A di ordine $n \geq 2$ si dice *riducibile* se esiste una matrice di permutazione Π e un intero k, $0 < k < n$, tale che

$$B = \Pi A \Pi^T = \begin{bmatrix} A_{11} & A_{12} \\ O & A_{22} \end{bmatrix} \begin{matrix} \} \quad k \text{ righe} \\ \} \; n-k \text{ righe} \end{matrix} \tag{9}$$

in cui $A_{11} \in \mathbf{C}^{k\times k}$ e $A_{22} \in \mathbf{C}^{(n-k)\times(n-k)}$. Se la matrice A non è riducibile, si dice che A è *irriducibile*. ∎

Se una matrice A è riducibile, possono esistere più matrici di permutazione Π che consentono di trasformare la matrice A in una matrice B

della forma (9). Se la matrice A del sistema lineare (8) è riducibile, poiché la matrice Π in (9) è ortogonale, risulta

$$\Pi A \Pi^T \Pi \mathbf{x} = \Pi \mathbf{b},$$

e ponendo $\mathbf{y} = \Pi \mathbf{x}$ e $\mathbf{c} = \Pi \mathbf{b}$, si ha

$$B\mathbf{y} = \mathbf{c}.$$

Partizionando i vettori

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \\ \mathbf{y}_2 \end{bmatrix} \begin{matrix} \} \quad k \text{ componenti} \\ \\ \} \; n-k \text{ componenti} \end{matrix} \qquad \mathbf{c} = \begin{bmatrix} \mathbf{c}_1 \\ \\ \mathbf{c}_2 \end{bmatrix} \begin{matrix} \} \quad k \text{ componenti} \\ \\ \} \; n-k \text{ componenti} \end{matrix}$$

dove $\mathbf{y}_1$, $\mathbf{c}_1 \in \mathbf{C}^k, \mathbf{y}_2$, $\mathbf{c}_2 \in \mathbf{C}^{n-k}$, il sistema (8) si può scrivere nella forma

$$\begin{cases} A_{11}\,\mathbf{y}_1 + A_{12}\,\mathbf{y}_2 = \mathbf{c}_1 \\ \\ \quad A_{22}\,\mathbf{y}_2 = \mathbf{c}_2. \end{cases}$$

La risoluzione del sistema (8) con matrice dei coefficienti di ordine n è così ricondotta alla risoluzione di due sistemi, il primo con matrice dei coefficienti di ordine $n-k$, il secondo con matrice dei coefficienti di ordine k.

Per determinare se una matrice A è riducibile, si può utilizzare il *grafo orientato* associato ad A, cioè un grafo che ha tanti nodi p_i, quant'è l'ordine di A, ed ha un arco orientato da p_i (nodo di partenza) a p_j (nodo di arrivo) per ogni elemento a_{ij} non nullo di A.

1.22 Esempio. Il grafo associato alla matrice

$$A = \begin{bmatrix} 1 & 3 & 0 \\ 0 & 2 & -1 \\ -1 & 0 & 2 \end{bmatrix} \tag{10}$$

è riportato nella figura 1.3. ∎

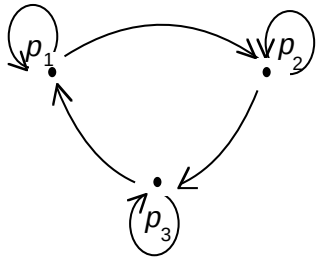


Fig. 1.3 - Grafo orientato associato alla matrice (10).

Due archi di un grafo orientato si dicono *contigui* se il nodo di arrivo del primo è il nodo di partenza del secondo. Una successione di archi orientati contigui si dice *cammino orientato*. Un cammino orientato si dice *chiuso* se il nodo di partenza del primo arco del cammino coincide con il nodo di arrivo dell'ultimo arco.

1.23 Definizione. Un grafo orientato si dice *fortemente connesso* se per ogni coppia di indici i, j, $1 \leq i, j \leq n$, con $i \neq j$, esiste un cammino orientato che parte dal nodo p_i e arriva al nodo p_j. ∎

1.24 Teorema. *Una matrice A è riducibile se e solo se il suo grafo orientato non è fortemente connesso.*

Dim. Si noti che il grafo orientato della matrice A e quello della matrice $B = \Pi A \Pi^T$ differiscono solamente per la diversa numerazione degli indici dei nodi p_i. Se la matrice A è riducibile, considerando la matrice B della (9) e un indice i, con $k < i \leq n$, risulta che non vi può essere alcun cammino che partendo dal nodo p_i arrivi ad un nodo p_j, $j \leq k$. Viceversa, se il grafo di A non è fortemente connesso, esiste un nodo p_j a partire dal quale non è possibile raggiungere almeno un altro nodo del grafo. Si indica con $\mathcal{P}$ l'insieme dei nodi che sono raggiungibili a partire da p_j e con $\mathcal{Q}$ l'insieme dei nodi che non sono raggiungibili a partire da p_j. Gli insiemi $\mathcal{P}$ e $\mathcal{Q}$ costituiscono una partizione dell'insieme dei nodi e $\mathcal{Q}$ è non vuoto. Inoltre non esistono cammini orientati che partendo da un nodo di $\mathcal{P}$ raggiungano nodi di $\mathcal{Q}$. Si riordinano i nodi, in modo tale che $\mathcal{Q} = \{p_1, p_2, \ldots, p_k\}$, con $k \geq 1$, e $\mathcal{P} = \{p_{k+1}, \ldots, p_n\}$. La matrice B ottenuta permutando conseguentemente righe e colonne di A è tale che $b_{ij} = 0$ se $i > k$ e $j \leq k$. ∎

Dal teorema 1.24 segue che se la matrice A è irriducibile, allora esiste un cammino orientato chiuso che tocca tutti i nodi del grafo.

1.25 Esempio. Dato il sistema lineare $A\mathbf{x} = \mathbf{b}$, dove

$$A = \begin{bmatrix} 1 & 0 & -1 & 0 \\ 2 & 3 & -2 & 1 \\ -1 & 0 & -2 & 0 \\ 1 & -1 & 1 & 4 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 1 \\ -2 \\ -1 \\ -2 \end{bmatrix}, \tag{11}$$

per verificare se la matrice A è riducibile, se ne disegna il grafo orientato, riportato in figura 1.4:

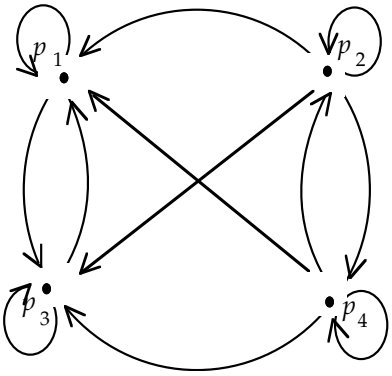


Fig. 1.4 - Grafo orientato associato alla matrice (11).

In questo grafo

al nodo p_1 arrivano cammini orientati provenienti dai nodi p_1, p_2, p_3, p_4;

al nodo p_2 arrivano cammini orientati provenienti dai nodi p_2, p_4;

al nodo p_3 arrivano cammini orientati provenienti dai nodi p_1, p_2, p_3, p_4;

al nodo p_4 arrivano cammini orientati provenienti dai nodi p_2, p_4.

Ne segue che scambiando fra loro i nodi p_1 e p_4, e mettendo così in testa i due nodi a cui non si arriva provenendo dagli altri due, la matrice di permutazione associata

$$\Pi = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix},$$

è tale che risulta

$$B = \Pi A \Pi^T = \begin{bmatrix} 4 & -1 & 1 & 1 \\ 1 & 3 & -2 & 2 \\ 0 & 0 & -2 & -1 \\ 0 & 0 & -1 & 1 \end{bmatrix}, \quad \mathbf{c} = \begin{bmatrix} -2 \\ -2 \\ -1 \\ 1 \end{bmatrix}.$$

Risolvendo allora i due sistemi

$$\begin{bmatrix} -2 & -1 \\ -1 & 1 \end{bmatrix} \mathbf{y}_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

e

$$\begin{bmatrix} 4 & -1 \\ 1 & 3 \end{bmatrix} \mathbf{y}_1 = \begin{bmatrix} -2 \\ -2 \end{bmatrix} - \begin{bmatrix} 1 & 1 \\ -2 & 2 \end{bmatrix} \mathbf{y}_2,$$

si ottiene la soluzione

$$\mathbf{y} = [-1, -1, 0, 1]^T,$$

da cui

$$\mathbf{x} = \Pi^T \mathbf{y} = [1, -1, 0, -1]^T. \qquad \blacksquare$$

Esercizi proposti

1.1 Si costruiscano due matrici A e $B \in \mathbf{C}^{2\times 2}$ tali che

$$AB \neq BA \neq A^H B \neq AB^H \neq A^H B^H \neq B^H A^H.$$

Si verifichi che

$$\det AB = \det BA = \det A \det B,$$
$$\det A^H B^H = \det B^H A^H = \overline{\det AB} = \overline{\det A}\ \ \overline{\det B}.$$

1.2 Si dice che due matrici A e $B \in \mathbf{C}^{n\times n}$ *commutano* se $AB = BA$.

a) Si costruiscano due matrici A e $B \in \mathbf{C}^{2\times 2}$ che commutano;

b) si dimostri che A e B commutano se A e B sono diagonali;

c) si dimostri che la relazione

$$(A+B)\ (A-B) = A^2 - B^2$$

vale se e solo se A e $B \in \mathbf{C}^{n\times n}$ commutano;

d) si dimostri che se $A \in \mathbf{C}^{n\times n}$ le due matrici $B = A^i$ e $C = A^j,\ i, j$ interi positivi, commutano;

e) si verifichi che le sole matrici di $\mathbf{C}^{n\times n}$ che commutano con ogni matrice di $\mathbf{C}^{n\times n}$ sono le matrici scalari;

f) si verifichi che se A e B commutano, allora anche

A^{-1}	e	B^{-1},	se A e B sono non singolari
A^H	e	B^H	
A^i	e	B^j,	$i,\ j$ interi positivi
A^{-1}	e	B,	se A è non singolare
A	e	B^{-1},	se B è non singolare

commutano.

1.3 Sia

$$p(x) = \alpha_0 x^k + \alpha_1 x^{k-1} + \ldots + \alpha_k, \quad \alpha_0, \alpha_1, \ldots, \alpha_k \in \mathbf{C}$$

un polinomio di grado k. Si definisce *polinomio della matrice* A la matrice

$$p(A) = \alpha_0 A^k + \alpha_1 A^{k-1} + \ldots + \alpha_k I.$$

a) Si dimostri che tutti i polinomi della matrice A commutano;

b) si dimostri che se $p(x) = q(x)s(x)$, in cui $p(x), q(x)$ ed $s(x)$ sono polinomi, allora

$$p(A) = q(A)s(A),$$

e in particolare, se

$$p(x) = \prod_{i=1}^{k}(x - x_i), \quad x_i \in \mathbf{C},$$

allora

$$p(A) = \prod_{i=1}^{k}(A - x_i I),$$

e quindi le matrici scalari $x_i I$, $i = 1, 2, \ldots, k$, possono essere considerate come *zeri* del polinomio $p(A)$.

c) Per i polinomi di matrici non vale un teorema analogo al teorema fondamentale dell'algebra, per cui ogni polinomio di grado k ha in $\mathbf{C}$ esattamente k zeri, se contati con la loro molteplicità. In $\mathbf{C}^{n\times n}$ infatti, l'uguaglianza $AB = 0$ può essere verificata anche da matrici A e $B \neq 0$. Si verifichi, ad esempio, che in $\mathbf{C}^{2\times 2}$ il polinomio

$$p(A) = A^2 - I$$

ha come zeri, oltre ad $A = I$ e $A = -I$, anche tutte le matrici della forma

$$\begin{bmatrix} a & b \\ c & -a \end{bmatrix}, \qquad \text{in cui} \quad a^2 + bc = 1.$$

Le matrici $A \in \mathbf{C}^{n\times n}$ tali che $A^2 = I$ sono dette matrici *involutorie*.

d) Si verifichi che la matrice unitaria G dell'esempio 1.1 non è involutoria, mentre la matrice

$$H = \begin{bmatrix} \cos\phi & \sin\phi \\ \sin\phi & -\cos\phi \end{bmatrix}, \quad \phi \in \mathbf{R}$$

è unitaria e involutoria.

e) Si verifichi che la matrice $A \in \mathbf{R}^{n\times n}$ i cui elementi sono

$$a_{ij} = \begin{cases} (-1)^{j-1}\binom{j-1}{i-1} & \text{per } i < j, \\ (-1)^{i-1} & \text{per } i = j, \\ 0 & \text{per } i > j, \end{cases}$$

è involutoria.

(Traccia: e) posto $B = A^2$, per $i > j$ è $b_{ij} = 0$, per $i = j$ è $b_{ij} = 1$ e per $i < j$ è

$$b_{ij} = \sum_{k=i}^{j} (-1)^{k+j} \binom{k-1}{i-1} \binom{j-1}{k-1}$$
$$= (-1)^{j+i} \frac{(j-1)!}{(i-1)!(j-i)!} \sum_{k=i}^{j} (-1)^{k-i} \frac{(j-i)!}{(k-i)!(j-k)!}$$

e ponendo $h = k - i$ e $r = j - i$, la sommatoria risulta uguale a

$$\sum_{h=0}^{r} (-1)^h \binom{r}{h} = 0.)$$

1.4 Si costruiscano due matrici A e $B \in \mathbf{C}^{2\times 2}$, $A, B \neq 0$ tali che $AB + BA = 0$. Si verifichi che per A e $B \in \mathbf{C}^{n\times n}$

$$(A + B)^2 = A^2 + B^2 \quad \text{se e solo se} \quad AB + BA = 0.$$

1.5 Si verifichi che se

$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix},$$

allora $A^k = 0$ per $k \geq 2$, e si dica per quali matrici $B \in \mathbf{C}^{2\times 2}$ vale la relazione

$$(B + \alpha A)^k = B^k + k\alpha B^{k-1} A.$$

1.6 Sia

$$A = \begin{bmatrix} \mathbf{i} & 0 \\ 0 & \mathbf{i} \end{bmatrix}.$$

Si costruisca A^k, k intero positivo. Esiste un k tale che $A^k = A$?

1.7 Sia

$$A = \begin{bmatrix} 1 & \alpha \\ 0 & 1 \end{bmatrix}, \ \alpha \in \mathbf{C}.$$

Si costruisca A^k, k intero (se k è negativo, si definisce $A^k = [A^{-k}]^{-1}$).

1.8 Sia $A \in \mathbf{C}^{2n\times 2n}$ la matrice a blocchi

$$A = \begin{bmatrix} P & Q \\ O & I \end{bmatrix},$$

in cui $P,\ Q \in \mathbf{C}^{n\times n}$ e $P-I$ è non singolare. Si dimostri che per ogni intero positivo k è

$$A^k = \begin{bmatrix} P^k & Q_k \\ O & I \end{bmatrix},$$

dove

$$Q_k = (P^k - I)\ (P-I)^{-1}Q.$$

Se P è non singolare, la relazione precedente vale anche per k intero negativo.

(Traccia: si proceda per induzione e si tenga conto che

$$P^{k-1} + P^{k-2} + \ldots + P + I = (P^k - I)\ (P-I)^{-1}.)$$

1.9 Una matrice $A \in \mathbf{C}^{n\times n}$ tale che $A = A^2$ si dice *idempotente*.

a) Si descriva il sottoinsieme di $\mathbf{C}^{2\times 2}$ delle matrici idempotenti;

b) si dimostri che se A è idempotente, anche $I - A$ è idempotente;

c) si dimostri che se A è idempotente, allora $A^k = A$ per ogni intero k positivo;

d) siano A e $B \in \mathbf{C}^{n\times n}$. Si dimostri che se $AB = A$ e $BA = B$, allora A e B sono idempotenti;

e) si dimostri che se A è involutoria (cioè $A^2 = I$, si veda l'esercizio 1.3), allora la matrice $B = \frac{1}{2}(A + I)$ è idempotente.

(Traccia: d) $ABA = A^2$ e anche $ABA = A$.)

1.10 Una matrice $A \in \mathbf{C}^{n\times n}$ è detta *nilpotente* se esiste un intero positivo k tale che $A^k = 0$. Il minimo intero g tale che $A^g = 0$ è detto *grado di nilpotenza* di A.

a) Si descriva il sottoinsieme di $\mathbf{C}^{2\times 2}$ delle matrici nilpotenti che hanno grado di nilpotenza 2;

b) si dimostri che se A e $B \in \mathbf{C}^{n\times n}$ sono nilpotenti e commutano, allora le matrici AB e $A + B$ sono nilpotenti.

c) si dimostri che una matrice triangolare inferiore (superiore) in senso stretto è nilpotente e si determini il grado di nilpotenza.

1.11 Sia $A = I + B$, dove $B \in \mathbf{C}^{n\times n}$ è una matrice triangolare in senso stretto. Si dimostri che la matrice A è invertibile e che

$$A^{-1} = I - B + B^2 - \ldots + (-1)^k B^k, \quad \text{dove } k \leq n-1.$$

1.12 Si dimostri che una matrice normale e triangolare è diagonale.

1.13 Si dica per quali valori dei parametri α e β la matrice

$$A = \begin{bmatrix} \frac{1}{2} & \frac{\sqrt{3}}{2}\mathbf{i} \\ \alpha & \beta \end{bmatrix}$$

è a) normale b) unitaria c) hermitiana d) definita positiva.

1.14 Sia $A \in \mathbf{C}^{n\times n}$ una matrice hermitiana. Si dimostri che

a) se $p(x)$ è un polinomio in x a coefficienti reali, allora $p(A)$ è una matrice hermitiana (si veda l'esercizio 1.3 per la definizione di polinomio di una matrice);

b) se A è non singolare, allora A^{-1} è hermitiana;

c) se $B \in \mathbf{C}^{n\times n}$, allora BAB^H è hermitiana.

d) le matrici $I + \mathbf{i}A$ e $I - \mathbf{i}A$ sono non singolari e le matrici

$$B = (I - \mathbf{i}A)(I + \mathbf{i}A)^{-1} \quad \text{e} \quad C = (I - \mathbf{i}A)^{-1}(I + \mathbf{i}A)$$

sono unitarie.

1.15 Siano A e $B \in \mathbf{C}^{n\times n}$ due matrici hermitiane.

a) Si faccia vedere con un esempio che la matrice AB non è sempre hermitiana.

b) Si dimostri che la matrice AB è hermitiana se e solo se A e B commutano.

1.16 Sia $A \in \mathbf{C}^{n\times n}$. Si verifichi che le matrici AA^H, A^HA, $A+A^H$, $\mathbf{i}(A-A^H)$ sono hermitiane. Inoltre le matrici A^HA e AA^H sono semidefinite positive (sono definite positive se A non è singolare).

1.17 Si dimostri che l'insieme delle matrici definite positive di $\mathbf{C}^{n\times n}$ è *convesso*, cioè che per ogni coppia $A,\ B$ di matrici definite positive la matrice

$$\alpha A + (1-\alpha)B, \quad 0 < \alpha < 1,$$

è definita positiva.

1.18 Sia $A \in \mathbf{C}^{n\times n}$ una matrice definita positiva. Si dimostri che l'elemento di massimo modulo è un elemento principale.

(Traccia: per ogni coppia di indici i e j si sfrutti la condizione che la sottomatrice di ordine 2 formata dalle corrispondenti righe e colonne ha determinante positivo.)

1.19 Sia $A \in \mathbf{C}^{n\times n}$. Si definisce *traccia* di A la somma dei termini principali di A:

$$\text{tr } A = \sum_{i=1}^{n} a_{ii}.$$

Si dimostri che

$$\begin{aligned}
&\text{tr } (A+B) = \text{tr } A + \text{tr } B,\ B \in \mathbf{C}^{n\times n},\\
&\text{tr } (\alpha A) = \alpha \text{ tr } A,\ \alpha \in \mathbf{C},\\
&\text{tr } (AB) = \text{tr } (BA),\ B \in \mathbf{C}^{n\times n},\\
&\text{tr } (AA^H) = \text{tr } (A^H A) = \sum_{i,j=1}^{n} |a_{ij}|^2.
\end{aligned}$$

Sfruttando queste relazioni si dimostri che non esistono due matrici A e $B \in \mathbf{C}^{n\times n}$ tali che

$$AB - BA = I.$$

1.20 Si dimostri che l'insieme delle matrici simmetriche è uno spazio vettoriale su $\mathbf{C}$ di dimensione $n(n+1)/2$, mentre l'insieme delle matrici hermitiane non costituisce uno spazio vettoriale su $\mathbf{C}$.

1.21 Una matrice $A \in \mathbf{C}^{n\times n}$ si dice *antihermitiana* se $A^H = -A$ ($A \in \mathbf{R}^{n\times n}$ si dice *antisimmetrica* se $A^T = -A$). Si dimostri che

a) se A è antihermitiana (risp. antisimmetrica), allora

$$a_{jj} = \mathbf{i}\theta_j,\ \theta_j \in \mathbf{R} \quad \text{(risp. } 0);$$

b) se $B \in \mathbf{C}^{n\times n}$ è hermitiana, allora la matrice $A = \mathbf{i}B$ è antihermitiana;

c) se $B \in \mathbf{C}^{n\times n}$ è una qualunque matrice, allora $A = B - B^H$ è antihermitiana (se $B \in \mathbf{R}^{n\times n}$, allora $A = B - B^T$ è antisimmetrica);

d) una qualunque matrice $B \in \mathbf{C}^{n\times n}$ può essere espressa come somma di una matrice hermitiana e di una antihermitiana;

e) se A è antihermitiana, allora A^2 è hermitiana;

f) A è antihermitiana se e solo se $\mathbf{x}^H A\mathbf{x} = \mathbf{i}\theta,\ \theta \in \mathbf{R}$, per ogni $\mathbf{x} \in \mathbf{C}^n$.

g) se A è antihermitiana, allora le matrici $I+A$ e $I-A$ sono non singolari e le matrici

$$B = (I-A)\,(I+A)^{-1} \quad \text{e} \quad C = (I-A)^{-1}\,(I+A)$$

sono unitarie;

h) se $A \in \mathbf{R}^{n\times n}$ è antisimmetrica e n è dispari, allora $\det A = 0$ (si veda l'esercizio 1.46 per il caso n pari).

(Traccia: d) si scriva $B = \frac{1}{2}(B + B^H) + \frac{1}{2}(B - B^H)$; f) per dimostrare la necessità della condizione, si verifichi che $\mathbf{x}^H A\mathbf{x} = -\overline{(\mathbf{x}^H A\mathbf{x})}$, per la sufficienza si considerino dei vettori $\mathbf{x}$ e $\mathbf{y}$ opportuni; g) per dimostrare la non singolarità di $I \pm A$, si verifichi che non esistono vettori $\mathbf{x} \neq \mathbf{0}$ tali che $\mathbf{x}^H(I \pm A)\mathbf{x} = \mathbf{0}$, per dimostrare che le matrici B e C sono unitarie, si noti che le due matrici $I + A$ e $I - A$ commutano; h) si ha

$$\det A = \det A^T = \det(-A) = (-1)^n \det A,$$

quindi se n è dispari è $\det A = 0$.)

1.22 Siano A e $B \in \mathbf{C}^{n\times n}$ due matrici unitarie. Allora anche le matrici AB e $B^{-1}AB$ sono unitarie.

1.23 Si dimostri che una matrice $A \in \mathbf{C}^{n\times n}$ è unitaria se e solo se

$$\mathbf{x}^H A^H A\mathbf{x} = \mathbf{x}^H\mathbf{x}, \quad \text{per ogni } \mathbf{x} \in \mathbf{C}^n.$$

(Traccia: per dimostrare la sufficienza della condizione, si noti che la matrice $I - A^H A$ è hermitiana e si scelgano dei vettori $\mathbf{x}$ opportuni.)

1.24 Sia $D \in \mathbf{C}^{n\times n}$ una matrice diagonale unitaria. Si verifichi che gli elementi principali di D sono della forma $\mathrm{e}^{\mathbf{i}\theta}$, $\theta \in \mathbf{R}$.

1.25 Una matrice $A \in \mathbf{C}^{n\times n}$ si dice *a banda di ampiezza* k, o anche $(2k+1)$-*diagonale*, se $a_{ij} = 0$ per $|i - j| > k$ ed esiste un elemento $a_{ij} \neq 0$ per $|i - j| = k$. Si dimostri che

a) se A è a banda di ampiezza k, allora A^2 è a banda di ampiezza al più $2k$, ..., A^i è a banda di ampiezza al più ik;

b) se A e $B \in \mathbf{C}^{n\times n}$ sono due matrici a banda di ampiezza rispettivamente k e h, allora $A + B$ e AB sono matrici a banda, e se ne determini l'ampiezza.

c) Si mostri con un esempio che se A è a banda e non singolare, A^{-1} può non essere a banda.

(Matrici a banda di particolare interesse sono le matrici tridiagonali, che hanno ampiezza di banda $k = 1$, si veda l'esercizio 1.50. Per il punto c) si veda ad esempio l'esercizio 1.53.)

1.26 Si dica sotto quali condizioni la seguente applicazione $< . , . >$ di $\mathbf{C}^n \times \mathbf{C}^n \to \mathbf{C}$

$$< \mathbf{x}, \mathbf{y} > = \mathbf{x}^H A \mathbf{y}$$

definisce un *prodotto scalare* su $\mathbf{C}^n$, cioè gode delle proprietà

1. $< \mathbf{x}, \mathbf{x} > \geq 0$ per ogni $\mathbf{x} \in \mathbf{C}^n$ e $< \mathbf{x}, \mathbf{x} > = 0$ se e solo se $\mathbf{x} = \mathbf{0}$,
2. $\overline{< \mathbf{x}, \mathbf{y} >} = < \mathbf{y}, \mathbf{x} >$,
3. $< \mathbf{x}, \alpha\mathbf{y} > = \alpha \ < \mathbf{x}, \mathbf{y} >$, per $\alpha \in \mathbf{C}$,
4. $< \mathbf{x}, \mathbf{y} + \mathbf{z} > = < \mathbf{x}, \mathbf{y} > + < \mathbf{x}, \mathbf{z} >$, per $\mathbf{z} \in \mathbf{C}^n$.

1.27 Si dimostri che per ogni prodotto scalare $< . , . >$ su $\mathbf{C}^n$ esiste una matrice $A \in \mathbf{C}^{n \times n}$ definita positiva tale che

$$< \mathbf{x}, \mathbf{y} > = \mathbf{x}^H A \mathbf{y}.$$

(Traccia: si ponga $\mathbf{x} = \sum_{i=1}^{n} x_i \mathbf{e}_i$ e $\mathbf{y} = \sum_{i=1}^{n} y_i \mathbf{e}_i$. Per le proprietà del prodotto scalare risulta

$$< \mathbf{x}, \mathbf{y} > = \sum_{i,j=1}^{n} \overline{x}_i y_j \ < \mathbf{e}_i, \mathbf{e}_j > .)$$

1.28 Se $< . , . >$ è un prodotto scalare su $\mathbf{C}^n$, si dimostri che se $< \mathbf{x}, \mathbf{y}_i > = 0$ per n vettori $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ linearmente indipendenti di $\mathbf{C}^n$, allora $\mathbf{x} = \mathbf{0}$.

1.29 Si dimostri che vale la seguente *legge del parallelogramma*:

$$< \mathbf{x}+\mathbf{y}, \mathbf{x}+\mathbf{y} > + < \mathbf{x}-\mathbf{y}, \mathbf{x}-\mathbf{y} > = 2(< \mathbf{x}, \mathbf{x} > + < \mathbf{y}, \mathbf{y} >).$$

(Traccia: si applichino le proprietà 3. e 4. della definizione).

1.30 Si dimostri la disuguaglianza di Cauchy-Schwarz

$$| \ < \mathbf{x}, \mathbf{y} > \ |^2 \leq < \mathbf{x}, \mathbf{x} > < \mathbf{y}, \mathbf{y} > .$$

(Traccia: per $\mathbf{y} \neq \mathbf{0}$ si ponga $\alpha = \dfrac{< \mathbf{y}, \mathbf{x} >}{< \mathbf{y}, \mathbf{y} >}$ e si sviluppi la disuguaglianza $< \mathbf{x} - \alpha\mathbf{y}, \mathbf{x} - \alpha\mathbf{y} > \geq 0$.)

1.31 Siano $A \in \mathbf{C}^{n \times n}$ e $\mathbf{x}_1 \in \mathbf{C}^n$. Si costruisca la successione di vettori

$$\mathbf{x}_{i+1} = A\mathbf{x}_i, \quad i = 1, 2, \ldots$$

Se per un k è $\mathbf{x}_k \neq \mathbf{0}$ e $\mathbf{x}_{k+1} = \mathbf{0}$, allora i vettori $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$ sono linearmente indipendenti.

1.32 Sia S il sottospazio di $\mathbf{C}^3$ generato dai vettori $[1, \mathbf{i}, \text{-}\mathbf{i}]^T$ e $[\mathbf{i}, \text{-}\mathbf{i}, 0]^T$. Si determini il sottospazio T tale che

$$S \oplus T = \mathbf{C}^3.$$

1.33 Si determini una base per il nucleo di

$$A = \begin{bmatrix} 1 & 1 & 2 & 2 \\ 1 & \text{-}1 & \text{-}1 & \text{-}2 \\ 1 & \text{-}3 & \text{-}4 & \text{-}6 \\ 2 & 0 & 1 & 0 \end{bmatrix}.$$

1.34 Siano S e T sottospazi di $\mathbf{C}^n$.

a) Si dimostri che $(S^\perp)^\perp = S$.

b) Si esprimano $(S+T)^\perp$ e $(S \cap T)^\perp$ in funzione di $S^\perp$ e $T^\perp$.

1.35 Sia $A \in \mathbf{C}^{m \times n}$. Si dimostri che il numero delle righe linearmente indipendenti di A è uguale al numero delle colonne linearmente indipendenti. (Traccia: siano r e c il numero delle righe e delle colonne linearmente indipendenti di A, e si consideri la matrice B formata dalle r righe linearmente indipendenti di A. Si dimostri che il numero c_B delle colonne linearmente indipendenti di B è tale che $c_B \leq r$ ed inoltre $c_B = c$, per cui $c \leq r$. Si applichi poi lo stesso ragionamento ad A^T.)

1.36 Si determini il rango delle seguenti matrici

$$\text{a) } I_n, \quad \text{b) } \begin{bmatrix} I_n \\ O \end{bmatrix}, \quad \text{c) } \begin{bmatrix} I_n & O \end{bmatrix}, \quad \text{d) } \begin{bmatrix} I_n & O \\ O & O \end{bmatrix}.$$

1.37 Siano A e $B \in \mathbf{C}^{n \times n}$. Si dimostrino le seguenti relazioni:

a) rango di $(A+B) \leq$ rango di A + rango di B;

b) rango di A + rango di $B - n$

$$\leq \text{ rango di } AB \leq \min \{ \text{ rango di } A, \text{ rango di } B \};$$

c) se A è idempotente, allora rango di A + rango di $(I - A) = n$;

d) le due proprietà seguenti sono equivalenti

1. per ogni $\mathbf{x}$ tale che $AB\mathbf{x} = \mathbf{0}$ segue che $B\mathbf{x} = \mathbf{0}$

2. rango di AB = rango di B;

e) $n \geq$ rango di $A \geq$ rango di $A^2 \geq \ldots$

f) se rango di A^k = rango di A^{k+1}, allora
rango di A^j = rango di A^k, per ogni $j \geq k$;

g) se A è antisimmetrica, il rango di A è un numero pari.

(Traccia: b) per la prima disuguaglianza, si noti che

$$\dim N(AB) \leq \dim N(A) + \dim N(B)$$

e si sfrutti la (7), per la seconda, si noti che le colonne (risp. le righe) di AB sono combinazioni lineari delle colonne di A (risp. delle righe di B); c) basta dimostrare che rango di $(I - A)$= dim $N(A)$ nell'ipotesi che $(I - A)A = 0$. Se $\mathbf{x} \in N(A)$, allora $(I - A)\mathbf{x} = \mathbf{x}$, quindi rango di $(I - A) \geq$ dim $N(A)$. Se inoltre $\mathbf{x} \in N(A)^\perp$, allora $A\mathbf{x} \neq \mathbf{0}$, e poiché $(I - A)A\mathbf{x} = \mathbf{0}$, segue che dim $N(I - A) \geq$ rango di A, cioè rango di $(I - A) \leq$ dim $N(A)$; g) si veda il punto h) dell'esercizio 1.21.)

1.38 Sia $A \in \mathbf{C}^{m\times n}$. Si verifichi che

$$N(A^H A) = N(A),$$
$$S(A^H A) = S(A^H), \quad \text{dove } S(A) \text{ è l'immagine di } A.$$

(Traccia: se $A\mathbf{x} = \mathbf{0}$, allora $A^H A\mathbf{x} = \mathbf{0}$ e quindi $N(A) \subset N(A^H A)$; viceversa se $A^H A\mathbf{x} = \mathbf{0}$ e per assurdo $A\mathbf{x} \neq \mathbf{0}$, il vettore $\mathbf{y} = A\mathbf{x}$ è tale che $\mathbf{y} \neq \mathbf{0}$, $\mathbf{y} \in S(A)$ e $\mathbf{y} \in N(A^H)$, ma ciò non è possibile perché $N(A^H) = S(A)^\perp$. La seconda relazione deriva dalla prima tenendo conto che $N(A) = S(A^H)^\perp$.)

1.39 Sia A la matrice triangolare a blocchi

$$A = \begin{bmatrix} A_{11} & A_{12} & \ldots & A_{1n} \\ & A_{22} & \ldots & A_{2n} \\ & & \ddots & \vdots \\ & & & A_{nn} \end{bmatrix},$$

con blocchi diagonali quadrati. Si dimostri che

$$\text{rango di } A \geq \sum_{i=1}^{n} \text{rango di } A_{ii},$$

inoltre vale il segno di uguaglianza se la matrice A è diagonale a blocchi.

1.40 Siano $\mathbf{x}, \mathbf{y}, \mathbf{u}, \mathbf{v} \in \mathbf{C}^n$. Si determinino il nucleo e il rango della matrice

$$A = \mathbf{x}\mathbf{y}^H + \mathbf{u}\mathbf{v}^H.$$

Più in generale, siano $\mathbf{u}_i, \mathbf{v}_i \in \mathbf{C}^n,\ i = 1, 2, \ldots, r$. Si dica qual è il rango della matrice

$$A = \sum_{i=1}^{r} \mathbf{u}_i\, \mathbf{v}_i^H. \tag{13}$$

Viceversa si dimostri che se il rango di A è r, allora esistono r coppie di vettori $\mathbf{u}_i, \mathbf{v}_i \in \mathbf{C}^n$ per cui vale la relazione (13).

1.41 Sia A la matrice diagonale a blocchi

$$A = \begin{bmatrix} A_{11} & & & \\ & A_{22} & & \\ & & \ddots & \\ & & & A_{nn} \end{bmatrix},$$

con blocchi diagonali quadrati. Si dimostri che

$$\det A = \det A_{11} \det A_{22} \ldots \det A_{nn},$$

e che la stessa relazione vale anche per le matrici triangolari a blocchi.

1.42 Si dimostri che se tutte le somme per righe degli elementi di A sono nulle, allora la matrice A è singolare.

(Traccia: il sistema $A\mathbf{x} = \mathbf{0}$ ammette la soluzione $\mathbf{x} = [1, 1, \ldots, 1]^T$.)

1.43 Siano $A_{11} \in \mathbf{C}^{n\times n}$ non singolare, $A_{22} \in \mathbf{C}^{m\times m}$, $A_{12} \in \mathbf{C}^{n\times m}$ e $A_{21} \in \mathbf{C}^{m\times n}$ e sia

$$A = \begin{bmatrix} A_{11} & A_{12} \\ & \\ A_{21} & A_{22} \end{bmatrix} \in \mathbf{C}^{(n+m)\times(n+m)}.$$

a) Si verifichi che

$$A = \begin{bmatrix} I & O \\ & \\ A_{21}A_{11}^{-1} & I \end{bmatrix} \begin{bmatrix} A_{11} & O \\ & \\ O & A_{22} - A_{21}A_{11}^{-1}A_{12} \end{bmatrix} \begin{bmatrix} I & A_{11}^{-1}A_{12} \\ & \\ O & I \end{bmatrix}.$$

La matrice $S = A_{22} - A_{21}A_{11}^{-1}A_{12}$ è detta *complemento di Schur* di A_{11} nella matrice A.

b) Si dimostri che vale la relazione

$$\det A = \det A_{11} \det(A_{22} - A_{21}A_{11}^{-1}A_{12})$$

e se $m = n$ e A_{11} commuta con A_{21} è

$$\det A = \det(A_{11}A_{22} - A_{21}A_{12}).$$

c) Si dimostri che

$$\text{rango di } A = n \quad \text{se e solo se } S = 0.$$

d) Se A è non singolare, si verifichi che

$$A^{-1} = \begin{bmatrix} A_{11}^{-1} + A_{11}^{-1}A_{12}S^{-1}A_{21}A_{11}^{-1} & -A_{11}^{-1}A_{12}S^{-1} \\ -S^{-1}A_{21}A_{11}^{-1} & S^{-1} \end{bmatrix}.$$

e) Si esamini il caso particolare di $m = 1$, cioè

$$A = \begin{bmatrix} A_{11} & \mathbf{u} \\ \mathbf{v}^H & \alpha \end{bmatrix}, \ \mathbf{u} \in, \ \mathbf{v} \in \mathbf{C}^n, \ \alpha \in \mathbf{C}.$$

f) Si dimostri che se A è definita positiva, allora S è definita positiva.

(Traccia: f) sia $\mathbf{x}_2 \in \mathbf{C}^m$, $\mathbf{x}_2 \neq \mathbf{0}$ e $\mathbf{x}_1 = -A_{11}^{-1}A_{12}\mathbf{x}_2$, posto $\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix}$, è $\mathbf{x}^H A\mathbf{x} = \mathbf{x}_2^H S\mathbf{x}_2$.)

1.44 Siano $\mathbf{u}, \mathbf{v} \in \mathbf{C}^n$. Allora

$$\det(I + \mathbf{u}\mathbf{v}^H) = 1 + \mathbf{v}^H\mathbf{u}$$

e

$$(I + \mathbf{u}\mathbf{v}^H)^{-1} = I - \frac{\mathbf{u}\mathbf{v}^H}{1 + \mathbf{v}^H\mathbf{u}}.$$

Di questa formula si può dare la seguente generalizzazione

$$\det(I_n + UV^H) = \det(I_m + V^HU),$$

dove $U, V \in \mathbf{C}^{n\times m}$ e

$$(I_n + UV^H)^{-1} = I_n - U(I_m + V^HU)^{-1}V^H.$$

(Traccia: si verifichi prima che

$$\begin{bmatrix} I & \mathbf{0} \\ -\mathbf{v}^H & 1 \end{bmatrix} \begin{bmatrix} I & -\mathbf{u} \\ \mathbf{v}^H & 1 \end{bmatrix} = \begin{bmatrix} I & -\mathbf{u} \\ \mathbf{0}^H & 1 + \mathbf{v}^H\mathbf{u} \end{bmatrix}$$

e che

$$\begin{bmatrix} I & -\mathbf{u} \\ \mathbf{v}^H & 1 \end{bmatrix} \begin{bmatrix} I & \mathbf{0} \\ -\mathbf{v}^H & 1 \end{bmatrix} = \begin{bmatrix} I + \mathbf{u}\mathbf{v}^H & -\mathbf{u} \\ \mathbf{0}^H & 1 \end{bmatrix}$$

e si applichi il teorema di Binet.)

1.45 Sia $A \in \mathbf{C}^{n\times n}$ una matrice non singolare e siano $\mathbf{u}, \mathbf{v} \in \mathbf{C}^n$. Si dimostri che

a) $\mathbf{v}^H A^{-1}\mathbf{u} = -1$, se e solo se la matrice $A + \mathbf{u}\mathbf{v}^H$ è singolare,

b) se la matrice $A + \mathbf{u}\mathbf{v}^H$ è non singolare, allora $\mathbf{v}^H A^{-1}\mathbf{u} \neq -1$ e vale la seguente formula di *Sherman-Morrison*

$$(A + \mathbf{u}\mathbf{v}^H)^{-1} = A^{-1} - \frac{A^{-1}\mathbf{u}\mathbf{v}^H A^{-1}}{1 + \mathbf{v}^H A^{-1}\mathbf{u}}.$$

Di questa formula si può dare la seguente generalizzazione di *Woodbury*: se $U,\ V \in \mathbf{C}^{n\times m}$ e la matrice $I + V^H A^{-1} U$ è non singolare, allora anche $A + UV^H$ è non singolare e risulta

$$(A + UV^H)^{-1} = A^{-1} - A^{-1}U(I + V^H A^{-1}U)^{-1}V^H A^{-1}.$$

(Traccia: si applichi la relazione dell'esercizio 1.44 alla matrice $A + \mathbf{u}\mathbf{v}^H = A(I + A^{-1}\mathbf{u}\mathbf{v}^H)$.)

1.46 Sia $A \in \mathbf{R}^{n\times n}$ antisimmetrica e sia n pari. Allora $\det A$ è il quadrato di un polinomio negli elementi della matrice (per la definizione di matrice antisimmetrica si veda l'esercizio 1.21).

(Traccia: si dimostri per induzione con n pari, incrementando n di 2 in 2. Per $n = 2$ la matrice A è della forma

$$\det \begin{bmatrix} 0 & -b \\ b & 0 \end{bmatrix}, \quad b \in \mathbf{R},$$

e quindi $\det A = b^2$. Per $n > 2$, si supponga che la tesi valga per $n - 2$ e si partizioni A nel modo seguente

$$A = \begin{bmatrix} A_{11} & -A_{21}^T \\ A_{21} & A_{22} \end{bmatrix},$$

in cui $A_{11} \in \mathbf{R}^{(n-2)\times(n-2)}$ e $A_{22} \in \mathbf{R}^{2\times 2}$ sono antisimmetriche. Se A_{22} è non singolare, allora per l'esercizio 1.43 si ha

$$\det A = \det A_{22} \det(A_{11} + A_{21}^T A_{22}^{-1} A_{21}).$$

Poiché le matrici A_{22} e $A_{11} + A_{21}^T A_{22}^{-1} A_{21}$ sono antisimmetriche di ordine 2 ed $n-2$, si può applicare l'ipotesi induttiva. Se A_{22} è singolare, si individui una sottomatrice principale di ordine 2 non singolare e si permutino righe e colonne di A in modo che tale sottomatrice si trovi nell'angolo a destra in basso. Se tutte le sottomatrici principali di ordine 2 sono singolari, allora risulta $A = O$.)

1.47 Siano A e B matrici partizionate a blocchi A_{ij} e B_{jk}, $i = 1, \ldots, p$, $j = 1, \ldots, q$, $k = 1, \ldots, r$. Si dimostri che se le dimensioni dei blocchi sono compatibili, allora la matrice AB risulta partizionata in $p \times r$ blocchi, in cui il blocco (i, k)-esimo è dato da

$$(AB)_{ik} = \sum_{j=1}^{q} A_{ij} B_{jk}.$$

1.48 Sia $A \in \mathbf{C}^{n \times n}$. Si dimostri che

a) valgono le seguenti proprietà

$$\begin{aligned}
&\text{adj } (A^H) = (\text{adj } A)^H, \\
&\text{adj } I = I, \\
&\text{adj } (\alpha A) = \alpha^{n-1} \text{adj } A, \quad \text{per } \alpha \in \mathbf{C}, \\
&\text{adj } (AB) = \text{adj } B \text{ adj } A, \quad \text{e quindi} \\
&\text{adj } A^{-1} = (\text{adj } A)^{-1}, \quad \text{se } A \text{ è non singolare}, \\
&A \text{ adj } A = (\text{adj } A)A = (\det A)I, \text{ e quindi} \\
&A \text{ è non singolare se e solo se } \det A \neq 0, \quad \text{inoltre} \\
&A^{-1} = \frac{1}{\det A} \text{ adj } A, \\
&\det(\text{adj } A) = (\det A)^{n-1}, \\
&\text{adj } (\text{adj } A) = (\det A)^{n-2} A;
\end{aligned}$$

b) se A è hermitiana, anche adj$A$ è hermitiana, se A è antihermitiana, adj$A$ è hermitiana se n è dispari e antihermitiana se n è pari;

c) se $\mathbf{u}$ e $\mathbf{v} \in \mathbf{C}^n$ e $\alpha \in \mathbf{C}$, allora vale

$$\det \begin{bmatrix} A & \mathbf{u} \\ \mathbf{v}^H & \alpha \end{bmatrix} = \alpha \det A - \mathbf{v}^H (\text{adj } A)\mathbf{u}.$$

(Traccia: a) si applichino le proprietà dei determinanti; b) se $A^H = A$, allora $(\text{adj } A)^H = \text{adj } A$, se $A^H = -A$, allora $(\text{adj } A)^H = (-1)^{n-1} \text{adj } A$; c) si calcoli il determinante con la regola di Laplace sull'ultima colonna.)

1.49 Si calcoli il determinante delle matrici $A \in \mathbf{C}^{n\times n}$, i cui elementi sono

a) $a_{ij} = 0$ per $j > n + 1 - i$;

b) $a_{ij} = \begin{cases} \alpha + \beta & \text{se } i = j, \\ \alpha & \text{se } i \neq j, \end{cases} \quad \alpha, \beta \in \mathbf{C}$;

c) $a_{ij} = \begin{cases} 2 & \text{se } i = j, \\ 1 & \text{se } |i - j| = 1, \\ 0 & \text{altrimenti,} \end{cases}$

(Traccia: a) $\det A = (-1)^{n(n-1)/2} a_{1n} a_{2,n-1} \dots a_{n1}$, b) si applichi l'esercizio 1.44, $\det A = \beta^{n-1}(n\alpha + \beta)$, c) si proceda per induzione su n, oppure si veda l'esercizio successivo, $\det A = n + 1$.)

1.50 Una matrice tridiagonale $A_n \in \mathbf{C}^{n\times n}$

$$A_n = \begin{bmatrix} \alpha_1 & \gamma_2 & & \\ \beta_2 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \gamma_n \\ & & \beta_n & \alpha_n \end{bmatrix},$$

in cui gli α_i, $i = 1, \dots, n$, β_i e γ_i, $i = 2, \dots, n$, sono numeri complessi, viene anche chiamata *matrice di Jacobi*.

a) Si dimostri che vale la relazioni ricorrente

$$\det A_n = \alpha_n \det A_{n-1} - \beta_n \gamma_n \det A_{n-2},$$

che consente di calcolare il determinante di A_n a partire dalle due relazioni iniziali

$$\det A_1 = \alpha_1, \qquad \det A_2 = \alpha_1 \alpha_2 - \beta_2 \gamma_2.$$

b) Per il caso particolare $\alpha_1 = \dots = \alpha_n = \gamma_2 = \dots = \gamma_n = 1$, $\beta_2 = \dots = \beta_n = -1$ (tali matrici si chiamano anche *matrici di Fibonacci*), si dimostri che $\det A_n$ è uguale all'n-esimo numero di Fibonacci.

c) Per il caso particolare $\alpha_1 = \dots = \alpha_n = 2\cos\theta$, $\theta \in \mathbf{R}$, $\beta_2 = \dots = \beta_n = \gamma_2 = \dots = \gamma_n = 1$, si verifichi che

$$\det A_n = \frac{\sin(n+1)\theta}{\sin\theta}.$$

d) Se $\alpha_i \in \mathbf{R}$, $i = 1, \dots, n$ e $\beta_i \gamma_i > 0$, $i = 2, \dots, n$, si costruisca una matrice $D \in \mathbf{C}^{n\times n}$ diagonale tale che la matrice $B = D^{-1}AD$ sia reale, simmetrica e tridiagonale.

(Traccia: a) si sviluppi $\det A_n$ con la regola di Laplace applicata all'ultima riga; b) si verifichi che in questo caso la relazione ricorrente del punto a) è quella che definisce i numeri di Fibonacci; c) si verifichi che

$$\det A_n = 2\cos\theta \det A_{n-1} - \det A_{n-2};$$

d) si imponga la condizione che

$$\frac{\beta_i d_{i-1,i-1}}{d_{ii}} \;=\; \frac{\overline{\gamma}_i\ \overline{d}_{ii}}{\overline{d}_{i-1,i-1}}\ , \quad \text{per } i = 2, \ldots, n,$$

e che tali valori siano reali.)

1.51 Si calcoli l'inversa della matrice A di elementi

$$a_{ij} = \begin{cases} 1 & \text{se } i = j, \\ -\dfrac{j}{i} & \text{se } i = j+1, \\ 0 & \text{altrimenti,} \end{cases}$$

(Traccia: si dimostri per induzione sull'ordine della matrice che gli elementi di $B^{-1} = A$ sono dati da

$$b_{ij} = \begin{cases} \dfrac{j}{i} & \text{se } j \le i, \\ 0 & \text{altrimenti.)} \end{cases}$$

1.52 Siano $x_1, x_2, \ldots, x_n,\ n$ numeri complessi. Si consideri la matrice A i cui elementi sono

$$a_{ij} = \begin{cases} x_{i-j+1} & \text{se } i \ge j, \\ 0 & \text{altrimenti.} \end{cases}$$

Si dimostri che la matrice $B = A^{-1}$ ha elementi del tipo

$$b_{ij} = \begin{cases} y_{i-j+1} & \text{se } i \ge j, \\ 0 & \text{altrimenti,} \end{cases}$$

e si esaminino in particolare i casi

(1) $x_i = i$, per $i = 1, 2, \ldots, n$,

(2) $x_1 = 1,\ x_2 = -\alpha,\ x_3 = \ldots = x_n = 0,\ \alpha \in \mathbf{C}$.

(Traccia: per mezzo della matrice

$$U = \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{bmatrix},$$

si può scrivere

$$A = x_1 I + x_2 U + x_3 U^2 + \ldots + x_n U^{n-1}.$$

Si verifichi che la matrice A^{-1} è della stessa forma della A, determinando le n costanti $y_1, y_2, \ldots, y_n$ tali che

$$A^{-1} = y_1 I + y_2 U + y_3 U^2 + \ldots + y_n U^{n-1}.$$

Per i casi particolari si ha:

(1) $$A = \begin{bmatrix} 1 & & & & \\ 2 & 1 & & & \\ 3 & 2 & 1 & & \\ \vdots & & \ddots & \ddots & \\ n & n-1 & \ldots & 2 & 1 \end{bmatrix}, \quad A^{-1} = \begin{bmatrix} 1 & & & & \\ -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \end{bmatrix},$$

(2) $$A = \begin{bmatrix} 1 & & & \\ -\alpha & 1 & & \\ & \ddots & \ddots & \\ & & -\alpha & 1 \end{bmatrix}, \quad A^{-1} = \begin{bmatrix} 1 & & & & \\ \alpha & 1 & & & \\ \alpha^2 & \alpha & 1 & & \\ \vdots & \ddots & \ddots & \ddots & \\ \alpha^{n-1} & \ldots & \alpha^2 & \alpha & 1 \end{bmatrix}.)$$

1.53 Siano $x_1, x_2, \ldots, x_n,\ n$ numeri complessi. Si consideri la matrice A i cui elementi sono

$$a_{ij} = \begin{cases} x_i & \text{se } i \geq j, \\ x_j & \text{altrimenti.} \end{cases}$$

Si dimostri che

a) $$\det A = x_n \prod_{i=1}^{n-1} (x_i - x_{i+1}),$$

per cui A è non singolare se $x_i \neq x_{i+1},\ i = 1, \ldots, n-1$ e $x_n \neq 0$;

b) la matrice A^{-1} è tridiagonale e se ne dia l'espressione;

c) si considerino in particolare i casi

(1) $x_i = i$, per $i = 1, 2, \ldots, n$,

(2) $x_1 = n - i + 1$, per $i = 1, 2, \ldots, n$.

(Traccia: a) si verifichi che, posto

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

la matrice $D = S^T A S$ risulta diagonale con gli elementi principali

$$d_{ii} = x_i - x_{i+1},\ i = 1, \ldots, n-1, \quad \text{e } d_{nn} = x_n.$$

b)

$$A^{-1} = SD^{-1}S^T = \begin{bmatrix} \alpha_1 & -\alpha_1 & & & \\ -\alpha_1 & \alpha_1 + \alpha_2 & -\alpha_2 & & \\ & -\alpha_2 & \alpha_2 + \alpha_3 & \ddots & \\ & & \ddots & \ddots & -\alpha_{n-1} \\ & & & -\alpha_{n-1} & \alpha_{n-1} + \alpha_n \end{bmatrix},$$

in cui $\alpha_i = \dfrac{1}{x_i - x_{i+1}}$, per $i = 1, \ldots, n-1$ e $\alpha_n = \dfrac{1}{x_n}$.

c) per i casi particolari si ha:

(1) $$A = \begin{bmatrix} 1 & 2 & 3 & \ldots & n \\ 2 & 2 & 3 & & n \\ 3 & 3 & 3 & & \vdots \\ \vdots & & & \ddots & \vdots \\ n & n & n & \ldots & n \end{bmatrix}, \quad A^{-1} = \begin{bmatrix} -1 & 1 & & & \\ 1 & -2 & \ddots & & \\ & \ddots & \ddots & 1 & \\ & & 1 & -2 & 1 \\ & & & 1 & \dfrac{1-n}{n} \end{bmatrix},$$

(2) $$A = \begin{bmatrix} n & n-1 & n-2 & \ldots & 1 \\ n-1 & n-1 & n-2 & & 1 \\ n-2 & n-2 & n-2 & & \vdots \\ \vdots & & & \ddots & \vdots \\ 1 & 1 & 1 & \ldots & 1 \end{bmatrix}, \quad A^{-1} = \begin{bmatrix} 1 & -1 & & & \\ -1 & 2 & \ddots & & \\ & \ddots & \ddots & -1 & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix}.$$

1.54 Siano $x_1, x_2, \ldots, x_n$, n numeri complessi. Si consideri la matrice A i cui elementi sono

$$a_{ij} = x_i^{j-1}, \quad i,\ j = 1, 2, \ldots, n,$$

detta matrice di *Vandermonde* di ordine n. Si dimostri che

$$\det A = \prod_{\substack{i,k=1,n \\ i<k}} (x_k - x_i),$$

e quindi $\det A \neq 0$ se i numeri x_i sono a due a due distinti.

(Traccia: si dimostri per induzione che, detto V_n il determinante della matrice A di ordine n, vale la relazione

$$V_n = (x_2 - x_1)(x_3 - x_1) \cdots (x_n - x_1) V_{n-1}.$$

Per questo, conviene sottrarre dalla k-esima colonna di A la $(k-1)$-esima moltiplicata per x_1, per $k = n, n-1, \ldots, 2$).

1.55 Sia $A \in \mathbf{R}^{n\times n}$ e si supponga che gli elementi di A siano funzioni $a_{ij} = a_{ij}(x)$ derivabili di una variabile reale x. Si dimostri che

$$\frac{\mathrm{d}}{\mathrm{d}x} (\det A) = \mathrm{tr}\ ((\mathrm{adj}\ A) A'),$$

dove A' è la matrice i cui elementi sono le derivate degli elementi di A.

(Traccia: si dimostri prima che

$$\frac{\partial}{\partial a_{ij}} (\det A) = (\mathrm{adj}\ A)_{ji}, \quad \text{e che} \quad \frac{\mathrm{d}}{\mathrm{d}x} (\det A) = \sum_{i,j=1}^{n} \frac{\partial}{\partial a_{ij}} (\det A) \frac{\mathrm{d}a_{ij}}{\mathrm{d}x}.)$$

1.56 Sia $\Pi \in \mathbf{R}^{n\times n}$ la matrice di permutazione ottenuta scambiando fra loro la i-esima e la j-esima riga della matrice identica.

a) Si dimostri che Π è ortogonale.

b) Per un vettore $\mathbf{x} \in \mathbf{C}^n$ e una matrice $A \in \mathbf{C}^{n\times n}$ si descrivano il vettore $\Pi\mathbf{x}$ e le matrici ΠA e $A\Pi$.

1.57 Si verifichi che la matrice A i cui elementi sono

$$a_{ij} = \begin{cases} 1 & \text{se } |i-j| = 1, \\ \\ 0 & \text{altrimenti,} \end{cases}$$

è irriducibile.

1.58 Una matrice $A \in \mathbf{C}^{n\times n}$ della forma

$$A = \begin{bmatrix} \alpha_1 & \alpha_2 & \alpha_3 & \dots & \alpha_n \\ \alpha_n & \alpha_1 & \alpha_2 & \dots & \alpha_{n-1} \\ \alpha_{n-1} & \alpha_n & \alpha_1 & \dots & \alpha_{n-2} \\ \vdots & & & & \vdots \\ \alpha_2 & \alpha_3 & \dots & \alpha_n & \alpha_1 \end{bmatrix},$$

dove $\alpha_1, \alpha_2, \dots, \alpha_n \in \mathbf{C}$, è detta *circolante*. Si dimostri che se per un $j \neq 1$ e tale che $j-1$ sia primo con n è $\alpha_j \neq 0$, allora A è irriducibile.

1.59 Una matrice $A \in \mathbf{C}^{n\times n}$ della forma

$$A = \begin{bmatrix} \alpha_1 & \alpha_2 & \alpha_3 & \cdots & \alpha_n \\ \gamma_2 & \beta_2 & 0 & \cdots & 0 \\ \gamma_3 & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \beta_{n-1} & 0 \\ \gamma_n & 0 & \cdots & 0 & \beta_n \end{bmatrix}$$

si dice matrice *ad albero*. Si dimostri che

$$\det A = \alpha_1 \prod_{i=2}^{n} \beta_i \; - \; \sum_{i=2}^{n} \alpha_i \gamma_i \prod_{\substack{j=2 \\ j\neq i}}^{n} \beta_j.$$

(Traccia: si dimostri prima la formula per $\beta_i \neq 0$, $i = 2, \dots, n$, e si estenda al caso generale per continuità. Si verifichi che $A = T(I + \mathbf{u}\mathbf{v}^H)$, dove

$$T = \begin{bmatrix} \alpha_1 & & & \\ \gamma_2 & \beta_2 & & \\ \vdots & & \ddots & \\ \gamma_n & & & \beta_n \end{bmatrix}, \quad \mathbf{u} = T^{-1}\mathbf{e}_1, \quad \mathbf{v} = \begin{bmatrix} 0 \\ \overline{\alpha}_2 \\ \vdots \\ \overline{\alpha}_n \end{bmatrix},$$

e si applichi l'esercizio 1.44.)

1.60 Siano $A \in \mathbf{C}^{n\times n}$ e $B \in \mathbf{C}^{m\times m}$. Si definisce *prodotto di Kronecker (o diretto o tensoriale)* di A e B la matrice di ordine nm

$$A \otimes B = \begin{bmatrix} a_{11}B & a_{12}B & \dots & a_{1n}B \\ a_{21}B & a_{22}B & \dots & a_{2n}B \\ \vdots & \vdots & & \dots \\ a_{n1}B & a_{n2}B & \dots & a_{nn}B \end{bmatrix}.$$

Si dimostri che:

a) la matrice $I_n \otimes B$ è diagonale a blocchi, con i blocchi principali uguali a B;

b) per matrici A, B e C di dimensioni opportune, valgono le seguenti proprietà

$$\begin{aligned}
&A \otimes (B + C) = A \otimes B + A \otimes C,\\
&(A + B) \otimes C = A \otimes C + B \otimes C,\\
&(\alpha A) \otimes B = A \otimes (\alpha B) = \alpha\ (A \otimes B),\ \alpha \in \mathbf{C},\\
&A \otimes (B \otimes C) = (A \otimes B) \otimes C,\\
&(A \otimes B)^H = A^H \otimes B^H,
\end{aligned}$$

c) $I_n \otimes I_m = I_{nm}$;

d) se $C \in \mathbf{C}^{n\times n}$ e $D \in \mathbf{C}^{m\times m}$, allora

$$[A \otimes B]\ [C \otimes D] = (AC) \otimes (BD)$$

e quindi

$$A \otimes B = (A \otimes I_m)\ (I_n \otimes B) = (I_n \otimes B)\ (A \otimes I_m);$$

e) $A \otimes B$ è non singolare se e solo se A e B sono non singolari, inoltre

$$(A \otimes B)^{-1} = A^{-1} \otimes B^{-1};$$

f) esiste una matrice di permutazione $\Pi \in \mathbf{R}^{nm\times nm}$ tale che

$$\Pi(A \otimes B)\Pi^T = B \otimes A;$$

g) $\det(I_n \otimes B) = (\det B)^n$, e quindi

$$\det(A \otimes B) = \det(B \otimes A) = (\det A)^m (\det B)^n;$$

h) sono chiuse rispetto all'operazione prodotto di Kronecker le seguenti classi di matrici

(1) matrici unitarie,

(2) matrici hermitiane,

(3) matrici definite positive.

Commento bibliografico

L'uso di tabelle per la rappresentazione di dati aventi caratteristiche comuni è antico, documentato fin dall'epoca egiziana e babilonese: tipico è

l'uso di rappresentare dati numerici in tabelle quadrate in modo che siano rispettate certe regole, come nei quadrati magici. Anche nella matematica cinese compaiono esempi di quadrati magici antichi più di duemila anni. Si deve però arrivare al 1693 per trovare il primo esempio di uso delle matrici in senso moderno: in una lettera di Leibniz a L'Hospital compare il sistema

$$\begin{aligned} 10 + 11x + 12y = 0, \\ 20 + 21x + 22y = 0, \end{aligned}$$

in cui i dati 10, 11, 12, 20, 21 e 22 non sono coefficienti, ma hanno un valore simbolico e stanno a indicare delle variabili, assumendo così il ruolo di indici, e si suggerisce anche un metodo di risoluzione del sistema che è assai simile al metodo di eliminazione. Questa lettera però non venne pubblicata fino al 1850.

Nei primi anni del 18^o secolo alla rappresentazione dei dati in tabella viene gradualmente associandosi anche l'idea di una grandezza, il determinante, che in qualche modo la caratterizzi. Uno degli scopritori di questa idea è Cramer, che la presenta in un lavoro del 1750, attirando su di essa l'attenzione dei matematici contemporanei. Nei circa 60 anni che seguono altri matematici contribuiscono allo sviluppo della teoria dei determinanti: da citare fra gli altri Vandermonde, che nel 1771 pubblica una memoria sul metodo di eliminazione per i sistemi lineari e Laplace, che nel 1772 utilizza il termine di *risultante*, per indicare il determinante.

Il termine *determinante* compare in una monografia che nel 1812 Cauchy presenta all'Accademia delle Scienze di Parigi sulla teoria dei determinanti, nello stesso giorno in cui anche Binet presenta il suo importante teorema sul determinante del prodotto di matrici. Grazie a Cauchy, i determinanti divengono di uso comune nella ricerca matematica: Jacobi, in particolare, impiega i determinanti per risolvere problemi di geometria, come quello del cambiamento di coordinate di una funzione quadratica (1827), per risolvere sistemi non lineari, introducendo la notazione del determinante *jacobiano* (1830), per risolvere equazioni differenziali e calcolare integrali (1827 e 1831).

Nella monumentale opera di Muir [11] è esposta una trattazione storica della teoria dei determinanti, con uno studio completo e critico dei 1859 lavori sull'argomento scritti fra il 1693 e il 1900. Muir ha anche fatto una raccolta sistematica [12] di tutte le proprietà dei determinanti note fino al 1933.

Allo sviluppo della teoria dei determinanti non segue in parallelo quello della teoria delle matrici. Occorre infatti arrivare alla metà del 19^o secolo per trovare uno studio autonomo sull'argomento: Hamilton nel 1853 e Cayley

nel 1854 e nel 1857 sviluppano le proprietà fondamentali dell'algebra delle matrici. Lo stesso termine di *matrice (matrix)* sembra sia stato utilizzato per la prima volta da Sylvester nel 1850. La teoria viene poi sviluppata negli anni successivi da Laguerre, Sylvester e Hermite. Fra il 1877 e il 1880 Frobenius raccoglie in modo sistematico tutti i risultati precedenti e ne sviluppa molti nuovi: a lui si devono alcuni concetti fondamentali, come quello di *rango* di una matrice. Agli inizi di questo secolo la parte fondamentale della teoria delle matrici è completata. Una cronologia assai dettagliata delle definizioni e dei teoremi fino agli anni 30 si trova nel libro di MacDuffee [10].

La nascita della nozione di combinazione lineare, di spazio vettoriale di dimensione finita, e la relazione (in notazione moderna)

$$\dim V + \dim W = \dim(V + W) + \dim(V \cap W)$$

si può far risalire a Grassman nel 1862. Ma solo nel 1888 Peano, disponendo del linguaggio insiemistico, dà una definizione di spazio vettoriale che corrisponde a quella attuale.

Per uno studio della storia dell'algebra lineare è importante quanto riportato nel libro di Dieudonné [3], in cui l'evolversi della teoria delle matrici e dei determinanti è inquadrato nel più generale sviluppo della matematica moderna.

Nella bibliografia che segue sono indicati alcuni testi, in parte classici, e in parte moderni, che trattano in modo specifico la teoria delle matrici. L'algebra lineare, che costituisce un argomento fondamentale anche per i corsi propedeutici di matematica, viene riportata anche nella maggior parte dei libri di algebra e di calcolo.

Bibliografia

[1] A. C. Aitken, *Determinants and Matrices*, Interscience, New York, 1956.

[2] S. Cherubino, *Calcolo delle matrici*, Edizioni Cremonese, Roma, 1957.

[3] J. Dieudonné, *Abrégé d'histoire des mathématiques 1700-1900*, Hermann, Paris, 1978.

[4] F. R. Gantmacher, *The Theory of Matrices*, vol. I e II. Chelsea, New York, 1959.

[5] P. R. Halmos, *Finite-Dimentional Vector Spaces*, Van Nostrand-Reinhold, Princeton, 1958.

[6] R. A. Horn, C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, 1985.

[7] P. Lancaster, *Theory of Matrices*, Academic Press New York, 1969.

[8] P. Lancaster, M. Tismenetsky, *The Theory of Matrices*, Academic Press New York, 1985.

[9] S. Lang, *Algebra lineare*, Boringhieri, Torino, 1968.

[10] C. C. MacDuffee, *The Theory of Matrices*, Chelsea, New York, 1946.

[11] T. Muir, *Theory of Determinants in the Historical Order of Development*, Dover, London, 1906.

[12] T. Muir, *A Treatise on the Theory of Determinants*, Dover, London, 1933.

[13] M. C. Pease, *Methods of Matrix Algebra*, Academic Press, New York, 1965.

Capitolo 2

AUTOVALORI E AUTOVETTORI

1. Definizioni

Siano $A \in \mathbf{C}^{n\times n}, \lambda \in \mathbf{C}$ e $\mathbf{x} \in \mathbf{C}^n, \mathbf{x} \neq \mathbf{0}$, tali che valga la relazione

$$A\mathbf{x} = \lambda\mathbf{x}. \tag{1}$$

Allora λ è detto *autovalore* di A ed $\mathbf{x}$ è detto *autovettore* corrispondente a λ. L'insieme degli autovalori di una matrice A costituisce lo *spettro* di A e il modulo massimo $\rho(A)$ degli autovalori di A è detto *raggio spettrale* di A.

Il sistema (1), che si può scrivere anche nella forma

$$(A - \lambda I)\mathbf{x} = \mathbf{0}, \tag{2}$$

ammette soluzioni non nulle se e solo se

$$\det(A - \lambda I) = 0. \tag{3}$$

Sviluppando $\det(A - \lambda I)$ risulta

$$\det(A - \lambda I) = P(\lambda) = a_0\lambda^n + a_1\lambda^{n-1} + \ldots + a_n,$$

in cui

$$a_0 = (\text{-}1)^n, \qquad a_i = (\text{-}1)^{n-i}\sigma_i, \quad i = 1, \ldots, n,$$

dove σ_i è la somma dei determinanti delle $\binom{n}{i}$ sottomatrici principali di A di ordine i. In particolare risulta:

$$a_1 = (\text{-}1)^{n-1}\operatorname{tr} A, \qquad a_n = \det A,$$

in cui si è indicato con $\operatorname{tr} A$ la *traccia* di A, cioè la somma degli elementi principali di A.

Dalle relazioni che legano i coefficienti e le radici di un'equazione algebrica risulta che:

$$\sum_{i=1}^{n} \lambda_i = \operatorname{tr} A \quad \text{e} \quad \prod_{i=1}^{n} \lambda_i = \det A. \tag{4}$$

Il polinomio $P(\lambda)$ è detto *polinomio caratteristico* di A e l'equazione $P(\lambda) = 0$ è detta *equazione caratteristica* di A.

Per il teorema fondamentale dell'algebra l'equazione caratteristica ha nel campo complesso n radici, tenendo conto della loro molteplicità. Quindi una matrice di ordine n ha, tenendo conto della loro molteplicità, n autovalori nel campo complesso.

Poiché gli autovettori sono soluzioni non nulle del sistema lineare omogeneo (2), un autovettore corrispondente ad un autovalore λ risulta determinato a meno di una costante moltiplicativa $\alpha \neq 0$, cioè se $\mathbf{x}$ è autovettore di A, anche $\alpha\mathbf{x}$ è autovettore di A, corrispondente allo stesso autovalore.

2.1 Esempio. Il polinomio caratteristico della matrice

$$A = \begin{bmatrix} 1 & 3 \\ 3 & 1 \end{bmatrix}$$

si ricava dal determinante

$$\det(A - \lambda I) = \det \begin{bmatrix} 1-\lambda & 3 \\ 3 & 1-\lambda \end{bmatrix} = \lambda^2 - 2\lambda - 8.$$

L'equazione caratteristica corrispondente

$$\lambda^2 - 2\lambda - 8 = 0$$

ha come radici $\lambda_1 = -2$ e $\lambda_2 = 4$, che sono gli autovalori della matrice A. L'autovettore corrispondente a $\lambda_1 = -2$ si calcola risolvendo il sistema (2) che in questo caso diventa

$$\begin{bmatrix} 3 & 3 \\ 3 & 3 \end{bmatrix} \quad \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \mathbf{0}.$$

Dalla prima equazione si ottiene

$$x_1 + x_2 = 0, \quad \text{da cui} \quad x_1 = -x_2,$$

da cui segue che qualunque vettore

$$\mathbf{x}_1 = \alpha \begin{bmatrix} 1 \\ \text{-}1 \end{bmatrix}$$

con $\alpha \neq 0$, è autovettore della matrice A corrispondente all'autovalore $\lambda_1 = -2$. L'autovettore corrispondente a $\lambda_2 = 4$ si determina risolvendo il sistema

$$\begin{bmatrix} \text{-}3 & 3 \\ 3 & \text{-}3 \end{bmatrix} \quad \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \mathbf{0}.$$

Dalla prima equazione si ottiene

$$-x_1 + x_2 = 0, \quad \text{da cui} \quad x_1 = x_2,$$

da cui segue che qualunque vettore

$$\mathbf{x}_1 = \alpha \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

con $\alpha \neq 0$, è autovettore della matrice A corrispondente all'autovalore $\lambda_2 = 4$. ∎

2. Proprietà degli autovalori

– Gli autovalori di una matrice A diagonale o triangolare (superiore o inferiore) sono uguali agli elementi principali. Infatti la matrice $A - \lambda I$ è ancora diagonale o triangolare e quindi il suo determinante è dato dal prodotto degli elementi principali.

– Se λ è un autovalore di una matrice A non singolare e $\mathbf{x}$ un autovettore corrispondente, allora risulta $\lambda \neq 0$ e $1/\lambda$ è autovalore di A^{-1} con $\mathbf{x}$ autovettore corrispondente. Infatti da

$$A\mathbf{x} = \lambda\mathbf{x}$$

si ha

$$\mathbf{x} = \lambda A^{-1}\mathbf{x}$$

e quindi

$$\lambda \neq 0 \quad \text{e} \quad A^{-1}\mathbf{x} = \frac{1}{\lambda}\,\mathbf{x}.$$

– Se λ è un autovalore di una matrice A, allora $\overline{\lambda}$ è autovalore di A^H e λ è autovalore di A^T. Infatti nel primo caso, poiché

$$\det A^H = \overline{\det A},$$

si ha

$$0 = \det(A - \lambda I) = \overline{\det(A - \lambda I)^H} = \overline{\det(A^H - \overline{\lambda} I)},$$

da cui

$$\det(A^H - \overline{\lambda} I) = 0.$$

Si procede in modo analogo per il secondo caso.

– Se λ è autovalore di una matrice unitaria A, cioè tale che $A^H A = AA^H = I$, allora risulta $|\lambda| = 1$. Infatti dalla relazione $A\mathbf{x} = \lambda\mathbf{x}$ si ha

$$(A\mathbf{x})^H = (\lambda\mathbf{x})^H$$

e quindi

$$\mathbf{x}^H A^H = \overline{\lambda}\mathbf{x}^H,$$

da cui si ha

$$\mathbf{x}^H A^H A\mathbf{x} = \overline{\lambda}\lambda\, \mathbf{x}^H\mathbf{x}.$$

Poiché A è unitaria, risulta

$$\mathbf{x}^H\mathbf{x} = \overline{\lambda}\lambda\mathbf{x}^H\mathbf{x},$$

e quindi, essendo $\mathbf{x}^H\mathbf{x} \neq 0$, segue

$$\overline{\lambda}\lambda = |\lambda|^2 = 1.$$

2.2 Esempio. La matrice G dell'esempio 1.1:

$$G = \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix}, \qquad \phi \in \mathbf{R},$$

è unitaria e quindi ha autovalori di modulo 1. Infatti dall'equazione caratteristica

$$\lambda^2 - 2\lambda\cos\phi + 1 = 0$$

risulta

$$\lambda_1 = \cos\phi + \mathbf{i}\sin\phi \quad \text{e} \quad \lambda_2 = \cos\phi - \mathbf{i}\sin\phi. \qquad \blacksquare$$

Particolarmente interessanti sono i polinomi di matrici. Sia

$$p(x) = \alpha_0 x^k + \alpha_1 x^{k-1} + \ldots + \alpha_k,$$

dove $\alpha_0, \alpha_1, \ldots, \alpha_k \in \mathbf{C}$, un polinomio di grado k nella variabile x e sia $A \in \mathbf{C}^{n\times n}$. Un *polinomio della matrice* A è una matrice della forma

$$p(A) = \alpha_0 A^k + \alpha_1 A^{k-1} + \ldots + \alpha_k I$$

(si veda l'esercizio 1.3). Se λ è un autovalore di A e $\mathbf{x}$ è un autovettore corrispondente, allora $p(\lambda)$ è autovalore di $p(A)$ e $\mathbf{x}$ è un autovettore corrispondente. Risulta infatti che

$$A^i\mathbf{x} = A^{i-1}A\mathbf{x} = A^{i-1}\lambda\mathbf{x} = \lambda A^{i-1}\mathbf{x} = \lambda A^{i-2}A\mathbf{x} = \ldots = \lambda^i\mathbf{x},$$

e quindi

$$p(A)\mathbf{x} = \alpha_0 A^k\mathbf{x} + \alpha_1 A^{k-1}\mathbf{x} + \ldots + \alpha_k\mathbf{x} = \alpha_0\lambda^k\mathbf{x} + \alpha_1\lambda^{k-1}\mathbf{x} + \ldots + \alpha_k\mathbf{x}$$
$$= p(\lambda)\mathbf{x}.$$

2.3 Esempio. La matrice

$$A = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

ha gli autovalori

$$\lambda_1 = 1 - \sqrt{2}, \quad \lambda_2 = 1, \quad \lambda_3 = 1 + \sqrt{2},$$

e i corrispondenti autovettori sono

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ -\sqrt{2} \\ 1 \end{bmatrix}, \quad \mathbf{x}_2 = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}, \quad \mathbf{x}_3 = \begin{bmatrix} 1 \\ \sqrt{2} \\ 1 \end{bmatrix}.$$

La matrice

$$B = 3A^2 - A + 2I = \begin{bmatrix} 7 & 5 & 3 \\ 5 & 10 & 5 \\ 3 & 5 & 7 \end{bmatrix},$$

ha gli autovalori

$$\mu_i = 3\lambda_i^2 - \lambda_i + 2, \quad i = 1, 2, 3,$$

cioè

$$\mu_1 = 10 - 5\sqrt{2}, \quad \mu_2 = 4, \quad \mu_3 = 10 + 5\sqrt{2};$$

gli autovettori di B sono gli stessi di A. ∎

2.4 Esempio. La matrice

$$A = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}$$

verifica la relazione

$$A^2 - 5A + 4I = 0. \tag{5}$$

Quindi, se λ è autovalore di A, $\lambda^2 - 5\lambda + 4$ è autovalore della matrice nulla. Perciò deve essere $\lambda^2 - 5\lambda + 4 = 0$, da cui segue che gli autovalori di A sono $\lambda_1 = 1$ e $\lambda_2 = 4$. Poiché uno di questi autovalori ha molteplicità 2,

e la somma degli autovalori, che è uguale alla traccia della matrice, è 6, λ_1 risulta di molteplicità 2.

La relazione (5) può essere utilizzata per calcolare la matrice inversa di A. Si ha infatti

$$A(5I - A) = 4I$$

da cui

$$A^{-1} = \frac{1}{4}\,(5I - A) = \frac{1}{4}\begin{bmatrix} 3 & -1 & -1 \\ -1 & 3 & -1 \\ -1 & -1 & 3 \end{bmatrix}. \qquad \blacksquare$$

Più in generale è possibile dimostrare il seguente

2.5 Teorema (di Cayley-Hamilton). *Sia $A \in \mathbf{C}^{n\times n}$ e sia $P(\lambda)$ il suo polinomio caratteristico. Allora*

$$P(A) = 0.$$

Dim. Per un $\lambda \in \mathbf{C}$ sia $C = A - \lambda I$. Sia B la matrice aggiunta di C, per cui vale (si veda l'esercizio 1.48)

$$CB = (\det C)I. \tag{6}$$

Gli elementi di B sono determinanti di sottomatrici di ordine $n-1$ della matrice $A-\lambda I$, e quindi sono polinomi in λ di grado al più $n-1$. La matrice B può essere così espressa

$$B = \lambda^{n-1}B_0 + \lambda^{n-2}B_1 + \ldots + B_{n-1},$$

dove B_j, $j = 0, 1, \ldots, n-1$ sono matrici di ordine n. Dalla (6) si ha

$$\begin{aligned}(\det C)I &= (A - \lambda I)(\lambda^{n-1}B_0 + \lambda^{n-2}B_1 + \ldots + B_{n-1}) \\ &= -\lambda^n B_0 + \lambda^{n-1}(AB_0 - B_1) + \lambda^{n-2}(AB_1 - B_2) + \ldots + AB_{n-1}.\end{aligned}$$

D'altra parte

$$\det C = \det(A - \lambda I) = P(\lambda) = a_0\lambda^n + a_1\lambda^{n-1} + \ldots + a_n,$$

per cui uguagliando termine a termine si ha:

$$\begin{aligned} a_0 I &= -B_0 \\ a_1 I &= AB_0 - B_1 \\ a_2 I &= AB_1 - B_2 \\ &\vdots \\ a_n I &= AB_{n-1}. \end{aligned}$$

Moltiplicando queste relazioni rispettivamente per $A^n, A^{n-1}, \ldots, I$ e sommando, si ottiene

$$a_0A^n + a_1A^{n-1} + a_2A^{n-2} + \ldots + a_nI$$
$$= -A^nB_0 + A^{n-1}(AB_0 - B_1) + A^{n-2}(AB_1 - B_2) + \ldots + AB_{n-1} = 0. \blacksquare$$

Dal teorema di Cayley-Hamilton segue quindi che una qualsiasi matrice A annulla il suo polinomio caratteristico $P(\lambda)$ e anche ogni altro polinomio di cui $P(\lambda)$ sia fattore.

Il polinomio *monico* (cioè con primo coefficiente uguale a 1) $\psi(\lambda)$ di grado minimo che è annullato da A è detto *polinomio minimo* di A ed è un fattore di $P(\lambda)$ e di ogni altro polinomio $p(\lambda)$ che sia annullato da A. Si ha infatti

$$p(\lambda) = \psi(\lambda)s(\lambda) + r(\lambda),$$

dove il grado di $r(\lambda)$ è minore di quello di $\psi(\lambda)$. Poiché

$$0 = p(A) = \psi(A)s(A) + r(A)$$

e $\psi(A) = 0$, ne segue che $r(A) = 0$, ed essendo $\psi(\lambda)$ il polinomio di grado minimo annullato da A, ne segue che $r(\lambda)$ è identicamente nullo.

Poiché $\psi(\lambda)$ è fattore di $P(\lambda)$, gli zeri di $\psi(\lambda)$ devono essere autovalori di A. Viceversa ogni autovalore di A è uno zero di $\psi(\lambda)$. Infatti se μ è autovalore di A, $\psi(\mu)$ è autovalore di $\psi(A)$, ed essendo $\psi(A) = 0$, ne segue che $\psi(\mu) = 0$.

Risulta perciò che $\psi(\lambda)$ è della forma

$$\psi(\lambda) = (\lambda - \lambda_1)^{n_1}(\lambda - \lambda_2)^{n_2} \cdots (\lambda - \lambda_p)^{n_p},$$

dove $\lambda_1, \lambda_2, \ldots, \lambda_p$ sono gli autovalori distinti di A, e $n_1 + n_2 + \ldots + n_p \leq n$. Se la matrice A ha tutti gli autovalori distinti, allora è

$$P(\lambda) = (-1)^n\psi(\lambda).$$

2.6 Esempio. La matrice A dell'esempio 2.4 annulla il polinomio

$$\psi(\lambda) = \lambda^2 - 5\lambda + 4,$$

che è il suo polinomio minimo. Infatti non esiste alcuna costante α tale che $A + \alpha I = 0$ e quindi nessun polinomio di grado 1 che sia annullato da A. $\blacksquare$

3. Proprietà degli autovettori

2.7 Teorema. *Autovettori corrispondenti ad autovalori distinti sono linearmente indipendenti.*

Dim. Siano $\lambda_1, \ldots, \lambda_m,\ m \leq n,\ m$ autovalori di $A \in \mathbf{C}^{n\times n}$ a due a due distinti, e siano $\mathbf{x}_1, \ldots, \mathbf{x}_m$ i corrispondenti autovettori. Si procede per induzione su m.

Per $m = 1, \mathbf{x}_1 \neq 0$, quindi $\mathbf{x}_1$ è linearmente indipendente.

Per $m > 1$, si supponga per assurdo che esista una combinazione lineare dei vettori $\mathbf{x}_1, \ldots, \mathbf{x}_m$ tale che

$$\sum_{i=1}^{m} \alpha_i\ \mathbf{x}_i = \mathbf{0}, \tag{7}$$

in cui non tutti gli α_i siano nulli e sia j tale che $\alpha_j \neq 0$. In tal caso esiste almeno un altro indice $k \neq j$, per cui $\alpha_k \neq 0$; altrimenti, se ciò non fosse, seguirebbe che $\mathbf{x}_j = 0$. Moltiplicando entrambi i membri della (7) per A, si ottiene

$$\mathbf{0} = A\sum_{i=1}^{m} \alpha_i\ \mathbf{x}_i = \sum_{i=1}^{m} \alpha_i\ A\mathbf{x}_i = \sum_{i=1}^{m} \alpha_i\ \lambda_i\ \mathbf{x}_i, \tag{8}$$

moltiplicando entrambi i membri della (7) per λ_j, si ottiene

$$\mathbf{0} = \lambda_j \sum_{i=1}^{m} \alpha_i\ \mathbf{x}_i = \sum_{i=1}^{m} \alpha_i\ \lambda_j\ \mathbf{x}_i. \tag{9}$$

Sottraendo membro a membro la (9) dalla (8), si ha:

$$\mathbf{0} = \sum_{i=1}^{m} \alpha_i(\lambda_i - \lambda_j)\mathbf{x}_i = \sum_{\substack{i=1\\ i\neq j}}^{m} \alpha_i(\lambda_i - \lambda_j)\mathbf{x}_i.$$

Si ottiene così una combinazione lineare nulla degli $m-1$ autovettori $\mathbf{x}_i \neq \mathbf{0},\quad i = 1, \ldots, m,\ i \neq j$, in cui $\lambda_i - \lambda_j \neq 0$ per $i \neq j$ e gli α_i per $i \neq j$ non sono tutti nulli, essendo $\alpha_k \neq 0$, ciò che è assurdo perché per l'ipotesi induttiva gli $m-1$ vettori sono linearmente indipendenti. ∎

Dal teorema 2.7 risulta che se una matrice A di ordine n ha n autovalori tutti distinti, allora A ha n autovettori linearmente indipendenti. Se la matrice A non ha n autovalori distinti, A può non avere n autovettori linearmente indipendenti, come risulta nell'esempio seguente.

2.8 Esempio. La matrice

$$A = \begin{bmatrix} 4 & 1 \\ -1 & 2 \end{bmatrix},$$

il cui polinomio caratteristico è $(\lambda-3)^2$, ha 3 come autovalore di molteplicità 2. Poiché ad esso corrispondono solo autovettori della forma

$$\alpha \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \quad \alpha \neq 0,$$

la matrice A non può avere due autovettori linearmente indipendenti. ∎

Le matrici con n autovalori distinti non sono le sole ad avere n autovettori linearmente indipendenti. Ad esempio la matrice identica I_n che ha 1 come autovalore di molteplicità n, ha i vettori $\mathbf{e}_i \in \mathbf{C}^n$, $i = 1, \ldots, n$, della base canonica di $\mathbf{C}^n$ come autovettori.

Nel caso in cui ad uno stesso autovalore corrispondano più autovettori linearmente indipendenti, essi generano un sottospazio lineare i cui vettori non nulli sono tutti autovettori della matrice corrispondenti allo stesso autovalore. Vale infatti il seguente

2.9 Teorema. *Sia $A \in \mathbf{C}^{n\times n}$, e siano $\mathbf{x}_1, \mathbf{x}_2 \ldots, \mathbf{x}_k$ k autovettori linearmente indipendenti, corrispondenti ad uno stesso autovalore λ di A. Allora un vettore $\mathbf{y} \in \mathbf{C}^n$, $\mathbf{y} \neq \mathbf{0}$, della forma*

$$\mathbf{y} = \sum_{j=1}^{k} \alpha_j \, \mathbf{x}_j$$

è autovettore di A.

Dim. Si ha infatti

$$A\mathbf{y} = A \sum_{j=1}^{k} \alpha_j \, \mathbf{x}_j = \sum_{j=1}^{k} \alpha_j A\mathbf{x}_j = \sum_{j=1}^{k} \alpha_j \lambda \mathbf{x}_j = \lambda \sum_{j=1}^{k} \alpha_j \, \mathbf{x}_j = \lambda \, \mathbf{y}. \quad ∎$$

2.10 Definizione. La molteplicità di un autovalore λ come radice dell'equazione caratteristica, è indicata con $\sigma(\lambda)$, ed è detta *molteplicità algebrica* di λ. Il massimo numero di autovettori linearmente indipendenti corrispondenti a λ è indicato con $\tau(\lambda)$ ed è detto *molteplicità geometrica* di λ. ∎

La molteplicità geometrica $\tau(\lambda)$ è uguale alla dimensione del sottospazio vettoriale generato dagli autovettori corrispondenti a λ, e quindi uguale alla dimensione dello spazio

$$N(A - \lambda I) = \{ \, \mathbf{x} \in \mathbf{C}^n \; : \; (A - \lambda I)\mathbf{x} = \mathbf{0} \, \},$$

nucleo di $A - \lambda I$. È evidente che

$$1 \leq \sigma(\lambda) \leq n \quad \text{e} \quad 1 \leq \tau(\lambda) \leq n.$$

2.11 Teorema. *Vale la disuguaglianza*

$$\tau(\lambda) \leq \sigma(\lambda).$$

Dim. Sia μ un autovalore di A con molteplicità algebrica $\sigma = \sigma(\mu)$ e geometrica $\tau = \tau(\mu)$. Per la (7) del capitolo 1 si ha

$$\text{rango di } (A - \mu I) = n - \text{ dim } N(A - \mu I) = n - \tau,$$

e quindi tutte le sottomatrici principali di ordine superiore a $n - \tau$ della matrice $A - \mu I$ sono singolari. Poiché il coefficiente del termine di grado i del polinomio caratteristico di una matrice è dato, a meno del segno, dalla somma dei determinanti delle sue sottomatrici principali di ordine $n - i$, ne segue che il polinomio caratteristico di $A - \mu I$ risulta della forma

$$p(\lambda) = \det[(A - \mu I) - \lambda I] = a_0\lambda^n + a_1\lambda^{n-1} + \ldots + a_k\lambda^{n-k},$$

dove $k \leq n-\tau$. Perciò l'equazione $p(\lambda) = 0$ ha la radice $\lambda = 0$ di molteplicità $n - k \geq \tau$. Posto $x = \lambda + \mu$, si ha

$$\begin{aligned}\det[(A - \mu I) - \lambda I] &= \det(A - xI) \\ &= a_0(x - \mu)^n + a_1(x - \mu)^{n-1} + \ldots + a_k(x - \mu)^{n-k},\end{aligned}$$

per cui la molteplicità di μ come radice dell'equazione caratteristica è $\sigma \geq \tau$. ∎

2.12 Esempio. La matrice $A \in \mathbf{R}^{n\times n}$, definita da

$$a_{ij} = \begin{cases} 1 & \text{per } j = i, \\ 1 & \text{per } j = i + 1, \\ 0 & \text{altrimenti,} \end{cases}$$

cioè

$$A = \begin{bmatrix} 1 & 1 & & & \\ & 1 & 1 & & \\ & & \ddots & \ddots & \\ & & & 1 & 1 \\ & & & & 1 \end{bmatrix},$$

ha 1 come autovalore di molteplicità algebrica n a cui corrispondono solo autovettori del tipo $\mathbf{x} = \alpha\mathbf{e}_1, \alpha \neq 0$. In questo caso risulta quindi $\tau(1) = 1$ e $\sigma(1) = n$. ∎

Nel caso della matrice identica I_n, che ha 1 come autovalore di molteplicità algebrica n, risulta $\tau(1) = \sigma(1) = n$.

4. Trasformazioni per similitudine

Data una base $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$ di $\mathbf{C}^n$ e una matrice $A \in \mathbf{C}^{n\times n}$, viene individuata univocamente l'applicazione lineare $\mathcal{L}: \mathbf{C}^n \to \mathbf{C}^n$, definita sugli elementi della base da

$$\mathcal{L}(\mathbf{u}_j) = \sum_{i=1}^{n} a_{ij}\ \mathbf{u}_i. \tag{10}$$

L'applicazione $\mathcal{L}$ risulta naturalmente estesa per linearità a tutti i vettori $\mathbf{x} \in \mathbf{C}^n$. Si considerino le matrici U e W le cui colonne sono rispettivamente i vettori $\mathbf{u}_j$ e $\mathcal{L}(\mathbf{u}_j)$. Allora la (10) può essere rappresentata nella forma:

$$W = UA. \tag{11}$$

Una stessa applicazione lineare $\mathcal{L}$ può essere rappresentata su due basi diverse $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$ e $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ e quindi da due matrici A e B, in generale diverse. Se V e Z sono le matrici le cui colonne sono rispettivamente formate dai vettori $\mathbf{v}_j$ e $\mathcal{L}(\mathbf{v}_j)$, vale la relazione analoga alla (11):

$$Z = VB. \tag{12}$$

Si vuole ora determinare la relazione che lega le due matrici A e B. Si supponga che i vettori $\mathbf{u}_i$ e $\mathbf{v}_i$, $i = 1, \ldots, n$, siano legati dalla relazione

$$\mathbf{v}_j = \sum_{i=1}^{n} s_{ij}\ \mathbf{u}_i, \quad j = 1, 2, \ldots, n, \tag{13}$$

che in notazione matriciale è rappresentata da

$$V = US,$$

dove la matrice non singolare S è la matrice del *cambiamento di base*. Sostituendo quest'ultima relazione nella (12) si ha

$$Z = USB. \tag{14}$$

D'altra parte, per la linearità dell'applicazione $\mathcal{L}$, dalla (13) si ha:

$$\mathcal{L}(\mathbf{v}_j) = \mathcal{L}\,(\sum_{i=1}^{n} s_{ij}\ \mathbf{u}_i) = \sum_{i=1}^{n} s_{ij}\ \ \mathcal{L}(\mathbf{u}_i)$$

e in notazione matriciale

$$Z = WS. \tag{15}$$

Sostituendo la (11) nella (15) si ha:

$$Z = UAS$$

da cui, per confronto con la (14), poiché U è non singolare, si ha:

$$AS = SB$$

e quindi

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

Se le due basi $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$ e $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ sono basi ortonormali, le matrici U e V sono unitarie, e allora anche la matrice S è unitaria; infatti:

$$I = V^H V = (US)^H (US) = S^H U^H US = S^H S.$$

In tal caso le due matrici A e B soddisfano la relazione

$$A = SBS^H.$$

2.13 Definizione. Due matrici $A, B \in \mathbf{C}^{n\times n}$ si dicono *simili* se esiste una matrice non singolare S per cui

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

La trasformazione che associa la matrice A alla matrice B viene detta *trasformazione per similitudine*. Se la matrice S è unitaria, la trasformazione viene detta *trasformazione per similitudine unitaria*. ∎

Si noti che la trasformazione per similitudine è una relazione di equivalenza, in quanto gode delle proprietà riflessiva, simmetrica e transitiva.

Data una matrice A, si consideri l'applicazione lineare $\mathcal{L}_A$ individuata dalla A e dalla base canonica $\mathbf{e}_i,\ i = 1, \ldots, n$ di $\mathbf{C}^n$; allora per ogni vettore $\mathbf{x} \in \mathbf{C}^n$ risulta

$$\mathcal{L}_A(\mathbf{x}) = A\mathbf{x}.$$

Quindi dalla relazione (1), riformulata in termini di applicazioni lineari, risulta che gli autovettori $\mathbf{x}$ di A sono i vettori che vengono trasformati dall'applicazione $\mathcal{L}_A$ in vettori proporzionali a se stessi. Cioè ogni autovettore individua una retta che è invariante per la trasformazione lineare $\mathcal{L}_A$. Le proprietà degli autovalori e autovettori sono dunque proprietà intrinseche dell'applicazione lineare e non legate solamente alla matrice che la rappresenta in una particolare base. Vale infatti il seguente teorema.

2.14 Teorema. *Due matrici simili hanno gli stessi autovalori con le stesse molteplicità algebriche e geometriche.*

Dim. Siano A e B matrici simili, cioè tali che $A = SBS^{-1}$. Si ha:

$$\begin{aligned}\det(A - \lambda I) &= \det(SBS^{-1} - \lambda SS^{-1}) = \det\left[S(B - \lambda I)S^{-1}\right] \\ &= \det S \det(B - \lambda I) \det(S^{-1}) = \det(B - \lambda I)\end{aligned}$$

per cui le due matrici hanno lo stesso polinomio caratteristico e quindi hanno gli stessi autovalori con le stesse molteplicità algebriche. Se $\mathbf{x}$ è autovettore di A corrispondente all'autovalore λ, risulta:

$$SBS^{-1}\mathbf{x} = \lambda\mathbf{x}$$

e quindi

$$BS^{-1}\mathbf{x} = \lambda S^{-1}\mathbf{x}.$$

Perciò il vettore $\mathbf{y} = S^{-1}\mathbf{x}$ è autovettore di B corrispondente a λ. Inoltre, essendo S^{-1} non singolare, se $\mathbf{x}_i,\ i = 1, \ldots, \tau(\lambda)$, sono autovettori linearmente indipendenti di A, anche $\mathbf{y}_i = S^{-1}\mathbf{x}_i,\ i = 1, \ldots, \tau(\lambda)$ sono linearmente indipendenti. Quindi A e B hanno gli stessi autovalori con le stesse molteplicità geometriche. ∎

Da questo teorema risulta che se due matrici sono simili, hanno uguali la traccia e il determinante.

2.15 Definizione. Una matrice A simile ad una matrice diagonale D si dice *diagonalizzabile*. ∎

2.16 Teorema. *Una matrice A di ordine n è diagonalizzabile se e solo se ha n autovettori linearmente indipendenti. Inoltre le colonne della matrice S, per cui $S^{-1}AS$ è diagonale, sono gli autovettori di A.*

Dim. Si suppone dapprima che A abbia n autovettori linearmente indipendenti $\mathbf{x}_1, \ldots, \mathbf{x}_n$, corrispondenti agli autovalori $\lambda_1, \ldots, \lambda_n$. Siano D la matrice diagonale avente λ_i come i-esimo elemento principale, e S la matrice la cui i-esima colonna è uguale a $\mathbf{x}_i$. Dalla relazione

$$A\mathbf{x}_i = \lambda_i\mathbf{x}_i, \quad i = 1, 2, \ldots, n,$$

si ha anche che

$$AS = SD. \tag{16}$$

Essendo S non singolare, perché formata da colonne linearmente indipendenti, esiste S^{-1}; quindi dalla (16) si ha

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

Viceversa, sia $A = SDS^{-1}$, D matrice diagonale con gli autovalori di A come elementi principali. Allora risulta $AS = SD$. Indicando con $\mathbf{s}_1, \ldots, \mathbf{s}_n$ le colonne di S, si ha:

$$A\,[\mathbf{s}_1 \,|\, \mathbf{s}_2 \,|\, \ldots \,|\, \mathbf{s}_n] = [\lambda_1\mathbf{s}_1 \,|\, \lambda_2\mathbf{s}_2 \,|\, \ldots \,|\, \lambda_n\mathbf{s}_n]$$

e quindi

$$A\mathbf{s}_i = \lambda_i\mathbf{s}_i, \quad i = 1, 2, \ldots, n.$$

Perciò le colonne di S sono n autovettori di A, che risultano linearmente indipendenti, perché S è non singolare. ∎

2.17 Esempio. Le matrici A e $B = 3A^2 - A + 2I$ dell'esempio 2.3, che hanno tre autovalori distinti, sono diagonalizzabili dalla stessa trasformazione per similitudine, in quanto hanno gli stessi tre autovettori linearmente indipendenti. Posto

$$S = \begin{bmatrix} 1 & 1 & 1 \\ -\sqrt{2} & 0 & \sqrt{2} \\ 1 & -1 & 1 \end{bmatrix},$$

si ha

$$S^{-1} = \begin{bmatrix} 1/4 & -\sqrt{2}/4 & 1/4 \\ 1/2 & 0 & -1/2 \\ 1/4 & \sqrt{2}/4 & 1/4 \end{bmatrix}$$

e

$$A = S \begin{bmatrix} 1-\sqrt{2} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1+\sqrt{2} \end{bmatrix} S^{-1},$$

$$B = S \begin{bmatrix} 10-5\sqrt{2} & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 10+5\sqrt{2} \end{bmatrix} S^{-1},$$

Anche la matrice A dell'esempio 2.4, pur non avendo 3 autovalori distinti ha 3 autovettori linearmente indipendenti, ed è quindi diagonalizzabile. Posto

$$S = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ -1 & -1 & 1 \end{bmatrix},$$

si ha

$$S^{-1} = \frac{1}{3} \begin{bmatrix} -1 & 2 & -1 \\ 2 & -1 & -1 \\ 1 & 1 & 1 \end{bmatrix}$$

e

$$A = S \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 4 \end{bmatrix} S^{-1}. \qquad \blacksquare$$

5. Forme canoniche

Dai teoremi 2.7 e 2.16 segue che se una matrice ha tutti gli autovalori distinti, allora è diagonalizzabile, in quanto ha n autovettori linearmente indipendenti. Se una matrice non ha tutti gli autovalori distinti, può non essere diagonalizzabile e questo accade se per almeno un autovalore di A la molteplicità geometrica è minore della corrispondente molteplicità algebrica. A tale proposito vale il seguente teorema (per la dimostrazione si veda [3]).

2.18 Teorema (Forma canonica o normale di Jordan). *Sia $A \in \mathbf{C}^{n\times n}$ e siano λ_i, $i = 1, \ldots, p$, i suoi autovalori distinti, con molteplicità algebrica $\sigma(\lambda_i)$ e molteplicità geometrica $\tau(\lambda_i)$. Allora A è simile ad una matrice diagonale a blocchi*

$$J = \begin{bmatrix} J_1 & & & \\ & J_2 & & \\ & & \ddots & \\ & & & J_p \end{bmatrix},$$

in cui il blocco J_i, relativo all'autovalore λ_i, ha ordine $\sigma(\lambda_i)$ ed è a sua volta diagonale a blocchi

$$J_i = \begin{bmatrix} C_i^{(1)} & & & \\ & C_i^{(2)} & & \\ & & \ddots & \\ & & & C_i^{(\tau(\lambda_i))} \end{bmatrix}, \quad i = 1, 2, \ldots, p,$$

e ognuno dei $\tau(\lambda_i)$ blocchi è della forma

$$C_i^{(j)} = \begin{bmatrix} \lambda_i & 1 & & \\ & \ddots & \ddots & \\ & & \lambda_i & 1 \\ & & & \lambda_i \end{bmatrix} \in \mathbf{C}^{\nu_i^{(j)} \times \nu_i^{(j)}}, \quad j = 1, 2, \ldots, \tau(\lambda_i),$$

dove gli interi $\nu_i^{(j)}$ sono tali che

$$\sum_{j=1}^{\tau(\lambda_i)} \nu_i^{(j)} = \sigma(\lambda_i).$$

La matrice J è detta forma canonica (o normale) di Jordan della matrice A, ed è unica, a meno dell'ordinamento dei blocchi che la compongono. ∎

Se gli autovalori λ_i di A sono tutti distinti, i blocchi J_i hanno tutti ordine 1, e quindi la matrice è diagonalizzabile. Se invece gli autovalori non sono tutti distinti, ma A ha n autovettori linearmente indipendenti, allora i blocchi J_i sono diagonali, e anche in questo caso la matrice è diagonalizzabile.

2.19 Esempio. Le matrici

$$A_1 = \begin{bmatrix} -4 & 7 & 4 & -8 & 6 & -3 \\ -5 & 7 & 5 & -8 & 6 & -3 \\ -4 & 4 & 6 & -7 & 6 & -3 \\ -3 & 3 & 3 & -4 & 6 & -3 \\ -2 & 2 & 2 & -4 & 6 & -2 \\ -1 & 1 & 1 & -2 & 2 & 1 \end{bmatrix}$$

e

$$A_2 = \begin{bmatrix} -4 & 12 & -10 & 8 & -6 & 4 \\ -5 & 12 & -9 & 8 & -6 & 4 \\ -4 & 8 & -6 & 8 & -6 & 4 \\ -3 & 6 & -16 & 8 & -5 & 4 \\ -2 & 4 & -4 & 4 & -2 & 4 \\ -1 & 2 & -2 & 2 & -2 & 4 \end{bmatrix}$$

hanno entrambe l'autovalore $\lambda = 2$ di molteplicità algebrica 6 e geometrica 3, e risulta

$$A_1 = SJ'S^{-1} = S \begin{bmatrix} 2 & 1 & 0 & & & \\ 0 & 2 & 1 & & & \\ 0 & 0 & 2 & & & \\ & & & 2 & 1 & \\ & & & 0 & 2 & \\ & & & & & 2 \end{bmatrix} S^{-1}$$

e

$$A_2 = SJ''S^{-1} = S \begin{bmatrix} 2 & 1 & & & & \\ 0 & 2 & & & & \\ & & 2 & 1 & & \\ & & 0 & 2 & & \\ & & & & 2 & 1 \\ & & & & 0 & 2 \end{bmatrix} S^{-1}.$$

In entrambi i casi si ha

$$S = \begin{bmatrix} 6 & 5 & 4 & 3 & 2 & 1 \\ 5 & 5 & 4 & 3 & 2 & 1 \\ 4 & 4 & 4 & 3 & 2 & 1 \\ 3 & 3 & 3 & 3 & 2 & 1 \\ 2 & 2 & 2 & 2 & 2 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}, \qquad S^{-1} = \begin{bmatrix} 1 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & -1 & 2 & -1 & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{bmatrix}.$$

■

Se la matrice A ha elementi reali, allora esiste una *forma normale reale di Jordan* di A, analoga a quella definita nel teorema 2.18, in cui i blocchi $C_i^{(j)}$ relativi ad autovalori reali hanno la stessa forma che nel teorema 2.18, mentre i blocchi $C_i^{(j)}$ relativi ad autovalori complessi risultano così modificati: in corrispondenza ad ogni coppia $\lambda_i = a_i + \mathbf{i}b_i$ e $\overline{\lambda}_i = a_i - \mathbf{i}b_i$ di autovalori complessi e coniugati di A le sottomatrici $C_i^{(j)}$ sono bidiagonali a blocchi della forma

$$C_i^{(j)} = \begin{bmatrix} E_i & I_2 & & & \\ & E_i & I_2 & & \\ & & \ddots & \ddots & \\ & & & \ddots & I_2 \\ & & & & E_i \end{bmatrix},$$

dove

$$E_i = \begin{bmatrix} a_i & -b_i \\ b_i & a_i \end{bmatrix}.$$

2.20 Esempio. La matrice

$$A = \begin{bmatrix} 8 & -16 & 13 & -3 \\ 6 & -12 & 10 & -2 \\ 4 & -9 & 9 & -3 \\ 2 & -5 & 5 & -1 \end{bmatrix}$$

ha la forma normale di Jordan

$$A = SJS^{-1} = S \begin{bmatrix} 1+\mathbf{i} & 1 & 0 & 0 \\ 0 & 1+\mathbf{i} & 0 & 0 \\ 0 & 0 & 1-\mathbf{i} & 1 \\ 0 & 0 & 0 & 1-\mathbf{i} \end{bmatrix} S^{-1},$$

dove

$$S = \begin{bmatrix} 4-3\mathbf{i} & 2-\mathbf{i} & 4+3\mathbf{i} & 2+\mathbf{i} \\ 3-3\mathbf{i} & 2-\mathbf{i} & 3+3\mathbf{i} & 2+\mathbf{i} \\ 2-2\mathbf{i} & 2-\mathbf{i} & 2+2\mathbf{i} & 2+\mathbf{i} \\ 1-\mathbf{i} & 1-\mathbf{i} & 1+\mathbf{i} & 1+\mathbf{i} \end{bmatrix}$$

e

$$S^{-1} = \frac{1}{2}\begin{bmatrix} 1-\mathbf{i} & -1+2\mathbf{i} & -\mathbf{i} & 0 \\ 0 & -1 & 2-\mathbf{i} & -1+2\mathbf{i} \\ 1+\mathbf{i} & -1-2\mathbf{i} & \mathbf{i} & 0 \\ 0 & -1 & 2+\mathbf{i} & -1-2\mathbf{i} \end{bmatrix}.$$

La matrice A, avendo elementi reali, può essere rappresentata anche nella forma normale reale di Jordan

$$A = ZJ_RZ^{-1} = Z\begin{bmatrix} 1 & -1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 1 & 1 \end{bmatrix} Z^{-1},$$

dove

$$Z = \begin{bmatrix} 4 & 3 & 2 & 1 \\ 3 & 3 & 2 & 1 \\ 2 & 2 & 2 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}, \qquad Z^{-1} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix}.$$

■

Dalla forma normale di Jordan si può ricavare il polinomio minimo di A. Infatti posto $A = SJS^{-1}$, dove J è la forma normale di Jordan di A, si ha per ogni intero k

$$A^k = \underbrace{SJS^{-1}SJS^{-1}\dots SJS^{-1}}_{k \text{ volte}} = SJ^kS^{-1}$$

e quindi per ogni polinomio $P(A)$ si ha

$$P(A) = SP(J)S^{-1}.$$

In particolare per il polinomio minimo $\psi(\lambda)$ di A, risulta

$$\psi(A) = S\psi(J)S^{-1},$$

e quindi il polinomio minimo di A e quello di J coincidono. Per la struttura diagonale a blocchi di J si ha

$$\psi(J) = \begin{bmatrix} \psi(J_1) & & & \\ & \psi(J_2) & & \\ & & \ddots & \\ & & & \psi(J_p) \end{bmatrix}.$$

Perciò

$$\psi(\lambda) = (\lambda-\lambda_1)^{n_1}(\lambda-\lambda_2)^{n_2}\dots(\lambda-\lambda_p)^{n_p},$$

deve avere gli esponenti $n_i,\ i = 1, 2, \ldots, p$, tali che $(\lambda - \lambda_i)^{n_i}$ sia il polinomio minimo di J_i, cioè gli n_i devono essere gli interi più piccoli per cui il polinomio $(\lambda - \lambda_i)^{n_i}$ sia annullato contemporaneamente da tutte le sottomatrici $C_i^{(j)},\ j = 1, 2, \ldots, \tau(\lambda_i)$. Ciò è vero se e solo se n_i è la dimensione massima delle sottomatrici $C_i^{(j)}$, per $j = 1, 2, \ldots, \tau(\lambda_i)$.

2.21 Esempio. Il polinomio minimo della matrice A_1 dell'esempio 2.19 è dato da

$$\psi(\lambda) = (\lambda - 2)^3,$$

e quello della matrice A_2 è dato da

$$\psi(\lambda) = (\lambda - 2)^2.$$ ∎

Fra le trasformazioni per similitudine che associano alla matrice B la matrice $A = SBS^{-1}$, hanno particolare importanza quelle per cui S è unitaria, cioè $S^H S = SS^H = I$. Il teorema che segue mostra come sia possibile, mediante una trasformazione per similitudine unitaria, ricondurre una qualsiasi matrice a una forma triangolare superiore.

2.22 Teorema (forma canonica o normale di Schur). *Sia $A \in \mathbf{C}^{n\times n}$ e siano $\lambda_1, \ldots, \lambda_n$ i suoi autovalori. Allora esiste una matrice unitaria U e una matrice triangolare superiore T i cui elementi principali sono i λ_i, tali che*

$$A = UTU^H.$$

Dim. Si procede per induzione. Per $n = 1$ la tesi vale con $U = [1]$. Per $n > 1$, sia $\mathbf{x}_1$ l'autovettore normalizzato corrispondente all'autovalore λ_1 e sia S lo spazio generato da $\mathbf{x}_1$. Indicata con $\mathbf{y}_2, \ldots, \mathbf{y}_n$ una base ortonormale dello spazio $S^\perp$, la matrice

$$Q = [\,\mathbf{x}_1 \,|\, \mathbf{y}_2 \,|\, \ldots \,|\, \mathbf{y}_n\,]$$

è unitaria e $Q^H \mathbf{x}_1 = \mathbf{e}_1$. Si considera la matrice

$$B = Q^H A Q$$

la cui prima colonna è

$$B\mathbf{e}_1 = Q^H A Q \mathbf{e}_1 = Q^H A \mathbf{x}_1 = Q^H \lambda_1 \mathbf{x}_1 = \lambda_1 Q^H \mathbf{x}_1 = \lambda_1 \mathbf{e}_1$$

e quindi B può essere partizionata nel modo seguente:

$$B = \begin{bmatrix} \lambda_1 & \mathbf{c}^H \\ \\ \mathbf{0} & A_1 \end{bmatrix},$$

dove $\mathbf{c} \in \mathbf{C}^{n-1}$ e $A_1 \in \mathbf{C}^{(n-1)\times(n-1)}$. Per l'ipotesi induttiva esiste una matrice unitaria $U_1 \in \mathbf{C}^{(n-1)\times(n-1)}$ tale che

$$A_1 = U_1 A_2 U_1^H,$$

dove $A_2 \in \mathbf{C}^{(n-1)\times(n-1)}$ è triangolare superiore. Allora risulta

$$A = QBQ^H = Q \begin{bmatrix} \lambda_1 & \mathbf{c}^H \\ \mathbf{0} & A_1 \end{bmatrix} Q^H = Q \begin{bmatrix} \lambda_1 & \mathbf{c}^H \\ \mathbf{0} & U_1 A_2 U_1^H \end{bmatrix} Q^H.$$

Indicando con $U_2 \in \mathbf{C}^{n\times n}$ la matrice unitaria

$$U_2 = \begin{bmatrix} 1 & \mathbf{0}^H \\ \mathbf{0} & U_1 \end{bmatrix},$$

si ha:

$$A = QU_2 \begin{bmatrix} \lambda_1 & \mathbf{c}^H U_1 \\ \mathbf{0} & A_2 \end{bmatrix} U_2^H Q^H.$$

Poiché la matrice $U = QU_2$ è ancora unitaria in quanto prodotto di matrici unitarie, risulta

$$A = U \begin{bmatrix} \lambda_1 & \mathbf{c}^H U_1 \\ \mathbf{0} & A_2 \end{bmatrix} U^H.$$

da cui la tesi, essendo A_2 matrice triangolare superiore. ∎

2.23 Esempio. La matrice

$$A = \frac{1}{2} \begin{bmatrix} 5 & -5 & 1 & -1 \\ 5 & -5 & 3 & 1 \\ -1 & -1 & -1 & -1 \\ 3 & -1 & 1 & 1 \end{bmatrix}$$

ha l'autovalore $\lambda_1 = \mathbf{i}$ con il corrispondente autovettore normalizzato

$$\mathbf{x}_1 = \frac{1}{2} [\, 1 \,,\, 1 \,,\, \mathbf{i} \,,\, -\mathbf{i} \,]^T.$$

Si considerano altri tre vettori $\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4 \in \mathbf{C}^4$ linearmente indipendenti fra di loro e da $\mathbf{x}_1$:

$$\mathbf{x}_2 = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{x}_3 = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix}, \quad \mathbf{x}_4 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}.$$

Si costruiscono poi, a partire dai vettori $\mathbf{x}_i,\ i = 1, ..., 4$, con il metodo di Gram-Schmidt, tre vettori $\mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4$

$$\mathbf{y}_2 = \frac{1}{2}\begin{bmatrix} 1 \\ 1 \\ \mathbf{-i} \\ \mathbf{i} \end{bmatrix}, \quad \mathbf{y}_3 = \frac{1}{2}\begin{bmatrix} \text{-}1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \quad \mathbf{y}_4 = \frac{1}{2}\begin{bmatrix} 1 \\ \text{-}1 \\ 1 \\ 1 \end{bmatrix},$$

tali che la matrice

$$Q = [\mathbf{x}_1 \mid \mathbf{y}_2 \mid \mathbf{y}_3 \mid \mathbf{y}_4] = \frac{1}{2}\begin{bmatrix} 1 & 1 & \text{-}1 & 1 \\ 1 & 1 & 1 & \text{-}1 \\ \mathbf{i} & \mathbf{-i} & 1 & 1 \\ \mathbf{-i} & \mathbf{i} & 1 & 1 \end{bmatrix}$$

è unitaria. Si ha poi

$$B = Q^H A Q = \begin{bmatrix} \mathbf{i} & 0 & \text{-}2 & 3+\mathbf{i} \\ 0 & \mathbf{-i} & \text{-}2 & 3-\mathbf{i} \\ 0 & 0 & 0 & 1 \\ 0 & 0 & \text{-}1 & 0 \end{bmatrix} = \begin{bmatrix} T_1 & C \\ O & A_1 \end{bmatrix},$$

in cui $T_1 \in \mathbf{C}^{2\times 2}$ è triangolare superiore (più precisamente in questo caso T_1 risulta diagonale). Si deve quindi riapplicare il procedimento alla matrice

$$A_1 = \begin{bmatrix} 0 & 1 \\ \text{-}1 & 0 \end{bmatrix}$$

che ha ancora l'autovalore $\mathbf{i}$, con l'autovettore normalizzato

$$\mathbf{z}_1 = \frac{1}{\sqrt{2}}\begin{bmatrix} \mathbf{-i} \\ 1 \end{bmatrix}.$$

Con il metodo di Gram-Schmidt si determina il vettore

$$\mathbf{z}_2 = \frac{1}{\sqrt{2}}\begin{bmatrix} \mathbf{i} \\ 1 \end{bmatrix}.$$

tale che la matrice

$$Q_1 = \frac{1}{\sqrt{2}}\begin{bmatrix} \mathbf{-i} & \mathbf{i} \\ 1 & 1 \end{bmatrix}$$

è unitaria e si ha

$$B_1 = Q_1^H A_1 Q_1 = \begin{bmatrix} \mathbf{i} & 0 \\ 0 & \mathbf{-i} \end{bmatrix}.$$

La forma normale di Schur della matrice A risulta quindi

$$A = QBQ^H = Q \begin{bmatrix} I_2 & O \\ O & Q_1 \end{bmatrix} \begin{bmatrix} T_1 & CQ_1 \\ O & Q_1^H A_1 Q_1 \end{bmatrix} \begin{bmatrix} I_2 & O \\ O & Q_1 \end{bmatrix}^H Q^H$$

$$= U \begin{bmatrix} \mathbf{i} & 0 & (3+3\mathbf{i})/\sqrt{2} & (3-\mathbf{i})/\sqrt{2} \\ 0 & \text{-}\mathbf{i} & (3+\mathbf{i})/\sqrt{2} & (3-3\mathbf{i})/\sqrt{2} \\ 0 & 0 & \mathbf{i} & 0 \\ 0 & 0 & 0 & \text{-}\mathbf{i} \end{bmatrix} U^H,$$

dove U è la matrice unitaria

$$U = Q \begin{bmatrix} I_2 & O \\ O & Q_1 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 1 & 1 & (1+\mathbf{i})/\sqrt{2} & (1-\mathbf{i})/\sqrt{2} \\ 1 & 1 & (\text{-}1-\mathbf{i})/\sqrt{2} & (\text{-}1+\mathbf{i})/\sqrt{2} \\ \mathbf{i} & \text{-}\mathbf{i} & (1-\mathbf{i})/\sqrt{2} & (1+\mathbf{i})/\sqrt{2} \\ \text{-}\mathbf{i} & \mathbf{i} & (1-\mathbf{i})/\sqrt{2} & (1+\mathbf{i})/\sqrt{2} \end{bmatrix}.$$

∎

Come nel caso della forma canonica di Jordan, anche nel caso della forma normale di Schur, se la matrice A ha elementi reali esiste la *forma normale reale di Schur*.

2.24 Teorema. *Se $A \in \mathbf{R}^{n\times n}$, esiste una matrice ortogonale $U \in \mathbf{R}^{n\times n}$ e una matrice $T \in \mathbf{R}^{n\times n}$ triangolare superiore a blocchi, della forma*

$$T = \begin{bmatrix} R_{11} & R_{12} & \cdots & R_{1m} \\ & R_{22} & \cdots & R_{2m} \\ & & \ddots & \vdots \\ & & & R_{mm} \end{bmatrix},$$

dove i blocchi R_{jj} per $j = 1, 2, \ldots, m$ hanno ordine 1 o 2. Se λ_j è autovalore reale di A, allora R_{jj} ha ordine 1 e coincide con $[\lambda_j]$, se λ_j è complesso, allora il blocco R_{jj} ha ordine 2 ed ha come autovalori λ_j e $\overline{\lambda}_j$. La somma delle dimensioni dei blocchi R_{jj}, $j = 1, 2, \ldots, m$ è pari ad n.

Dim. La dimostrazione viene fatta per induzione, in modo analogo a quella del teorema 2.22. Se l'autovalore λ_1 è reale, si ripete il ragionamento fatto per il caso complesso. Se invece $\lambda_1 = \mu_1 + \mathbf{i}\nu_1$, $\mu_1, \nu_1 \in \mathbf{R}$, $\nu_1 \neq 0$, si considera il corrispondente autovettore $\mathbf{x}_1 + \mathbf{i}\mathbf{y}_1$, $\mathbf{x}_1, \mathbf{y}_1 \in \mathbf{R}^n$, in cui si può supporre che il vettore $\mathbf{x}_1$ sia normalizzato. Poiché

$$A(\mathbf{x}_1 + \mathbf{i}\mathbf{y}_1) = A\mathbf{x}_1 + \mathbf{i}A\mathbf{y}_1 = (\mu_1\mathbf{x}_1 - \nu_1\mathbf{y}_1) + \mathbf{i}(\mu_1\mathbf{y}_1 + \nu_1\mathbf{x}_1),$$

risulta

$$A[\,\mathbf{x}_1\,|\,\mathbf{y}_1\,] = [\,\mathbf{x}_1\,|\,\mathbf{y}_1\,] \begin{bmatrix} \mu_1 & \nu_1 \\ \text{-}\nu_1 & \mu_1 \end{bmatrix}. \tag{17}$$

I due vettori $\mathbf{x}_1$ e $\mathbf{y}_1$ sono linearmente indipendenti: infatti, se ciò non fosse, esisterebbe una costante $\alpha \neq 0$ per cui $\mathbf{y}_1 = \alpha\mathbf{x}_1$, e quindi

$$\mathbf{x}_1 + \mathbf{i}\mathbf{y}_1 = \mathbf{x}_1 + \mathbf{i}\alpha\mathbf{x}_1 = (1 + \mathbf{i}\alpha)\,\mathbf{x}_1$$

e il vettore reale $\mathbf{x}_1$ risulterebbe essere un autovettore reale di A corrispondente all'autovalore complesso λ_1, ciò che è assurdo perché A ha elementi reali.

Si costruisce un vettore normalizzato $\mathbf{z}_1$, ortogonale al vettore $\mathbf{x}_1$, ponendo

$$\mathbf{z}_1 = \beta\mathbf{x}_1 + \gamma\mathbf{y}_1, \quad \gamma = \frac{1}{\sqrt{\mathbf{y}_1^T\mathbf{y}_1 - (\mathbf{x}_1^T\mathbf{y}_1)^2}}, \quad \beta = \text{-}\gamma(\mathbf{x}_1^T\mathbf{y}_1).$$

Si ha perciò

$$[\,\mathbf{x}_1\,|\,\mathbf{z}_1\,] = [\,\mathbf{x}_1\,|\,\mathbf{y}_1\,]\,W, \quad \text{dove } W = \begin{bmatrix} 1 & \beta \\ 0 & \gamma \end{bmatrix}. \tag{18}$$

Si costruisce poi una matrice $Q \in \mathbf{R}^{n\times n}$ ortogonale, le cui prime due colonne siano $\mathbf{x}_1$ e $\mathbf{z}_1$:

$$Q = [\,\mathbf{x}_1\,|\,\mathbf{z}_1\,|\,\mathbf{y}_3\,|\,\ldots\,|\,\mathbf{y}_n\,].$$

Proseguendo la dimostrazione in modo analogo a quanto fatto nel teorema 2.22, per le prime due colonne della matrice $B = Q^TAQ$ si ha dalle (17) e (18):

$$B\,[\,\mathbf{e}_1\,|\,\mathbf{e}_2\,] = Q^TA\,[\,\mathbf{x}_1\,|\,\mathbf{z}_1\,] = Q^TA\,[\,\mathbf{x}_1\,|\,\mathbf{y}_1\,]W$$

$$= Q^T[\,\mathbf{x}_1\,|\,\mathbf{y}_1\,] \begin{bmatrix} \mu_1 & \nu_1 \\ \text{-}\nu_1 & \mu_1 \end{bmatrix} W = Q^T[\,\mathbf{x}_1\,|\,\mathbf{z}_1\,]\,W^{-1} \begin{bmatrix} \mu_1 & \nu_1 \\ \text{-}\nu_1 & \mu_1 \end{bmatrix} W.$$

Poiché Q è ortogonale, le prime due colonne di B possono essere scritte nel modo seguente

$$B\,[\,\mathbf{e}_1\,|\,\mathbf{e}_2\,] = \begin{bmatrix} I_2 \\ \\ O \end{bmatrix} W^{-1} \begin{bmatrix} \mu_1 & \nu_1 \\ \text{-}\nu_1 & \mu_1 \end{bmatrix} W = \begin{bmatrix} R_{11} \\ \\ O \end{bmatrix} \begin{matrix} \} & 2 \text{ righe} \\ \\ \} & n-2 \text{ righe,} \end{matrix}$$

in cui il blocco

$$R_{11} = W^{-1} \begin{bmatrix} \mu_1 & \nu_1 \\ \text{-}\nu_1 & \mu_1 \end{bmatrix} W$$

è reale e ha come autovalori λ_1 e $\overline{\lambda}_1$. La dimostrazione prosegue sfruttando l'ipotesi induttiva come nel teorema 2.22. ∎

2.25 Esempio. Si determina la forma normale reale di Schur della matrice

$$A = \frac{1}{2}\begin{bmatrix} 5 & -5 & 1 & -1 \\ 5 & -5 & 3 & 1 \\ -1 & -1 & -1 & -1 \\ 3 & -1 & 1 & 1 \end{bmatrix}$$

dell'esempio 2.23. La matrice A ha l'autovalore $\lambda_1 = \mathbf{i}$ con il corrispondente autovettore

$$\mathbf{x}_1 + \mathbf{i}\mathbf{y}_1 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \\ \mathbf{i} \\ \mathbf{-i} \end{bmatrix} = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} + \mathbf{i}\frac{1}{\sqrt{2}}\begin{bmatrix} 0 \\ 0 \\ 1 \\ -1 \end{bmatrix}.$$

In questo caso i due vettori $\mathbf{x}_1$ e $\mathbf{y}_1$ sono ortonormali. Si considerano poi gli altri due vettori ortonormali

$$\mathbf{y}_3 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -1 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{y}_4 = \frac{1}{\sqrt{2}}\begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}.$$

La matrice

$$U = [\,\mathbf{x}_1 \,|\, \mathbf{y}_1 \,|\, \mathbf{y}_3 \,|\, \mathbf{y}_4\,]$$

risulta così ortogonale, ed è tale che

$$A = U\begin{bmatrix} 0 & 1 & 5 & 1 \\ -1 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{bmatrix}U^T.$$

Si è cosí determinata la forma normale reale di Schur di A. ∎

Un caso particolarmente importante è quello in cui le matrici sono hermitiane.

2.26 Teorema. *Sia A una matrice hermitiana di ordine n, cioè $A = A^H$ e siano $\lambda_1, \ldots, \lambda_n$ i suoi autovalori. Allora esiste una matrice unitaria U tale che*

$$A = U \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix} U^H,$$

cioè la matrice A è diagonalizzabile. Inoltre i $\lambda_i,\ i = 1, \ldots, n$ sono reali e le colonne di U costituiscono un insieme di autovettori ortonormali.

Dim. Per il teorema 2.22 si ha $T = U^H A U$, dove T è una matrice triangolare superiore e U è unitaria. Poiché $A = A^H$, si ha

$$T^H = (U^H A U)^H = U^H A^H U = U^H A U = T,$$

cioè la matrice triangolare T risulta essere una matrice diagonale con gli elementi pricipali reali e per il teorema 2.16 le colonne di U, che sono ortonormali perché U è unitaria, risultano essere gli autovettori di A. ∎

Se la matrice A è reale e simmetrica, la matrice U risulta reale, e quindi è ortogonale.

2.27 Esempio. La matrice

$$A = \begin{bmatrix} 1 & \mathbf{i} & 0 \\ \text{-}\mathbf{i} & 2 & \text{-}\mathbf{i} \\ 0 & \mathbf{i} & 1 \end{bmatrix}$$

ha autovalori $\lambda_1 = 0$, $\lambda_2 = 1$ e $\lambda_3 = 3$, con i corrispondenti autovettori

$$\mathbf{x}_1 = \alpha_1 \begin{bmatrix} 1 \\ \mathbf{i} \\ 1 \end{bmatrix}, \quad \mathbf{x}_2 = \alpha_2 \begin{bmatrix} 1 \\ 0 \\ \text{-}1 \end{bmatrix}, \quad \mathbf{x}_3 = \alpha_3 \begin{bmatrix} 1 \\ \text{-}2\mathbf{i} \\ 1 \end{bmatrix}, \quad \alpha_1,\ \alpha_2,\ \alpha_3 \neq 0,$$

che sono fra loro ortogonali e che risultano normalizzati ponendo $\alpha_1 = 1/\sqrt{3}$, $\alpha_2 = 1/\sqrt{2}$ e $\alpha_3 = 1/\sqrt{6}$. Per cui, in questo caso, la matrice

$$U = [\, \mathbf{x}_1 \,|\, \mathbf{x}_2 \,|\, \mathbf{x}_3 \,],$$

data da

$$U = \frac{1}{\sqrt{6}} \begin{bmatrix} \sqrt{2} & \sqrt{3} & 1 \\ \mathbf{i}\sqrt{2} & 0 & \text{-}2\mathbf{i} \\ \sqrt{2} & \text{-}\sqrt{3} & 1 \end{bmatrix},$$

è unitaria e si ha

$$A = U \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{bmatrix} U^H.$$ ∎

Una classe più ampia di matrici che comprende, come casi particolari, le matrici hermitiane e le matrici unitarie, è quella delle matrici normali, cioè tali che $A^H A = A A^H$. Questa classe di matrici è particolarmente importante perché è quella che comprende tutte e sole le matrici diagonalizzabili con trasformazioni per similitudine unitarie. Vale infatti il

2.28 Teorema. *Una matrice $A \in \mathbf{C}^{n\times n}$ è normale, cioè $A^HA = AA^H$, se e solo se esiste una matrice unitaria U tale che*

$$A = U \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix} U^H,$$

in cui $\lambda_1, \ldots, \lambda_n$ sono gli autovalori di A. La matrice U ha per colonne gli autovettori della matrice A, che quindi sono a due a due ortonormali.

Dim. Si supponga dapprima che A sia normale. Per il teorema 2.22 esiste una matrice U unitaria tale che

$$T = U^HAU,$$

con T matrice triangolare superiore e si ha:

$$\begin{aligned} T^HT &= U^HA^HUU^HAU = U^HA^HAU, \\ TT^H &= U^HAUU^HA^HU = U^HAA^HU. \end{aligned}$$

Poiché A è normale, ne segue che

$$T^HT = TT^H, \tag{19}$$

e quindi anche T è normale. Si dimostra per induzione su n che T è diagonale. Se $n = 1$ questo è ovvio. Se $n > 1$, poiché T è triangolare superiore, per l'elemento p_{11} della matrice $P = T^HT = TT^H$, si ha

$$p_{11} = \bar{t}_{11}t_{11} = |\lambda_1|^2 \quad \text{e} \quad p_{11} = \sum_{j=1}^{n} t_{1j}\bar{t}_{1j} = |\lambda_1|^2 + \sum_{j=2}^{n} |t_{1j}|^2,$$

da cui

$$t_{1j} = 0, \quad \text{per} \quad j = 2, \ldots, n,$$

cioè la prima riga di T ha tutti gli elementi nulli eccetto quello principale. Indicata con T_{n-1} la sottomatrice ottenuta da T cancellando la prima riga e la prima colonna, dalla (19) segue che

$$T_{n-1}^HT_{n-1} = T_{n-1}T_{n-1}^H.$$

Per l'ipotesi induttiva T_{n-1} è diagonale, quindi anche T risulta diagonale.

Viceversa, sia A diagonalizzabile con una trasformazione per similitudine unitaria:

$$A = UDU^H,$$

con D diagonale. Si ha:

$$A^H A = UD^H U^H UDU^H = UD^H DU^H,$$

$$AA^H = UDU^H UD^H U^H = UDD^H U^H.$$

Poiché D è diagonale, $D^H D$ e DD^H sono diagonali, con elementi principali uguali a $\overline{\lambda}_i \lambda_i$; risulta allora $D^H D = DD^H$ e quindi

$$A^H A = AA^H. \qquad \blacksquare$$

Nel caso in cui la matrice A è normale e ha elementi reali, la sua forma normale reale di Schur risulta

$$A = UTU^T,$$

dove T e U sono matrici reali, U è ortogonale e T è diagonale a blocchi di ordine 1 o 2.

2.29 Esempio. La matrice $A \in \mathbf{R}^{4\times 4}$

$$A = \begin{bmatrix} 4 & -5 & 0 & 3 \\ 0 & 4 & -3 & -5 \\ 5 & -3 & 4 & 0 \\ 3 & 0 & 5 & 4 \end{bmatrix}$$

è normale, perché risulta $A^T A = AA^T$, e quindi è diagonalizzabile con trasformazioni per similitudine unitarie. Posto

$$U = \frac{1}{2} \begin{bmatrix} 1 & 1 & 1 & 1 \\ -1 & -\mathbf{i} & \mathbf{i} & 1 \\ 1 & -\mathbf{i} & \mathbf{i} & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix},$$

risulta

$$A = U \begin{bmatrix} 12 & & & \\ & 1+5\mathbf{i} & & \\ & & 1-5\mathbf{i} & \\ & & & 2 \end{bmatrix} U^H.$$

A può essere rappresentata anche nella forma normale reale di Schur. Poiché

$$\begin{bmatrix} 1+5\mathbf{i} & 0 \\ 0 & 1-5\mathbf{i} \end{bmatrix} = V \begin{bmatrix} 1 & -5 \\ 5 & 1 \end{bmatrix} V^H,$$

dove V è la matrice unitaria

$$V = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -\mathbf{i} \\ 1 & \mathbf{i} \end{bmatrix},$$

si ha

$$A = Z \begin{bmatrix} 12 & 0 & 0 & 0 \\ 0 & 1 & -5 & 0 \\ 0 & 5 & 1 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix} Z^T,$$

dove la matrice ortogonale Z è data da

$$Z = U \begin{bmatrix} 1 & \mathbf{0}^H & 0 \\ \mathbf{0} & V & \mathbf{0} \\ 0 & \mathbf{0}^H & 1 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 1 & \sqrt{2} & 0 & 1 \\ -1 & 0 & -\sqrt{2} & 1 \\ 1 & 0 & -\sqrt{2} & -1 \\ 1 & -\sqrt{2} & 0 & 1 \end{bmatrix}.$$

∎

Per il teorema di Schur è possibile caratterizzare completamente gli autovalori dei polinomi di matrici. È infatti immediato dimostrare il seguente teorema.

2.30 Teorema. *Sia $A = UTU^H$ la forma normale di Schur della matrice A. Se $p(x)$ è un polinomio in x, allora $p(A) = Up(T)U^H$ e gli autovalori di $p(A)$ sono tutti e soli i numeri $p(\lambda)$, dove λ è autovalore di A.* ∎

Siano $p(x)$ e $q(x)$ due polinomi nella variabile x, tali che $q(\lambda) \neq 0$ per ogni autovalore λ di A, e si consideri la funzione razionale $f(x) = p(x)/q(x)$. Per il teorema 2.30, la matrice $q(A)$ risulta non singolare ed è quindi possibile definire

$$f(A) = [q(A)]^{-1} p(A).$$

Per la matrice $f(A)$ vale un risultato analogo a quello del teorema 2.30:

$$f(A) = U f(T) U^H. \tag{20}$$

6. Alcune proprietà delle matrici definite positive

2.31 Teorema. *Sia A una matrice hermitiana di ordine n e siano $\lambda_1, \ldots, \lambda_n$ i suoi autovalori. Allora A è definita positiva se e solo se $\lambda_i > 0,\ i = 1, \ldots, n$.*

Dim. Si dimostra prima che se A è definita positiva, allora i suoi autovalori sono positivi. Poiché A è hermitiana, i suoi autovalori sono tutti reali. Se λ è un autovalore e $\mathbf{x} \neq \mathbf{0}$ è un autovettore corrispondente, da $A\mathbf{x} = \lambda\mathbf{x}$, premoltiplicando per $\mathbf{x}^H$, si ottiene

$$\mathbf{x}^H A\mathbf{x} = \lambda \mathbf{x}^H \mathbf{x}.$$

Poiché A è definita positiva, il primo membro è positivo, ed essendo $\mathbf{x}^H\mathbf{x} > 0$, risulta $\lambda > 0$.

Viceversa, poiché A è hermitiana, risulta $A = UDU^H$, con U matrice unitaria e D matrice diagonale avente come elementi principali gli autovalori $\lambda_i,\ i = 1, \ldots n$, di A. Se $\mathbf{x} \in \mathbf{C}^n, \mathbf{x} \neq \mathbf{0}$, si ha:

$$\mathbf{x}^H A\mathbf{x} = \mathbf{x}^H UDU^H\mathbf{x} = \mathbf{y}^H D\mathbf{y}, \tag{21}$$

dove il vettore $\mathbf{y} = U^H\mathbf{x}$ non può essere uguale a $\mathbf{0}$, perché U è non singolare. Dalla (21) si ha:

$$\mathbf{x}^H A\mathbf{x} = (\overline{y}_1, \ldots, \overline{y}_n) \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$
$$= \lambda_1\overline{y}_1 y_1 + \ldots + \lambda_n\overline{y}_n y_n = \lambda_1|y_1|^2 + \cdots + \lambda_n|y_n|^2 > 0$$

poiché gli autovalori λ_i sono tutti positivi e gli $|y_i|$ non sono tutti nulli. ∎

Poiché il prodotto degli autovalori di una matrice è uguale al determinante, dal teorema 2.31 segue che il determinante di una matrice definita positiva è positivo. Inoltre dal teorema 2.31 segue che l'inversa di una matrice definita positiva è ancora definita positiva. Infatti l'inversa A^{-1} di una matrice hermitiana A è hermitiana, e gli autovalori di A^{-1} sono positivi in quanto reciproci di quelli di A.

2.32 Esempio. La matrice hermitiana

$$A = \begin{bmatrix} 1 & \mathbf{i} & 0 \\ \mathbf{-i} & 2 & \mathbf{-2i} \\ 0 & \mathbf{2i} & 5 \end{bmatrix}$$

è definita positiva, infatti per ogni vettore $\mathbf{x} \neq \mathbf{0}$ si ha

$$\mathbf{x}^H A\mathbf{x} = |x_1 + \mathbf{i}x_2|^2 + |x_2 - 2\mathbf{i}x_3|^2 + |x_3|^2 > 0,$$

e il suo polinomio caratteristico è dato da

$$P(\lambda) = -\lambda^3 + 8\lambda^2 - 12\lambda + 1. \tag{22}$$

Esaminando il grafico di $P(\lambda)$ riportato nella figura 2.1 e calcolando i valori che questo polinomio assume nei punti 0, 1, 2, 7:

$$P(0) = 1,\ P(1) = -4,\ P(2) = 1,\ P(7) = -34,$$

risulta che il polinomio ha 3 zeri reali compresi negli intervalli

$$(0,1),\ (1,2) \text{ e } (2,7),$$

e quindi gli autovalori di A sono tutti positivi. ∎

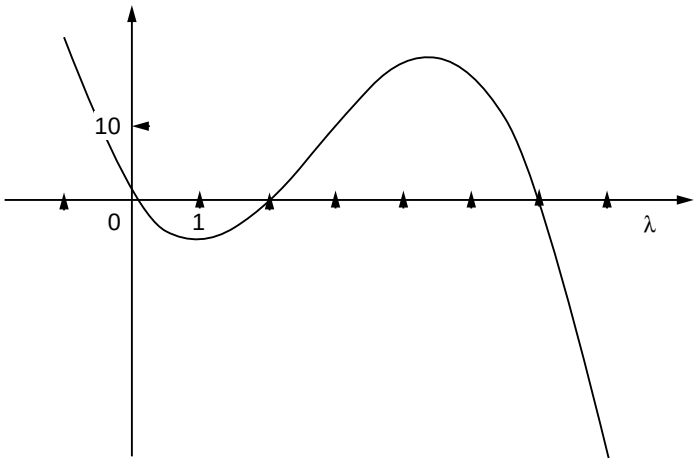


Fig. 2.1 - Grafico del polinomio(22).

2.33 Teorema. *Una matrice hermitiana A è definita positiva se e solo se i determinanti di tutte le sottomatrici principali di testa di A (e quindi anche il determinante di A) sono positivi.*

Dim. Se A è definita positiva, allora la tesi segue direttamente dal teorema 1.14. Viceversa, si suppone che i determinanti di tutte le sottomatrici principali di testa di A siano positivi e si procede per induzione su n. Per $n = 1$ il risultato è banale. Per $n > 1$, siano $\lambda_1, \ldots, \lambda_n$ gli autovalori di A. Poiché per ipotesi il prodotto dei λ_i è positivo, si dimostra che non può esistere

un numero pari di autovalori negativi e quindi tutti i λ_i sono positivi. Si supponga, per assurdo, che esistano m autovalori negativi, con $m \geq 2$, m numero pari (si può supporre, senza violare la generalità, che tali autovalori siano i primi m). Sia U la matrice unitaria tale che $A = UDU^H$, in cui D è la matrice diagonale i cui elementi principali sono i λ_i, ordinati come indicato. Allora è possibile costruire due vettori $\mathbf{x}, \mathbf{y} \in \mathbf{C}^n$, tali che

$$\mathbf{x}, \mathbf{y} \neq \mathbf{0},\ x_n = 0,\ y_{m+1} = y_{m+2} = \ldots = y_n = 0,\ \mathbf{y} = U^H\mathbf{x}.$$

Infatti, partizionando la matrice U^H e i vettori $\mathbf{x}, \mathbf{y}$ nel modo seguente:

$$U^H = \begin{bmatrix} V & \mathbf{v} \\ W & \mathbf{w} \end{bmatrix} \begin{matrix} \} & m \text{ righe} \\ \} & n-m \text{ righe} \end{matrix}$$

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ 0 \end{bmatrix} \begin{matrix} \} & n-1 \text{ componenti} \\ \} & 1 \text{ componente} \end{matrix} \qquad \mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{0} \end{bmatrix} \begin{matrix} \} & m \text{ componenti} \\ \} & n-m \text{ componenti} \end{matrix}$$

dalla condizione $\mathbf{y} = U^H\mathbf{x}$ si ottiene

$$V\mathbf{x}_1 = \mathbf{y}_1$$

$$W\mathbf{x}_1 = \mathbf{0}.$$

Poiché nella matrice W il numero $(n-1)$ di colonne è maggiore del numero di righe $(n-m, m \geq 2)$, è sempre possibile determinare una combinazione lineare nulla a coefficienti non tutti nulli delle colonne di W. Esiste allora un vettore $\mathbf{x}_1 \neq \mathbf{0}$ formato da tali coefficienti, tale che $W\mathbf{x}_1 = \mathbf{0}$ e $\mathbf{y}_1 = V\mathbf{x}_1 \neq \mathbf{0}$, perché altrimenti le prime $n-1$ colonne di U^H risulterebbero linearmente dipendenti. Ne segue la relazione

$$\mathbf{x}^H A\mathbf{x} = \mathbf{x}^H UDU^H\mathbf{x} = \mathbf{y}^H D\mathbf{y} = \sum_{i=1}^{n} \lambda_i |y_i|^2 = \sum_{i=1}^{m} \lambda_i |y_i|^2 < 0,$$

che è assurda perché

$$\mathbf{x}^H A\mathbf{x} = \mathbf{x}_1^H A_{n-1}\mathbf{x}_1,$$

dove $A_{n-1} \in \mathbf{C}^{(n-1)\times(n-1)}$ è la sottomatrice principale di testa di ordine $n-1$, che è definita positiva per l'ipotesi induttiva. ∎

7. Localizzazione degli autovalori

In questo paragrafo vengono dati tre teoremi che consentono di individuare zone del piano complesso in cui si trovano gli autovalori di una matrice.

2.34 Definizione. Sia $A \in \mathbf{C}^{n\times n}$. I cerchi del piano complesso

$$K_i = \{\, z \in \mathbf{C} \,:\, |z - a_{ii}| \le \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| \,\},\ i = 1, 2, \ldots, n,$$

di centro a_{ii} e raggio $r_i = \sum\limits_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|$ sono detti *cerchi di Gerschgorin*. Vale il seguente

2.35 Teorema (primo teorema di Gerschgorin). *Gli autovalori della matrice A di ordine n sono tutti contenuti in*

$$\bigcup_{i=1,\ldots,n} K_i.$$

Dim. Sia λ un autovalore di A e $\mathbf{x}$ un autovettore corrispondente, ossia

$$A\mathbf{x} = \lambda\,\mathbf{x},\ \mathbf{x} \neq \mathbf{0}.$$

Allora si ha:

$$\sum_{j=1}^{n} a_{ij}\, x_j = \lambda\, x_i,\ i = 1, \ldots, n,$$

da cui

$$(\lambda - a_{ii})x_i = \sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij}\, x_j,\ i = 1, \ldots, n. \tag{23}$$

Sia x_p la componente di $\mathbf{x}$ di massimo modulo, cioè quella per cui

$$|x_p| = \max_{j=1,\ldots,n} |x_j| \neq 0, \tag{24}$$

e, ponendo $i = p$ nella (23), si ha:

$$(\lambda - a_{pp})x_p = \sum_{\substack{j=1 \\ j\neq p}}^{n} a_{pj}\, x_j;$$

da cui:

$$|\lambda - a_{pp}|\,|x_p| \leq \sum_{\substack{j=1 \\ j\neq p}}^{n} |a_{pj}|\,|x_j|, \tag{25}$$

e, per la (24),

$$|\lambda - a_{pp}|\,|x_p| \leq \sum_{\substack{j=1 \\ j\neq p}}^{n} |a_{pj}|\,|x_p|;$$

infine dividendo per $|x_p| > 0$, si ottiene

$$|\lambda - a_{pp}| \leq \sum_{\substack{j=1 \\ j\neq p}}^{n} |a_{pj}|, \tag{26}$$

e quindi $\lambda \in K_p$. Si osservi che, poiché a priori non è noto il valore dell'indice p, è possibile solo dire che λ appartiene all'unione di tutti i cerchi K_i. ∎

Poiché il teorema precedente può essere applicato anche alla matrice A^T, che ha gli stessi autovalori della matrice A, risulta che gli autovalori di A appartengono anche all'unione dei cerchi

$$H_i = \{\, z \in \mathbf{C} \;:\; |z - a_{ii}| \leq \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ji}| \,\},\ i = 1, 2, \ldots, n,$$

e quindi gli autovalori di A appartengono all'insieme

$$\Big(\bigcup_{i=1,\ldots,n} K_i\Big) \bigcap \Big(\bigcup_{i=1,\ldots,n} H_i\Big).$$

2.36 Esempio. Si consideri la matrice

$$A = \begin{bmatrix} 15 & -2 & 2 \\ 1 & 10 & -3 \\ -2 & 1 & 0 \end{bmatrix} \tag{27}$$

alla quale sono associati i cerchi

$$\begin{aligned} K_1 &= \{z \in \mathbf{C} \;:\; |z - 15| \leq 4\}, \\ K_2 &= \{z \in \mathbf{C} \;:\; |z - 10| \leq 4\}, \\ K_3 &= \{z \in \mathbf{C} \;:\; |z| \leq 3\}, \end{aligned}$$

rappresentati nella figura 2.2.

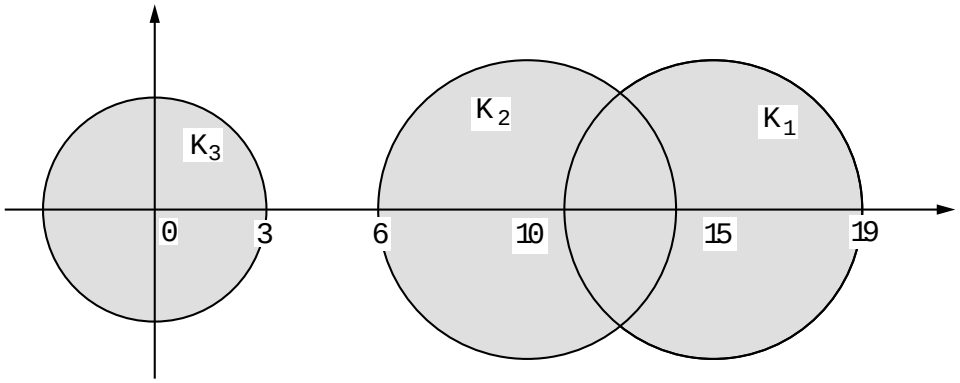


Fig. 2.2 - Cerchi di Gerschgorin associati alla matrice A in (27).

Quindi gli autovalori stanno nelle aree grigie. Si considerino poi i cerchi

$$H_1 = \{z \in \mathbf{C} \ : \ |z - 15| \leq 3\},$$
$$H_2 = \{z \in \mathbf{C} \ : \ |z - 10| \leq 3\},$$
$$H_3 = \{z \in \mathbf{C} \ : \ |z| \leq 5\},$$

associati alla matrice A^T e rappresentati nella figura 2.3.

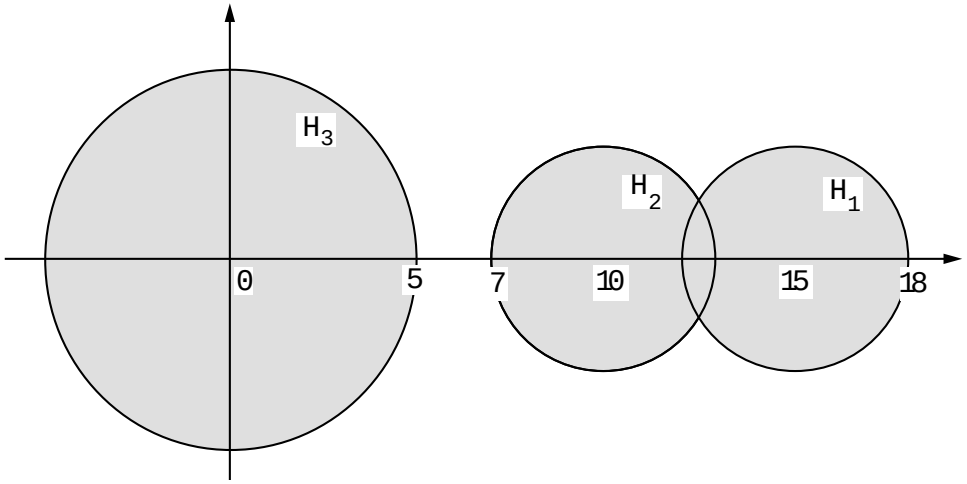


Fig. 2.3 - Cerchi di Gerschgorin associati alla matrice A^T in (27).

Quindi gli autovalori di A stanno nell'intersezione dei due insiemi di cerchi, rappresentata nella figura 2.4. ∎

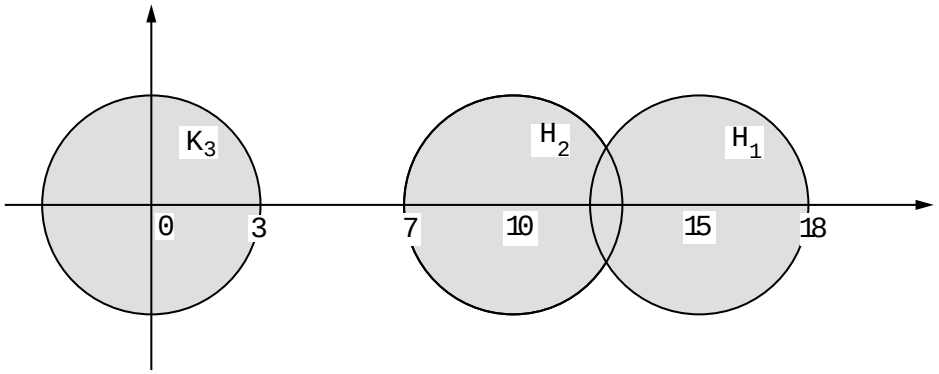


Fig. 2.4 - Intersezione dei cerchi di Gerschgorin delle figure 2.2 e 2.3.

2.37 Teorema (secondo teorema di Gerschgorin). *Se l'unione M_1 di k cerchi di Gerschgorin è disgiunta dall'unione M_2 dei rimanenti $n-k$, allora k autovalori appartengono a M_1 e $n-k$ autovalori appartengono a M_2.*

Dim. Si può supporre, senza ledere la generalità, che i cerchi che costituiscono M_1 siano i primi k, cioè che

$$M_1 = \bigcup_{i=1,\ldots,k} K_i \quad \text{e} \quad M_2 = \bigcup_{i=k+1,\ldots,n} K_i.$$

Siano D e R le matrici di elementi

$$d_{ij} = \begin{cases} a_{ij} & \text{se } i = j, \\ 0 & \text{se } i \neq j, \end{cases} \qquad r_{ij} = \begin{cases} 0 & \text{se } i = j, \\ a_{ij} & \text{se } i \neq j, \end{cases}$$

per cui $A = D + R$. La matrice

$$A(t) = D + tR, \quad t \in [0,1],$$

i cui elementi sono funzioni continue di t, ha autovalori che sono funzioni continue di t, in quanto zeri del polinomio caratteristico i cui coefficienti sono funzioni continue degli elementi di $A(t)$. Infatti gli zeri di un polinomio, sono funzioni continue dei coefficienti (si veda [5]). Per ogni $t \in [0,1]$ i primi k cerchi di Gerschgorin di $A(t)$ sono contenuti in M_1 perché hanno gli stessi centri dei cerchi K_i, $1,\ldots,k$, e raggio crescente con t, e analogamente i restanti $n-k$ cerchi di Gerschgorin di $A(t)$ sono contenuti in M_2. Poiché l'unione dei primi k cerchi di Gerschgorin di $A(t)$ è disgiunta dall'unione dei restanti cerchi di Gerschgorin, facendo variare con continuità t nell'intervallo [0,1], gli autovalori di $A(t)$ non possono passare da un insieme all'altro fra loro disgiunti. Per $t = 0$ in M_1 e M_2 stanno rispettivamente k e $n-k$

autovalori di $A(t)$, perché $A(0) = D$ e gli autovalori coincidono con i centri dei cerchi di Gerschgorin (che sono gli elementi principali di A). Quindi per ogni $t \in [0,1]$, e in particolare per $t = 1$ per cui $A(1) = A$, in M_1 stanno k autovalori e in M_2 stanno $n-k$ autovalori. ∎

Dei tre autovalori della matrice A dell'esempio 2.36, uno è contenuto in K_3, mentre gli altri due appartengono ad $H_1 \cup H_2$. I due autovalori contenuti in $H_1 \cup H_2$ possono essere reali o complessi e hanno modulo compreso fra 7 e 18. L'autovalore contenuto in K_3 è reale; infatti, se avesse parte immaginaria non nulla, anche il suo coniugato dovrebbe essere un autovalore di A, essendo zero di un polinomio a coefficienti reali.

Per le matrici irriducibili vi è poi un altro teorema che precisa ulteriormente la localizzazione degli autovalori.

2.38 Teorema (terzo teorema di Gerschgorin). *Se la matrice A di ordine n è irriducibile, ogni autovalore λ, che sta sulla frontiera dei cerchi di Gerschgorin a cui appartiene, sta sulla frontiera di tutti i cerchi di Gerschgorin. In particolare questo vale per gli autovalori che appartengono alla frontiera dell'unione dei cerchi di Gerschgorin.*

Dim. Sia $\mathbf{x}$ un autovettore corrispondente all'autovalore λ, e sia x_p la sua componente di massimo modulo

$$|x_p| = \max_{j=1,\ldots,n} |x_j|.$$

Procedendo come nella dimostrazione del teorema 2.35, si ha $\lambda \in K_p$. Poiché per ipotesi λ sta sulla frontiera di K_p, la (26) deve valere con il segno di uguaglianza:

$$|\lambda - a_{pp}| = \sum_{\substack{j=1 \\ j \neq p}}^{n} |a_{pj}|.$$

Ne segue che il segno di uguaglianza deve valere anche per la (25) e quindi $|x_j| = |x_p|$, per tutti gli indici j per cui $a_{pj} \neq 0$. Per l'ipotesi di irriducibilità, esiste però almeno un indice $r, r \neq p$, per cui $a_{pr} \neq 0$, e poiché

$$|x_r| = \max_{j=1,\ldots,n} |x_j|,$$

si può riapplicare per l'indice r il procedimento appena seguito per l'indice p. Così procedendo si arriva alla conclusione che λ sta sulla frontiera di K_r ed inoltre che $|x_j| = |x_r|$, per tutti gli indici j per cui $a_{rj} \neq 0$. Il procedimento si può ripetere poi per un altro indice $s, s \neq r$, per cui $a_{rs} \neq 0$, e così via per tutti i rimanenti indici, in quanto per la irriducibilità di A, esiste un cammino orientato che tocca tutti i nodi del grafo di A. ∎

2.39 Esempio. La matrice

$$F = \begin{bmatrix} 0 & \cdot & \cdots & 0 & -a_0 \\ 1 & 0 & \cdots & 0 & -a_1 \\ & \ddots & \ddots & \vdots & \vdots \\ & & \ddots & 0 & -a_{n-2} \\ & & & 1 & -a_{n-1} \end{bmatrix}$$

è detta *matrice di Frobenius*. Calcolando il $\det(F - \lambda I)$ con la regola di Laplace applicata all'ultima colonna, risulta che

$$\det(F - \lambda I) = (-1)^n \left(\lambda^n + \sum_{i=0}^{n-1} a_i \, \lambda^i \right).$$

Inoltre il polinomio minimo coincide, a meno del segno, con il polinomio caratteristico. Infatti se per assurdo fosse

$$\psi(\lambda) = \lambda^k + \alpha_0 \lambda^{k-1} + \ldots + \alpha_{k-1}, \quad \text{con } k < n,$$

allora, moltiplicando la matrice $\psi(F)$ per il primo vettore della base canonica $\mathbf{e}_1$, risulterebbe

$$\begin{aligned} \psi(F)\mathbf{e}_1 &= F^k \mathbf{e}_1 + \alpha_0 F^{k-1} \mathbf{e}_1 + \ldots + \alpha_{k-1} \mathbf{e}_1 \\ &= \mathbf{e}_{k+1} + \alpha_0 \mathbf{e}_k + \ldots + \alpha_{k-1} \mathbf{e}_1 \end{aligned}$$

e quindi $\psi(F)\mathbf{e}_1$ sarebbe uguale al vettore le cui prime $k+1$ componenti sono nell'ordine

$$\alpha_{k-1}, \ldots, \alpha_0, 1,$$

e ciò è assurdo perché $\psi(F) = 0$.

Il teorema di Gerschgorin, applicato alle matrici F e F^T permette di dare al modulo degli zeri λ_i del polinomio

$$\lambda^n + \sum_{i=0}^{n-1} a_i \, \lambda^i$$

le seguenti limitazioni

$$\begin{aligned} &|\lambda_i| \leq \max\{ \, |a_0|, 1 + |a_1|, \ldots, 1 + |a_{n-1}| \, \} \\ &|\lambda_i| \leq \max\{ \, 1, \, \sum_{i=0}^{n-1} |a_i| \, \}. \end{aligned}$$

∎

8. Predominanza diagonale

Un'altra classe importante di matrici è quella delle matrici a predominanza diagonale, che si presentano spesso nella risoluzione numerica di problemi differenziali.

2.40 Definizioni. Una matrice $A \in \mathbf{C}^{n\times n}$ si dice *a predominanza diagonale* se per ogni $i = 1, \ldots, n$ risulta

$$|a_{ii}| \geq \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|$$

ed esiste almeno un indice s per cui

$$|a_{ss}| > \sum_{\substack{j=1 \\ j\neq s}}^{n} |a_{sj}|. \tag{28}$$

Una matrice $A \in \mathbf{C}^{n\times n}$ si dice *a predominanza diagonale in senso stretto* se per ogni $i = 1, \ldots, n$ risulta

$$|a_{ii}| > \sum_{\substack{j=1 \\ j\neq i}} |a_{ij}|.$$

Le due definizioni di predominanza diagonale e di predominanza diagonale in senso stretto si possono dare anche per colonne, considerando le somme per colonne anziché per righe e in tal caso si specifica che la *predominanza diagonale (predominanza diagonale in senso stretto) è per colonne.* ∎

2.41 Teorema. *Se $A \in \mathbf{C}^{n\times n}$ è una matrice a predominanza diagonale in senso stretto, oppure a predominanza diagonale e irriducibile, allora A è non singolare. Se inoltre A ha elementi principali tutti reali e positivi, allora gli autovalori di A hanno parte reale positiva e se A è anche hermitiana, allora A è definita positiva.*

Dim. Se A è a predominanza diagonale in senso stretto, dal teorema 2.35 risulta che i cerchi di Gerschgorin, avendo raggio minore della distanza del centro dall'origine del piano complesso, non possono includere l'origine, e quindi A non può avere un autovalore nullo.

Se A è a predominanza diagonale ed è irriducibile, è possibile che l'origine appartenga alla frontiera di un cerchio di Gerschgorin. Se però un autovalore di A fosse nullo, allora per il teorema 2.38 l'origine dovrebbe

appartenere alla frontiera di tutti i cerchi di Gerschgorin, ma ciò, per la (28), non può essere vero per l's-esimo cerchio.

Inoltre, se A ha predominanza diagonale in senso stretto o ha predominanza diagonale ed è irriducibile, e se gli elementi principali di A sono tutti positivi, nessun cerchio può contenere numeri complessi a parte reale negativa. Quindi se A è anche hermitiana, cioè con autovalori reali, questi devono essere positivi, e per il teorema 2.31 risulta che la matrice è definita positiva. ∎

Poiché gli autovalori della matrice A sono uguali a quelli della matrice A^T, le tesi del teorema 2.41 valgono anche nel caso in cui la predominanza diagonale (la predominanza diagonale in senso stretto) della matrice sia per colonne.

Esercizi proposti

2.1 Si calcolino gli autovalori e gli autovettori delle matrici

$$a)\quad \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}, \qquad b)\quad \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$

$$c)\quad \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}, \qquad d)\quad \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & -1 \\ 1 & 1 & 2 \end{bmatrix}.$$

2.2 Si determinino il polinomio caratteristico e il polinomio minimo delle seguenti matrici

$$A_1 = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 1 & 2 & 2 \\ 2 & 1 & 2 \\ 2 & 2 & 1 \end{bmatrix}, \quad A_3 = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 1 & 2 \\ 1 & 1 & 2 \end{bmatrix}.$$

2.3 Si dica quante e quali sono le matrici di ordine 6, non simili fra di loro, il cui polinomio caratteristico è dato da

$$P(\lambda) = (3-\lambda)^4(1-\lambda)^2,$$

e per ogni matrice si scriva il polinomio minimo.

2.4 Si dica se la matrice

$$A = \begin{bmatrix} 17 & 2 & 2 \\ 1 & 10 & 3 \\ 1 & 2 & 0 \end{bmatrix}$$

ha autovalori complessi.
(Traccia: si sfrutti il 2^o teorema di Gerschgorin.)

2.5 Si dica se la matrice

$$A = \begin{bmatrix} 3 & 0 & 1 & 0 \\ 0 & 11 & 1 & -1 \\ 1 & 1 & 10 & 2 \\ 0 & -1 & 2 & 20 \end{bmatrix}$$

ha un autovalore maggiore o uguale a 20.
(Traccia: si verifichi che $\det(A - 20I) < 0$.)

2.6 Si verifichi che il raggio spettrale della matrice

$$A = \begin{bmatrix} 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 \end{bmatrix}$$

è maggiore di 2.

2.7 Si determini la forma normale di Schur e la forma normale reale di Schur delle seguenti matrici:

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix}, \qquad B = \begin{bmatrix} 4 & 1 & -8 \\ 7 & 4 & 4 \\ 4 & -8 & 1 \end{bmatrix}, \qquad C = \begin{bmatrix} 1 & 8 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 \\ 0 & 4 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

2.8 È data la matrice

$$A = \begin{bmatrix} 1 & -\frac{1}{2} & -\frac{1}{2}\mathbf{i} & 0 \\ -\frac{1}{2} & \frac{3}{2}\mathbf{i} & \mathbf{i} & 0 \\ 0 & -\frac{1}{2}\mathbf{i} & 5+\mathbf{i} & \frac{1}{2}\mathbf{i} \\ -\mathbf{i} & 0 & 0 & 4\mathbf{i} \end{bmatrix}.$$

Si dica

a) se la matrice A è singolare;

b) se esiste un autovalore uguale a $\rho(A)$;

c) se esiste un solo autovalore di modulo massimo.

(Traccia: a) e b) si disegnino i cerchi di Gerschgorin e si noti che la matrice è irriducibile; c) si consideri la matrice $B = S^{-1}AS$, con

$$S = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 3 & \\ & & & 1 \end{bmatrix}.)$$

2.9 Sia α un numero reale tale che $0 < \alpha < 1$ e sia

$$A = \begin{bmatrix} 1 & \alpha & & & & & & \alpha \\ \alpha & 1 & 2\alpha & & & & & \\ & 2\alpha & 1 & & & & & \\ & & & 2 & \alpha & & & \\ & & & \alpha & 2 & & & \\ & & & & & -0.5 & 0.1 & -0.2 \\ & & & & & 0.1 & -1 & 0 \\ \alpha & & & & & -0.2 & 0 & 2 \end{bmatrix}.$$

Si verifichi che vi sono 3 autovalori di A nell'intervallo $[1 - 3\alpha, 1 + 3\alpha]$ se $\alpha < 0.25$.

(Traccia: come per il punto c) dell'esercizio 2.8, si consideri un'opportuna matrice S.)

2.10 Sia $A \in \mathbf{R}^{n\times n}$ la matrice i cui elementi sono dati da

$$a_{ij} = \begin{cases} i & \text{per } i = j, \\ 1 & \text{per } i < j, \\ 0 & \text{per } i > j. \end{cases}$$

Si determinino gli autovalori e gli autovettori di A e A^T.

(Traccia: tenendo presente l'esercizio 1.52, si dimostri che $A = XDX^{-1}$ e $A^T = X^{-T}DX^T$, dove

$$X = \begin{bmatrix} 1 & 1 & \dots & 1 \\ & 1 & \dots & 1 \\ & & \ddots & \vdots \\ & & & 1 \end{bmatrix}, \qquad D = \begin{bmatrix} 1 & & & \\ & 2 & & \\ & & \ddots & \\ & & & n \end{bmatrix}.)$$

2.11 Sia $F \in \mathbf{C}^{n\times n}$ una matrice di Frobenius (per la definizione e le proprietà si veda l'esempio 2.39). Si dimostri che se F ha autovalori λ_i,

$i = 1, \ldots, n$, distinti, allora

$$\text{a)} \qquad F\mathbf{x}_i = \lambda_i \mathbf{x}_i, \qquad \text{dove} \qquad \mathbf{x}_i = \prod_{\substack{j=1 \\ j \neq i}}^{n} (F - \lambda_j I)\mathbf{e}_1 \neq \mathbf{0},$$

$$\text{b)} \qquad F^T \mathbf{y}_i = \lambda_i \mathbf{y}_i, \qquad \text{dove} \qquad \mathbf{y}_i = [1 \ , \ \lambda_i \ , \ \lambda_i^2 \ , \ldots , \ \lambda_i^{n-1}]^T.$$

Si esamini in particolare la matrice

$$F = \begin{bmatrix} \mathbf{0} & \text{-}1 \\ \\ I_4 & \text{-}\mathbf{u}^T \end{bmatrix}, \qquad \mathbf{u} \in \mathbf{R}^4,$$

dove $\mathbf{u} = [1 \ , \ 2 \ , \ 2 \ , \ 1]^T$.

(Traccia: a) basta verificare che $(F - \lambda_i I)\mathbf{x}_i = \mathbf{0}$, per questo si osservi che

$$(F - \lambda_i I) \prod_{j=1}^{n} (F - \lambda_j I) = \psi(F),$$

dove $\psi(\lambda)$ è il polinomio minimo di F; si verifichi inoltre che l'ultima componente di $\mathbf{x}_i$ è uguale a 1 per $i = 1, \ldots, n$; b) si consideri il prodotto $F^T \mathbf{y}_i$.)

2.12 Sia $A \in \mathbf{C}^{n \times n}$.

a) Si dimostri che A e A^T hanno gli stessi autovalori;

b) si dica se le matrici A e A^T sono simili;

c) si costruisca una matrice $A \in \mathbf{C}^{2 \times 2}$ tale che gli autovettori di A e A^T non siano gli stessi;

d) si dimostri che se $\mathbf{x}$ è autovettore di A corrispondente all'autovalore λ_i e $\mathbf{y}$ è autovettore di A^T corrispondente all'autovalore $\lambda_j \neq \lambda_i$, allora $\mathbf{y}^T \mathbf{x} = 0$ e quindi se $\mathbf{x}$ e $\mathbf{y}$ sono reali, allora sono anche ortogonali.

(Traccia: d) dalle relazioni $A\mathbf{x} = \lambda_i \mathbf{x}$ e $A^T \mathbf{y} = \lambda_j \mathbf{y}$ segue che $\lambda_i \mathbf{y}^T \mathbf{x} = \mathbf{y}^T A \mathbf{x} = \lambda_j \mathbf{y}^T \mathbf{x}$.)

2.13 Siano $A, B \in \mathbf{C}^{n \times n}$. Si dimostri che le matrici AB e BA hanno gli stessi autovalori. Se A e B sono singolari, è possibile che AB e BA, pur avendo gli stessi autovalori, non siano simili: si esamini il caso

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \qquad B = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}.$$

(Traccia: se una delle due matrici, ad esempio A, è non singolare, è $AB = A(BA)A^{-1}$, e quindi le matrici AB e BA sono simili. In generale, se $\lambda \neq 0$ è autovalore di AB, cioè $AB\mathbf{x} = \lambda\mathbf{x}$, $\mathbf{x} \neq \mathbf{0}$, allora $B\mathbf{x} \neq \mathbf{0}$ e $(BA)B\mathbf{x} = \lambda B\mathbf{x}$. Quindi λ è autovalore di BA. Se AB è singolare, anche BA lo è.)

2.14 Si dimostri, con un controesempio, che non tutte le matrici i cui autovalori hanno modulo 1 sono unitarie.

2.15 Si dimostri che gli autovalori della matrice $A \in \mathbf{C}^{n\times n}$ i cui elementi sono dati da

$$a_{ij} = \begin{cases} \alpha_i & \text{se } i = 1,\ j = n \text{ e } j = i-1,\ \text{per } i = 2,\dots,n, \\ \\ 0 & \text{altrimenti,} \end{cases}$$

dove $\alpha_1\alpha_2\cdots\alpha_n = 1$, sono i numeri

$$\lambda_k = \cos\frac{2\pi k}{n} + \mathbf{i}\sin\frac{2\pi k}{n},\ k = 1, 2, \dots, n.$$

(Traccia: si calcoli il polinomio caratteristico.)

2.16 Sia A la matrice

$$A = \begin{bmatrix} 1 & \alpha \\ 0 & 1 \end{bmatrix},$$

dove $\alpha \in \mathbf{C}$ è non nullo. Si dimostri che A non può essere diagonalizzata.

2.17 Si dimostri che il polinomio minimo di una qualsiasi matrice di permutazione di ordine n divide il polinomio $\lambda^k - 1$, per un opportuno k intero. Si esamini in particolare il caso

$$\Pi = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}.$$

(Traccia: sia Π la matrice di permutazione, si dimostri che esiste k tale che $\Pi^k = I$. Nel caso particolare è $\Pi^6 = I$ e $\psi(\lambda) = (\lambda+1)(\lambda^3-1)$ divide $\lambda^6 - 1$.)

2.18 Siano A e $B \in \mathbf{C}^{n\times n}$ due matrici che commutano. Si dimostri che

a) A e B hanno almeno un autovettore in comune;

b) se A e B sono diagonalizzabili, allora A e B hanno in comune n autovettori linearmente indipendenti; si esamini, come controesempio il caso

$$A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \qquad B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix};$$

c) se A e B sono diagonalizzabili e B ha autovalori distinti, allora A appartiene allo spazio vettoriale generato da $I, B, \ldots, B^{n-1}$; si esamini in particolare il caso

$$B = \begin{bmatrix} 0 & 1 & & \\ 1 & 0 & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & 0 \end{bmatrix};$$

d) se A ha autovalori distinti e $A = UTU^H$, $U \in \mathbf{C}^{n\times n}$ unitaria e $T \in \mathbf{C}^{n\times n}$ triangolare superiore, allora la matrice $U^H BU$ è triangolare superiore.

(Traccia: a) sia $A\mathbf{x} = \lambda\mathbf{x},\ \mathbf{x} \neq \mathbf{0}$; se $B\mathbf{x} = \mathbf{0}$, allora $\mathbf{x}$ è autovettore anche di B; altrimenti è $AB^k\mathbf{x} = B^kA\mathbf{x} = \lambda B^k\mathbf{x}$ per ogni intero $k \geq 1$ e quindi tutti i vettori $B^k\mathbf{x}$ sono autovettori di A corrispondenti all'autovalore λ. Si consideri il sottospazio S di $\mathbf{C}^n$ generato dai vettori $B^k\mathbf{x},\ k = 0, 1, \ldots$ Poiché se $\mathbf{z} \in S$ allora $B\mathbf{z} \in S$, esiste una costante $\mu \in \mathbf{C}$ e un vettore $\mathbf{y} \in S,\ \mathbf{y} \neq \mathbf{0}$, tale che $B\mathbf{y} = \mu\mathbf{y}$. b) siano $\lambda_1, \ldots, \lambda_m$ gli autovalori distinti di A, di molteplicità geometrica $\tau_1, \ldots, \tau_m$. Allora

$$A = T \begin{bmatrix} D_{11} & & & \\ & D_{22} & & \\ & & \ddots & \\ & & & D_{mm} \end{bmatrix} T^{-1},$$

dove $D_{ii} = \lambda_i I_{\tau_i}$ per l'ipotesi di diagonalizzabilità. Per $i = 1, \ldots, m$ si consideri il sottospazio S_i di $\mathbf{C}^n$ generato dai τ_i autovettori linearmente indipendenti $\mathbf{t}_1^{(i)}, \ldots, \mathbf{t}_{\tau_i}^{(i)}$ corrispondenti a λ_i che sono colonne di T. Per quanto visto al punto a) i vettori $B\mathbf{t}_j^{(i)}$ appartengono ad S_i per $j = 1, \ldots, \tau_i$, cioè

$$B\mathbf{t}_j^{(i)} = \sum_{r=1}^{\tau_i} \alpha_{jr}\mathbf{t}_r^{(i)}.$$

Queste relazioni possono essere scritte nella forma $B = TPT^{-1}$, dove P è diagonale a blocchi, con i blocchi principali P_i di ordine τ_i. Poiché B è diagonalizzabile, anche P lo è e quindi per ogni blocco P_i si ha che $P_i =$

$V_i Z_i V_i^{-1}$, dove Z_i è diagonale. c) poiché A e B hanno una base comune di autovettori basta dimostrare che esistono n costanti $\alpha_1, \alpha_2, \ldots, \alpha_n$ tali che

$$\sum_{i=1}^{n} \alpha_i \mu_j^{i-1} = \lambda_j,$$

dove $\mu_1, \ldots, \mu_n$ sono gli autovalori di B, a due a due distinti (si veda per questo l'esercizio 1.54). Nel caso particolare la matrice A commuta con B se

$$a_{i-1,j} + a_{i+1,j} = a_{i,j-1} + a_{i,j+1},$$

(avendo posto $a_{rs} = 0$ se $r < 0,\ s < 0,\ r > n$ oppure $s > n$). d) se $A\mathbf{x} = \lambda\mathbf{x}$, $\mathbf{x}^H\mathbf{x} = 1$, poiché per la commutatività $B\mathbf{x}$, se non nullo, è autovettore di A corrispondente allo stesso autovalore λ, che è distinto dagli altri autovalori, ne segue che $B\mathbf{x}$ è proporzionale a $\mathbf{x}$ e quindi $B\mathbf{x} = \mu\mathbf{x}$; se $B\mathbf{x} = \mathbf{0}$ allora $\mathbf{x}$ è autovettore di B corrispondente all'autovalore nullo. Si prosegua per induzione, seguendo la linea della dimostrazione del teorema di Schur e notando che una matrice unitaria Q con la prima colonna uguale a $\mathbf{x}$ è tale che

$$Q^H A Q = \begin{bmatrix} \lambda & \mathbf{c}^H \\ \mathbf{0} & A_1 \end{bmatrix}, \qquad Q^H B Q = \begin{bmatrix} \mu & \mathbf{d}^H \\ \mathbf{0} & B_1 \end{bmatrix},$$

e A_1 e B_1 commutano.)

2.19 Sia $A \in \mathbf{C}^{n\times n}$ normale. Si dimostri che:

a) se $p(x)$ è un polinomio in x, allora $p(A)$ è una matrice normale (si veda l'esercizio 1.3 per la definizione di polinomio di una matrice);

b) si può scrivere $A = B + \mathbf{i}C$, dove B e C sono matrici hermitiane che commutano e gli autovalori λ di A sono della forma $\lambda = \alpha + \mathbf{i}\beta$, dove α e $\beta \in \mathbf{R}$ sono gli autovalori di B e di C e A, B e C hanno gli stessi autovettori;

c) A è normale se e solo se ogni autovettore di A è anche autovettore di A^H.

(Traccia: a) si verifichi che A^k, per k intero positivo, è una matrice normale; b) si ponga $B = \dfrac{1}{2}(A + A^H)$ e $C = -\dfrac{\mathbf{i}}{2}(A - A^H)$, e si utilizzi il risultato b) dell'esercizio 2.18; c) sia $A = UDU^H$ la forma normale di Schur di A, allora $A^H = UD^HU^H$, viceversa si dimostri prima che A ha n autovettori linearmente indipendenti sfruttando la forma normale di Jordan di A e si verifichi poi che $A^HA\mathbf{x} = AA^H\mathbf{x}$ per ogni autovettore $\mathbf{x}$.)

2.20 Sia A la matrice diagonale a blocchi

$$A = \begin{bmatrix} A_{11} & & & \\ & A_{22} & & \\ & & \ddots & \\ & & & A_{nn} \end{bmatrix},$$

con blocchi diagonali quadrati.

a) Si dimostri che lo spettro degli autovalori di A è dato dall'unione degli spettri dei blocchi $A_{11}, A_{22}, \ldots, A_{nn}$;

b) si dimostri che la stessa proprietà vale se la matrice A è triangolare a blocchi;

c) si dica come sono fatti gli autovettori di A.

2.21 Sia $A \in \mathbf{C}^{n\times n}$ hermitiana e sia $\mathbf{x}$ un vettore non nullo. Si dimostri che se A ha k autovalori distinti, con $k < n$, allora il sottospazio S generato dai $k+1$ vettori $\mathbf{x}, A\mathbf{x}, \ldots, A^{k-1}\mathbf{x}, A^k\mathbf{x}$ ha dimensione minore o uguale a k. (Traccia: siano $\lambda_1, \ldots, \lambda_k$ gli autovalori distinti di A, di molteplicità algebrica $m_1, \ldots, m_k$. Risulta $A = UDU^H$, dove U è unitaria e

$$D = \begin{bmatrix} \lambda_1 I_{m_1} & & & \\ & \lambda_2 I_{m_2} & & \\ & & \ddots & \\ & & & \lambda_k I_{m_k} \end{bmatrix}.$$

Si dimostri che esiste un vettore $\mathbf{y} \in \mathbf{C}^{k+1}$, $\mathbf{y} \neq \mathbf{0}$, tale che

$$\begin{bmatrix} 1 & \lambda_1 & \ldots & \lambda_1^k \\ 1 & \lambda_2 & \ldots & \lambda_2^k \\ \vdots & & & \vdots \\ 1 & \lambda_k & \ldots & \lambda_k^k \end{bmatrix} \mathbf{y} = \mathbf{0}.$$

Posto $\mathbf{z} = U^H\mathbf{x}$ e $B = [\mathbf{z} \mid D\mathbf{z} \mid \ldots \mid D^k\mathbf{z}] \in \mathbf{C}^{n\times(k+1)}$, il vettore $\mathbf{y}$ è tale che $B\mathbf{y} = \mathbf{0}$, per cui la matrice

$$[\mathbf{x} \mid A\mathbf{x} \mid \ldots \mid A^k\mathbf{x}] = UB$$

ha rango minore o uguale a k.)

2.22 Siano $A \in \mathbf{C}^{m\times n}$ e $B \in \mathbf{C}^{n\times m}$, con $m \geq n$. Si dimostri che

a) $2n$ autovalori della matrice

$$\begin{bmatrix} O_m & A \\ B & O_n \end{bmatrix}$$

sono le radici quadrate degli autovalori di BA, gli altri $m-n$ sono nulli;

b) n autovalori della matrice AB coincidono con quelli della matrice BA, gli altri $m-n$ sono nulli; quindi se $m=n$ le due matrici AB e BA hanno gli stessi autovalori (si confronti con l'esercizio 2.13).

(Traccia: a) si verifichi che

$$\begin{bmatrix} I_m & O \\ \lambda^{-1}B & I_n \end{bmatrix} \begin{bmatrix} -\lambda I_m & A \\ B & -\lambda I_n \end{bmatrix} = \begin{bmatrix} -\lambda I_m & A \\ O & \lambda^{-1}BA - \lambda I_n \end{bmatrix},$$

da cui per il teorema di Binet risulta

$$P(\lambda) = \det \begin{bmatrix} -\lambda I_m & A \\ B & -\lambda I_n \end{bmatrix} = \det(-\lambda I_m)\ \det(\lambda^{-1}BA - \lambda I_n)$$

$$= (-\lambda)^m \lambda^{-n} \det(BA - \lambda^2 I_n) = (\text{-}1)^m \lambda^{m-n} \det(BA - \lambda^2 I_n).$$

b) oltre alla relazione del caso a), considerando il prodotto

$$\begin{bmatrix} I_m & \lambda^{-1}A \\ O & I_n \end{bmatrix} \begin{bmatrix} -\lambda I_m & A \\ B & -\lambda I_n \end{bmatrix} = \begin{bmatrix} \lambda^{-1}AB - \lambda I_m & O \\ B & -\lambda I_n \end{bmatrix},$$

si ottiene la relazione

$$P(\lambda) = (\text{-}1)^n \lambda^{n-m} \det(AB - \lambda^2 I_m).$$

Uguagliando le due relazioni e ponendo $\mu = \lambda^2$ si ha

$$\det(AB - \mu I_m) = (-\mu)^{m-n} \det(BA - \mu I_n).\ )$$

2.23 Siano $\mathbf{x}, \mathbf{y} \in \mathbf{C}^n$. Si determinino gli autovalori e gli autovettori della matrice

$$A = \mathbf{x}\mathbf{y}^H.$$

(Traccia: si tenga conto che la matrice $\mathbf{x}\mathbf{y}^H$ ha rango 1 e che $\mathbf{x}\mathbf{y}^H\mathbf{x} = (\mathbf{y}^H\mathbf{x})\mathbf{x}$; quali sono i vettori $\mathbf{z} \neq \mathbf{0}$ tali che $\mathbf{x}\mathbf{y}^H\mathbf{z} = \mathbf{0}$? Oppure si sfrutti il punto b) dell'esercizio 2.22.)

2.24 Siano $\mathbf{x}, \mathbf{y}, \mathbf{u}, \mathbf{v} \in \mathbf{C}^n$ e

$$A = \mathbf{x}\mathbf{y}^H + \mathbf{u}\mathbf{v}^H, \qquad B = \begin{bmatrix} \mathbf{y}^H\mathbf{x} & \mathbf{y}^H\mathbf{u} \\ \mathbf{v}^H\mathbf{x} & \mathbf{v}^H\mathbf{u} \end{bmatrix}.$$

Si dica che relazione c'è fra gli autovalori di A e quelli di B.

(Traccia: si sfrutti il punto b) dell'esercizio 2.22, ponendo

$$A = [\mathbf{x} \mid \mathbf{u}], \qquad B^H = [\mathbf{y} \mid \mathbf{v}]. \ )$$

2.25 Siano $\mathbf{x}, \mathbf{y}, \mathbf{u}, \mathbf{v} \in \mathbf{C}^n$. Si scrivano il polinomio caratteristico e il polinomio minimo delle seguenti matrici

$$a) \quad \mathbf{x}\mathbf{y}^H, \quad b) \quad I + \mathbf{x}\mathbf{y}^H, \quad c) \quad \mathbf{x}\mathbf{y}^H + \mathbf{u}\mathbf{v}^H, \quad d) \quad I + \mathbf{x}\mathbf{y}^H + \mathbf{u}\mathbf{v}^H.$$

2.26 Siano A e $B \in \mathbf{C}^{n\times n}$. Si dimostri che

a) lo spettro degli autovalori della matrice

$$C = \begin{bmatrix} A & B \\ B & A \end{bmatrix}$$

è costituito dall'unione degli spettri delle matrici

$$A + B \quad \text{e} \quad A - B,$$

e gli autovettori di C sono i vettori

$$\begin{bmatrix} \mathbf{x} \\ \mathbf{x} \end{bmatrix} \quad \text{e} \quad \begin{bmatrix} \mathbf{y} \\ -\mathbf{y} \end{bmatrix},$$

in cui $\mathbf{x}$ e $\mathbf{y}$ sono gli autovettori di $A + B$ e $A - B$;

b) lo spettro degli autovalori della matrice

$$C = \begin{bmatrix} A & -B \\ B & A \end{bmatrix}$$

è costituito dall'unione degli spettri delle matrici

$$A + \mathbf{i}B \quad \text{e} \quad A - \mathbf{i}B,$$

e gli autovettori sono i vettori

$$\begin{bmatrix} \mathbf{x} \\ -\mathbf{i}\mathbf{x} \end{bmatrix} \quad \text{e} \quad \begin{bmatrix} \mathbf{y} \\ \mathbf{i}\mathbf{y} \end{bmatrix}.$$

(Traccia: si consideri la matrice

$$S = \begin{bmatrix} I_n & I_n \\ I_n & -I_n \end{bmatrix}$$

e si costruisca la matrice $D = S^{-1}CS$; b) come per il punto a) considerando la matrice

$$S = \begin{bmatrix} I_n & I_n \\ -\mathbf{i}I_n & \mathbf{i}I_n \end{bmatrix}.)$$

2.27 Siano A e $B \in \mathbf{R}^{n\times n}$ e si considerino le matrici

$$V = A + \mathbf{i}B, \qquad W = \begin{bmatrix} A & -B \\ B & A \end{bmatrix}.$$

Si dimostri che:

a) V è normale se e solo se W è normale;

b) V è hermitiana se e solo se W è simmetrica;

c) V è definita positiva se e solo se W è definita positiva;

d) V è unitaria se e solo se W è ortogonale;

e) se gli autovalori di V sono reali, allora anche gli autovalori di W sono reali e coincidono con quelli di V.

(Traccia: e) si verifichi che gli autovalori di $A + \mathbf{i}B$ coincidono con quelli di $A - \mathbf{i}B$ e si applichi l'esercizio 2.26.)

2.28 Sia $B \in \mathbf{C}^{n\times n}$ una matrice definita positiva. Si dimostri che

$$\det B \leq \prod_{j=1}^{n} b_{jj},$$

in cui il segno di uguaglianza vale solo nel caso di matrici diagonali. Sfruttando questa relazione si dimostri la seguente *disuguaglianza di Hadamard* per una matrice $A \in \mathbf{C}^{n\times n}$

$$|\det A|^2 \leq \prod_{j=1}^{n} \sum_{i=1}^{n} |a_{ij}|^2,$$

e quindi la maggiorazione

$$|\det A| \leq (M\sqrt{n})^n,$$

in cui $M = \max_{i,j=1,\ldots,n} |a_{ij}|$.

(Traccia: si proceda per induzione; per $n > 1$ si ha

$$\det B = b_{11} \det Z + \det \begin{bmatrix} 0 & \mathbf{v}^H \\ \mathbf{v} & Z \end{bmatrix},$$

in cui Z è la sottomatrice principale di B ottenuta cancellando la prima riga e la prima colonna e $\mathbf{v}$ è il vettore formato dalla prima colonna di B, escluso il primo elemento. Poiché (si veda l'esercizio 1.43) è

$$\det \begin{bmatrix} 0 & \mathbf{v}^H \\ \mathbf{v} & Z \end{bmatrix} = \det Z \det(-\mathbf{v}^H Z^{-1} \mathbf{v}) \leq 0$$

perché Z è definita positiva, ne segue che $\det B \leq b_{11} \det Z$. Per la disuguaglianza di Hadamard, se $\det A \neq 0$, si applichi la relazione precedente alla matrice definita positiva $B = AA^H$.)

2.29 Sia $A \in \mathbf{C}^{m \times n}$, $m \geq n$, e B la matrice

$$B = \begin{bmatrix} I_m & A \\ A^H & I_n \end{bmatrix}.$$

Si dimostri che B è definita positiva se e solo se $\rho(A^H A) < 1$.

(Traccia: per l'esercizio 2.22 gli autovalori di B diversi da 1 sono dati da $1 \pm \sqrt{\mu}$, dove μ è autovalore di $A^H A$.)

2.30 Sia $A \in \mathbf{C}^{n \times n}$ è una matrice non singolare. Si dimostri che

a) esistono una matrice U unitaria e una matrice B definita positiva, tali che
$$A = BU;$$

b) A è normale, se e solo se le matrici U e B commutano.

(Traccia: a) sia B tale che $B^2 = AA^H$; b) $A^H A = U^H B^2 U$.)

2.31 Sia $A \in \mathbf{C}^{n \times n}$ idempotente. Si dimostri che

a) A è diagonalizzabile;

b) i suoi autovalori sono uguali a 0 o a 1 e il rango di A è uguale alla sua traccia;

c) ogni matrice diagonalizzabile, avente solo autovalori uguali a 0 o a 1 è idempotente;

d) la sola matrice idempotente non singolare è la matrice identica;

e) si consideri in particolare il caso

$$A = \begin{bmatrix} 2 & 1 & 1 \\ 6 & 3 & 3 \\ -8 & -4 & -4 \end{bmatrix}.$$

(Per la definizione e le proprietà delle matrici idempotenti, si veda l'esercizio 1.9). (Traccia: a) e b) si consideri la forma normale di Jordan; d) $I = A^{-1}A = A^{-1}A^2 = A$.)

2.32 Sia $A \in \mathbf{C}^{n\times n}$. Si dimostri che

a) A è nilpotente se e solo se i suoi autovalori sono nulli;

b) A è nilpotente se e solo se

$$\text{tr } A^k = 0 \quad \text{per ogni } k \text{ intero positivo;}$$

c) se A è nilpotente e non nulla, A non è diagonalizzabile;

d) se A è nilpotente, allora esiste un intero k tale che per ogni $\alpha \in \mathbf{C}$ è

$$\det(\alpha I - A) = \alpha^k;$$

e) se A è nilpotente e A e $B \in \mathbf{C}^{n\times n}$ commutano, le due matrici $A + B$ e B hanno gli stessi autovalori. Si consideri in particolare il caso

$$A = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & -1 & -1 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & -2 & 3 \\ -3 & 0 & -8 \\ 5 & 5 & 8 \end{bmatrix}.$$

(Per la definizione e le proprietà delle matrici nilpotenti, si veda l'esercizio 1.10). (Traccia: e) se λ è autovalore di $A + B$, cioè $(A + B)\mathbf{x} = \lambda\mathbf{x},\ \mathbf{x} \neq \mathbf{0}$, sia k il minimo intero tale che $A^k\mathbf{x} = \mathbf{0}$; da $A^{k-1}(A + B)\mathbf{x} = \lambda A^{k-1}\mathbf{x}$ segue per la commutatività che $B\mathbf{y} = \lambda\mathbf{y}$, in cui $\mathbf{y} = A^{k-1}\mathbf{x}$; il viceversa è analogo.)

2.33 Sia $A \in \mathbf{C}^{n\times n}$ antihermitiana.

a) Si dimostri che i suoi autovalori o sono nulli o sono numeri immaginari puri;

b) se A è antisimmetrica, si dimostri che la forma normale reale di Schur di A è

$$A = UDU^H,$$

in cui U è ortogonale e D è la matrice diagonale a blocchi

$$D = \begin{bmatrix} D_{11} & & & \\ & D_{22} & & \\ & & \ddots & \\ & & & D_{mm} \end{bmatrix},$$

così formata: se n è pari, allora $m = \dfrac{n}{2}$ e

$$D_{ii} = \begin{bmatrix} 0 & r_i \\ -r_i & 0 \end{bmatrix}, \quad r_i \in \mathbf{R}, \text{ per } i = 1, \ldots, m,$$

e se n è dispari, allora $m = \dfrac{n+1}{2}$ e

$$D_{ii} = \begin{bmatrix} 0 & r_i \\ -r_i & 0 \end{bmatrix}, \quad r_i \in \mathbf{R}, \text{ per } i = 1, \ldots, m-1, \qquad D_{mm} = [0],$$

dove alcuni degli r_i possono essere nulli.

(Traccia: b) si tenga conto del fatto che anche D è antisimmetrica.)

2.34 Siano A e $B \in \mathbf{C}^{n\times n}$ due matrici hermitiane. Si dimostri che

a) se λ è autovalore di AB, anche $\overline{\lambda}$ lo è;

b) gli autovalori di $AB - BA$ sono nulli o numeri immaginari puri.

(Traccia: a) si noti che $(AB)^H = BA$ e si sfrutti l'esercizio 2.22; b) la matrice $AB - BA$ è antihermitiana e si sfrutti l'esercizio 2.33.)

2.35 Una matrice $A \in \mathbf{R}^{n\times n}$ si dice *centrosimmetrica* se

$$JAJ = A \qquad \text{dove} \qquad J = \begin{bmatrix} & & & 1 \\ & & \iddots & \\ & 1 & & \\ 1 & & & \end{bmatrix}.$$

a) Si scriva una matrice $A \in \mathbf{R}^{2\times 2}$ centrosimmetrica;

b) si determini la dimensione dello spazio vettoriale generato dalle matrici centrosimmetriche;

c) se A è centrosimmetrica, si dimostri che per ogni autovalore di A esiste un autovettore $\mathbf{x}$ tale che $\mathbf{x} = J\mathbf{x}$ (detto *centrosimmetrico*) o tale che $\mathbf{x} = -J\mathbf{x}$ (detto *centroantisimmetrico*);

d) se A è centrosimmetrica ed n è un numero pari, si determini una matrice $Q \in \mathbf{C}^{n\times n}$ ortogonale, tale che

$$Q^H A Q = \begin{bmatrix} A_1 & O \\ O & A_2 \end{bmatrix},$$

dove A_1 e $A_2 \in \mathbf{C}^{n/2\times n/2}$. Si esamini anche il caso di n dispari.

(Traccia: c) se $A\mathbf{x} = \lambda\mathbf{x}$, $\mathbf{x} \neq \mathbf{0}$, allora $JAJJ\mathbf{x} = \lambda J\mathbf{x}$, quindi $A\mathbf{y} = \lambda\mathbf{y}$, con $\mathbf{y} = J\mathbf{x}$. Inoltre almeno uno dei vettori $\mathbf{x} + J\mathbf{x}$ e $\mathbf{x} - J\mathbf{x}$ è non nullo ed è $\mathbf{x} + J\mathbf{x}$ centrosimmetrico e $\mathbf{x} - J\mathbf{x}$ centroantisimmetrico; d) se n è pari vale

$$A = \begin{bmatrix} B_1 & B_2 J \\ J B_2 & J B_1 J \end{bmatrix},$$

dove B_1, B_2, $J \in \mathbf{R}^{n/2 \times n/2}$. Si applichi l'esercizio 2.26 essendo

$$A = \begin{bmatrix} I & O \\ O & J \end{bmatrix} \begin{bmatrix} B_1 & B_2 \\ B_2 & B_1 \end{bmatrix} \begin{bmatrix} I & O \\ O & J \end{bmatrix}. \;)$$

2.36 Una matrice $A \in \mathbf{R}^{n \times n}$ si dice *persimmetrica* se è simmetrica rispetto alla diagonale secondaria, cioè se $a_{ij} = a_{n-j+1,n-i+1}$. Si dimostri che:

a) A è persimmetrica se e solo se JA è simmetrica;

b) A è persimmetrica se e solo se AJ è simmetrica;

c) A è persimmetrica se e solo se $JAJ = A^T$;

d) se A è simmetrica e persimmetrica, allora A è centrosimmetrica.

e) Si verifichi che sono persimmetriche le matrici di *Toeplitz*, cioè matrici definite per mezzo di $2n-1$ scalari $r_{-n+1}, \ldots, r_0, \ldots, r_{n-1}$, nel modo seguente

$$a_{ij} = r_{j-i}, \quad i,\ j = 1, \ldots, n.$$

f) Si verifichi che la matrice

$$A_{2n} = \begin{bmatrix} a_n & & & & & & & b_n \\ & a_{n-1} & & & & & b_{n-1} & \\ & & \ddots & & & \iddots & & \\ & & & a_1 & b_1 & & & \\ & & & -b_1 & a_1 & & & \\ & & \iddots & & & \ddots & & \\ & -b_{n-1} & & & & & a_{n-1} & \\ -b_n & & & & & & & a_n \end{bmatrix}.$$

è persimmetrica e se ne calcoli gli autovalori. (Per la definizione di matrice centrosimmetrica e per la matrice J si veda l'esercizio 2.35). (Risposta: f) il polinomio caratteristico della matrice A_{2n} è dato da

$$p_{2n}(\lambda) = \prod_{k=1}^{n} [(a_k - \lambda)^2 + b_k]. \;)$$

2.37 Sia

$$\omega = \cos\frac{2\pi}{n} + \mathbf{i}\sin\frac{2\pi}{n}$$

una radice n-esima dell'unità: si dimostri che

a) ω è *primitiva*, cioè che vale $\omega^k = 1$ se e solo se k è un multiplo intero di n;

b) $$\sum_{k=0}^{n-1} \omega^{jk} = \begin{cases} n & \text{se } \omega^j = 1, \text{ cioè se } j = 0 \bmod n, \\ 0 & \text{se } \omega^j \neq 1, \text{ cioè se } j \neq 0 \bmod n; \end{cases}$$

c) si dimostri che la matrice

$$\Omega = \frac{1}{\sqrt{n}} \begin{bmatrix} 1 & 1 & \dots & 1 \\ 1 & \omega & \dots & \omega^{n-1} \\ 1 & \omega^2 & \dots & \omega^{2(n-1)} \\ \vdots & \vdots & & \vdots \\ 1 & \omega^{n-1} & \dots & \omega^{(n-1)(n-1)} \end{bmatrix},$$

il cui elemento (i,j)-esimo è dato da

$$\omega_{ij} = \frac{1}{\sqrt{n}}\,\omega^{(i-1)(j-1)}, \quad \text{per } i,j = 1,\dots,n,$$

è unitaria;

d) sia $A \in \mathbf{C}^{n\times n}$ la matrice circolante

$$A = \begin{bmatrix} \alpha_1 & \alpha_2 & \alpha_3 & \dots & \alpha_n \\ \alpha_n & \alpha_1 & \alpha_2 & \dots & \alpha_{n-1} \\ \alpha_{n-1} & \alpha_n & \alpha_1 & \dots & \alpha_{n-2} \\ \vdots & & & & \vdots \\ \alpha_2 & \alpha_3 & \dots & \alpha_n & \alpha_1 \end{bmatrix},$$

dove $\alpha_1, \alpha_2, \dots, \alpha_n \in \mathbf{C}$, definita nell'esercizio 1.58; si dimostri che $A = \Omega D \Omega^H$, dove D è la matrice diagonale il cui i-esimo elemento principale è l'autovalore di A

$$\lambda_i = \sum_{j=1}^{n} \alpha_j \omega^{(i-1)(j-1)}, \qquad i = 1,2,\dots,n.$$

Quindi la matrice A è normale.

e) Si determinino in particolare gli autovalori delle matrici

$$A = \begin{bmatrix} \text{-2} & 1 & 0 & 1 \\ 1 & \text{-2} & 1 & 0 \\ 0 & 1 & \text{-2} & 1 \\ 1 & 0 & 1 & \text{-2} \end{bmatrix},$$

$$B = \begin{bmatrix} \alpha & \beta & \dots & \beta \\ \beta & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \beta \\ \beta & \dots & \beta & \alpha \end{bmatrix}.$$

La trasformazione $\mathbf{u} \to \mathbf{v} = \dfrac{1}{\sqrt{n}}\, \Omega^H \mathbf{u}$, con $\mathbf{u}, \mathbf{v} \in \mathbf{C}^n$, è detta *trasformata discreta di Fourier*. La trasformazione $\mathbf{v} \to \mathbf{u} = \sqrt{n}\, \Omega \mathbf{v}$ è detta *trasformata inversa discreta di Fourier*.
(Traccia: b) dalla relazione

$$(1 + z + z^2 + \ldots + z^{n-1})\,(z-1) = z^n - 1$$

si ha che, ponendo $z = \omega^j$, è $z^n = 1$ e se $z \neq 1$, ne segue $1+z+\ldots+z^{n-1} = 0$; c) l'elemento (i,j)-esimo di $\Omega^H \Omega$ è dato da

$$\begin{aligned} &\frac{1}{n} \sum_{k=1}^{n} \overline{\omega}^{\,(i-1)(k-1)} \omega^{(k-1)(j-1)} \\ =&\frac{1}{n} \sum_{k=0}^{n-1} \omega^{-(i-1)k} \omega^{k(j-1)} = \frac{1}{n} \sum_{k=0}^{n-1} \omega^{k(j-i)} = \begin{cases} 1 & \text{se } j = i, \\ 0 & \text{se } j \neq i; \end{cases} \end{aligned}$$

d) posto $C = \begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \\ 1 & & & 0 \end{bmatrix}$, vale $A = \sum_{i=1}^{n} \alpha_i C^{i-1}$. È quindi sufficiente dimostrare che $\Omega^H C \Omega = \begin{bmatrix} 1 & & & & \\ & \omega & & & \\ & & \omega^2 & & \\ & & & \ddots & \\ & & & & \omega^{n-1} \end{bmatrix}$, utilizzando il risultato del punto b); e) gli autovalori di A sono 0 e -4 di molteplicità 1 e -2 di molteplicità 2; gli autovalori di B sono $\lambda_i = \alpha - \beta$ per $i = 1, 2, \ldots, n-1$ e $\lambda_n = \alpha + (n-1)\beta$.)

2.38 Siano $A_1, A_2, \ldots, A_m \in \mathbf{C}^{n \times n}$ e $A \in \mathbf{C}^{nm \times nm}$

$$A = \begin{bmatrix} O & O & \dots & O & -A_m \\ I & O & \dots & O & -A_{m-1} \\ O & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & O & -A_2 \\ O & \dots & O & I & -A_1 \end{bmatrix}.$$

Si dimostri che il polinomio caratteristico di A è dato da

$$\det(P(\lambda)), \quad \text{dove } P(\lambda) = \lambda^m I + \lambda^{m-1} A_1 + \cdots + A_m.$$

Inoltre se

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_m \end{bmatrix}$$

è un autovettore di A corrispondente ad un autovalore λ, allora

$$P(\lambda)\mathbf{x}_m = 0.$$

(Traccia: si costruisca una matrice $B \in \mathbf{C}^{nm\times nm}$ bidiagonale a blocchi, con i blocchi diagonali uguali a I_n, tale che la matrice BA sia triangolare superiore a blocchi.)

2.39 Sia $A_n \in \mathbf{C}^{n\times n}$, per $n \geq 2$, la matrice (particolare *matrice ad albero*, si veda la definizione dell'esercizio 1.59)

$$A_n = \begin{bmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_{n-1} & \alpha_n \\ \alpha_2 & \beta & \cdots & 0 & 0 \\ \vdots & & \ddots & & \vdots \\ \alpha_{n-1} & 0 & \cdots & \beta & 0 \\ \alpha_n & 0 & \cdots & 0 & \beta \end{bmatrix}.$$

Si dimostri che il polinomio caratteristico di A_n soddisfa alla seguente relazione ricorrente

$$P_n(\lambda) = (\beta - \lambda)P_{n-1}(\lambda) - \alpha_n^2(\beta - \lambda)^{n-2}, \quad P_1(\lambda) = \alpha_1 - \lambda,$$

e si determinino gli autovalori di A_n.

(Risposta: $\lambda_{1,2} = \dfrac{1}{2}(\alpha_1 + \beta) \pm \sqrt{\dfrac{1}{4}(\alpha_1 - \beta)^2 + \alpha_2^2 + \cdots + \alpha_n^2}$, e $\lambda_k = \beta$, per $k = 3, \ldots, n$.)

2.40 Sia A la matrice tridiagonale (di Jacobi)

$$A_n = \begin{bmatrix} \alpha_1 & \gamma_2 & & \\ \beta_2 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \gamma_n \\ & & \beta_n & \alpha_n \end{bmatrix},$$

in cui gli $\alpha_i,\ i = 1, \ldots, n,\ \beta_i$ e $\gamma_i,\ i = 2, \ldots, n$, sono numeri complessi. Si dimostri che

a) il polinomio caratteristico di A può essere ricavato mediante una relazione ricorrente a tre termini;

b) se λ è un autovalore di $A,\ \lambda$ è anche autovalore della matrice ottenuta da A cambiando di segno tutti gli elementi non principali;

c) se gli $\alpha_i,\ i = 1, 2, \ldots, n$ sono reali e i prodotti $\beta_i\gamma_i,\ i = 2, \ldots, n$ sono reali positivi, allora gli autovalori di A sono reali;

d) se $\alpha_i = \alpha$, per $i = 1, 2, \ldots, n$, allora ogni autovalore λ di A verifica la relazione

$$|\lambda - \alpha| \leq 2\sqrt{\max_{i=1,\ldots,n} |\beta_i| \max_{i=1,\ldots,n} |\gamma_i|}\ ;$$

e) si determinino gli autovalori e gli autovettori della matrice tridiagonale B i cui elementi sono

$$b_{ij} = \begin{cases} 1 & \text{se } i = j - 1 \text{ e } i = j + 1, \\ \\ 0 & \text{altrimenti;} \end{cases}$$

f) si faccia vedere come il calcolo degli autovalori di una matrice tridiagonale A in cui $\alpha_i = \alpha$, per $i = 1, 2, \ldots, n$, e $\beta_i = \beta,\ \gamma_i = \gamma$ per $i = 2, \ldots, n$, possa essere ricondotto al calcolo degli autovalori della matrice B del punto e). Si esamini in particolare la matrice C i cui elementi sono

$$c_{ij} = \begin{cases} -1 & \text{se } i = j - 1, \\ 1 & \text{se } i = j + 1, \\ 0 & \text{altrimenti;} \end{cases}$$

g) si dica per quali valori del parametro β la matrice A i cui elementi sono

$$a_{ij} = \begin{cases} 1 & \text{se } i = j, \\ \beta & \text{se } |i - j| = 1, \\ 0 & \text{altrimenti,} \end{cases}$$

è definita positiva.

(Traccia: a) si sviluppi $\det(A_n - \lambda I)$ con la regola di Laplace applicata all'ultima riga: si ottiene

$$p_1(\lambda) = \alpha_1 - \lambda, \quad p_2(\lambda) = (\alpha_1 - \lambda)(\alpha_2 - \lambda) - \beta_2\gamma_2,$$

$$p_n(\lambda) = (\alpha_n - \lambda)p_{n-1}(\lambda) - \beta_n\gamma_n p_{n-2}(\lambda);$$

b) si consideri la matrice HAH^{-1}, dove H è la matrice diagonale il cui i-esimo elemento principale è $(-1)^i$; c) si veda il punto d) dell'esercizio 1.50; d) si applichi il primo teorema di Gerschgorin alla matrice $B = DAD^{-1}$ a elementi complessi, tale che $B = B^T$, in cui D è diagonale; e) indicato con λ un autovalore di B e con $\mathbf{x} = [x_1, x_2, \ldots, x_n]^T$ un autovettore corrispondente, si ha che

$$x_{j-1} - \lambda x_j + x_{j+1} = 0, \quad \text{per } j = 1, 2, \ldots, n,$$

con le condizioni $x_0 = x_{n+1} = 0$, che è un'equazione alle differenze con condizioni al contorno. Si cercano soluzioni di tale equazione del tipo $x_i = t^i$, $t \neq 0$; sostituendo risulta $t^{-1} - \lambda + t = 0$, cioè $\lambda = t + 1/t$, in cui t e $1/t$ sono le soluzioni dell'equazione di secondo grado $y^2 - \lambda y + 1 = 0$; la soluzione dell'equazione alle differenze è allora $x_j = c_1 t^j + c_2 t^{-j}$, dove c_1 e c_2 sono tali che $x_0 = c_1 + c_2 = 0$ e $x_{n+1} = c_1 t^{n+1} + c_2 t^{-(n+1)} = 0$. Ne segue che $c_1 = -c_2$ e, dovendo essere $c_1 \neq 0$, è $t^{2(n+1)} = 1$, da cui si ottengono per t i k valori complessi

$$t_k = \cos \frac{k\pi}{n+1} + \mathbf{i} \sin \frac{k\pi}{n+1}, \quad k = 1, 2, \ldots, n.$$

Ne segue che gli autovalori di B sono

$$\lambda_k = t_k + \frac{1}{t_k} = 2 \cos \frac{k\pi}{n+1}, \quad k = 1, 2, \ldots, n,$$

e i corrispondenti autovettori $\mathbf{x}^{(k)}$ hanno le componenti

$$x_j^{(k)} = \sin \frac{kj\pi}{n+1}, \quad j = 1, 2, \ldots, n.$$

f) si determini una matrice diagonale D tale che

$$D(A - \alpha I)D^{-1} = \sqrt{\beta\gamma}B.$$

Per la matrice C, il j-esimo elemento principale di D è $d_{jj} = \mathbf{i}^j$; g) gli autovalori di A sono $\lambda_k = 1 - 2\beta \cos \dfrac{k\pi}{n+1}$, $k = 1, \ldots, n$, e quindi A è definita positiva se e solo se $|\beta| < \dfrac{1}{2} \sec \dfrac{\pi}{n+1}$.)

2.41 Siano $A \in \mathbf{C}^{n\times n}$ e $B \in \mathbf{C}^{m\times m}$. Si dimostri che se λ_i, $i = 1, 2, \ldots, n$ e μ_j, $j = 1, 2, \ldots, m$ sono gli autovalori rispettivamente di A e B, allora

a) gli autovalori di $A \otimes B$ (si veda l'esercizio 1.60) sono dati da $\lambda_i \mu_j$, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m$, e quindi le matrici $A \otimes B$ e $B \otimes A$ hanno gli stessi autovalori e tr $A \otimes B$=(tr A)(tr B);

b) gli autovalori di $(I_m \otimes A) + (B \otimes I_n)$ sono dati da $\lambda_i + \mu_j,\ i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, m$.

(Traccia: a) siano $A = SJ_1S^{-1}$ e $B = TJ_2T^{-1}$ le forme normali di Jordan di A e di B; si verifichi che

$$A \otimes B = (S \otimes T)\,(J_1 \otimes J_2)\,(S \otimes T)^{-1},$$

in cui $J_1 \otimes J_2$ è triangolare superiore con elementi diagonali uguali a $\lambda_i\mu_j$, $i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, m$; si proceda in modo analogo per il punto b).)

2.42 Siano A, B e $C \in \mathbf{C}^{n\times n}$.

a) Si determinino le condizioni necessarie e sufficienti affinché il sistema lineare

$$AX + XB = C$$

abbia un'unica soluzione $X \in \mathbf{C}^{n\times n}$.

b) Si esamini in particolare il caso in cui $B = A$ e si risolva il sistema per

$$A = \begin{bmatrix} 1 & 1 \\ 0 & \text{-}2 \end{bmatrix}, \qquad C = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}.$$

(Traccia: a) il sistema può essere scritto nella forma

$$(I \otimes A + B^T \otimes I) \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{bmatrix} = \begin{bmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \\ \vdots \\ \mathbf{c}_n \end{bmatrix},$$

in cui $\mathbf{x}_i$ e $\mathbf{c}_i,\ i = 1, 2, \ldots, n$, sono le colonne delle matrici X e C. Gli autovalori di $I \otimes A + B^T \otimes I$ sono date dalle somme degli autovalori di A e di B (esercizio 2.41). Quindi il sistema ha un'unica soluzione se e solo se non esistono autovalori di A tali che $-\lambda$ sia autovalore di B; b) se $B = A$, la condizione è che A non abbia coppie di autovalori λ e $-\lambda$.)

2.43 La serie

$$I + A + \frac{1}{2!}A^2 + \frac{1}{3!}A^3 + \cdots + \frac{1}{n!}A^n + \cdots$$

converge per ogni matrice $A \in \mathbf{C}^{n\times n}$. La somma della serie si indica con e^A. Si verifichi che

a) se $A = SBS^{-1}$, allora $\mathrm{e}^A = S\mathrm{e}^BS^{-1}$;

b) se $\lambda_1, \lambda_2, \ldots, \lambda_n$ sono gli autovalori di A, allora e^{λ_1}, e^{λ_2}, ..., e^{λ_n} sono gli autovalori di e^A;

c) se A e $B \in \mathbf{C}^{n\times n}$ commutano, allora $e^{A+B}=e^A\ e^B$;

d) se $U \in \mathbf{C}^{n\times n}$ è una matrice unitaria, allora esiste una matrice A tale che $U = e^{iA}$;

e) per la matrice

$$A = \begin{bmatrix} 1+\alpha & -\alpha \\ \alpha & 1-\alpha \end{bmatrix}$$

si determinino per ogni k i coefficienti p e q tali che

$$A^k = pI + qA,$$

e si calcoli e^A.

(Risposta: e) $p = 1-k,\ q = k$, $e^A = eA$.)

2.44 Sia $A \in \mathbf{C}^{n\times n}$ e posto

$$s_i = \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|, \quad \text{per } i = 1, 2, \ldots, n,$$

si definisca l'insieme

$$C_{ij} = \{\ z \in \mathbf{C} \text{ tali che } |z - a_{ii}|\ |z - a_{jj}| \leq s_i s_j\ \}$$

detto *ovale di Cassini*. Si dimostri che gli autovalori di A appartengono all'unione di tutti gli ovali di Cassini di A:

$$\bigcup_{\substack{i,j=1 \\ i>j}}^{n} C_{ij}.$$

Si disegni l'unione degli ovali di Cassini della matrice

$$\begin{bmatrix} 6 & 2 & -1 \\ 1 & 2 & 1 \\ -3 & 1 & 12 \end{bmatrix}$$

e per confronto l'unione dei cerchi di Gerschgorin.

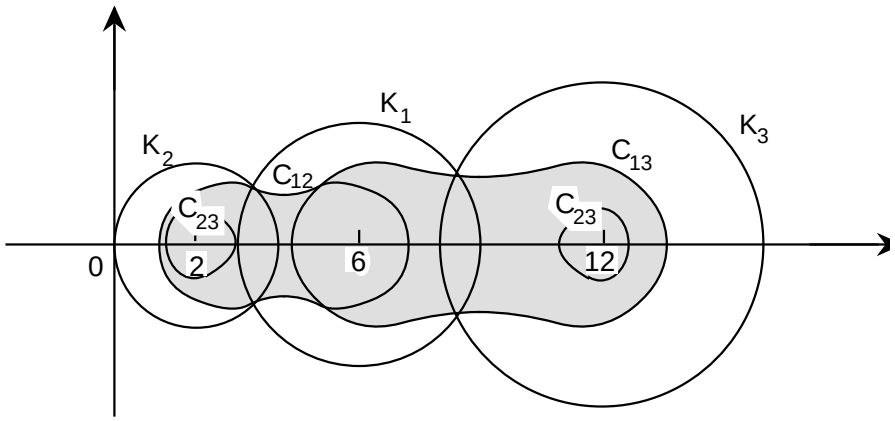
(Traccia: si segua la dimostrazione del teorema 2.35. Siano p e q gli indici delle componenti di massimo modulo dell'autovettore $\mathbf{x}$ corrispondente all'autovalore λ, tali che $|x_p| \geq |x_q|$, e si scriva la (23) per $i = p$ e $i = q$:

$$(\lambda - a_{pp})x_p = \sum_{\substack{j=1 \\ j\neq p}}^{n} a_{pj}x_j \quad \text{e} \quad (\lambda - a_{qq})x_q = \sum_{\substack{j=1 \\ j\neq q}}^{n} a_{qj}x_j.$$

Se $x_q = 0$, allora $x_j = 0$ per ogni $j \neq p$ e quindi $\lambda = a_{pp}$. Altrimenti si ha passando ai moduli

$$|\lambda - a_{pp}|\,|x_p| \leq \sum_{\substack{j=1 \\ j\neq p}}^{n} |a_{pj}|\,|x_j| \leq \sum_{\substack{j=1 \\ j\neq p}}^{n} |a_{pj}|\,|x_q|,$$

$$|\lambda - a_{qq}|\,|x_q| \leq \sum_{\substack{j=1 \\ j\neq q}}^{n} |a_{qj}|\,|x_j| \leq \sum_{\substack{j=1 \\ j\neq q}}^{n} |a_{qj}|\,|x_p|.\ )$$



Commento bibliografico

La nozione di autovalore appare per la prima volta nel 18^o secolo, non riguardo alle trasformazioni lineari, ma nella teoria delle equazioni differenziali lineari omogenee a coefficienti costanti. Nel 1762 Lagrange, affrontando il problema del moto di un sistema a n parametri in prossimità di una posizione di equilibrio, è condotto a risolvere un'equazione algebrica di grado n, la cosiddetta *equazione caratteristica*. Nel 1774 un problema analogo gli si presenta nello studio del movimento dei pianeti. La teoria degli autovalori in analisi si sviluppa successivamente ad opera soprattutto di Sturm e Liouville. Per le trasformazioni lineari la teoria degli autovalori si sviluppa

nell'ambito di quella dei determinanti, ad opera inizialmente di Cauchy, che applica alle matrici il concetto di equazione caratteristica e che nel 1829 dimostra che gli autovalori di una matrice reale e simmetrica sono reali. Nel 1853 Hamilton enuncia il risultato noto come teorema di Cayley-Hamilton per il caso particolare dei quaternioni e nel 1854 Brioschi dimostra che gli autovalori di una matrice ortogonale hanno modulo 1.

Probabilmente nessun altro termine ha avuto maggiore difficoltà ad affermarsi quanto il termine *autovalore*, nonostante la sua vasta diffusione nei più svariati campi della matematica (equazioni della forma $A\mathbf{x} = \lambda\mathbf{x}$ si presentano ad esempio nella determinazione degli assi principali di una quadrica, nella teoria delle oscillazioni, nella teoria della correlazione e nei problemi di filtraggio di segnali, nella risoluzione di equazioni differenziali e integrali, nella teoria delle catene di Markov). Inizialmente nel 1851 Sylvester usa per autovalore e autovettore i termini *latent root* e *latent point*, derivando questa definizione dal fatto che se $\mathbf{x}$ è un autovettore di A, allora il vettore $A\mathbf{x}$ è sovrapposto al vettore $\mathbf{x}$. Questi termini sopravvivono fin verso il 1950, ma affiancati dall'inizio degli anni '40 dai termini *characteristic root* (o *value* o *number*) e *characteristic vector*. Anche questi termini, a loro volta, lasciano gradualmente il campo ai termini *eigenvalue* e *eigenvector*, mutuati dai termini tedeschi *Eigenwert* e *Eigenvektor* (letteralmente *valore proprio* e *vettore proprio*, e anche in francese si usano i termini *valeur propre* e *vecteur propre*).

La terminologia italiana ha naturalmente seguito quella inglese, adottando nel tempo i termini di *radici* o *valori latenti, radici proprie* o *caratteristiche*, e attualmente di *autovalori*.

Nella seconda metà dell'ottocento una parte notevole dell'attenzione dei matematici si rivolge alla classificazione delle forme quadratiche in più variabili attraverso trasformazioni lineari che riconducono le forme a tipi semplici, e quindi alla determinazione di varie forme canoniche o normali delle matrici. Una forma normale molto simile a quella che Jordan pubblica nel 1870 era già stata trovata da Weierstrass nel 1868, e altre forme vengono trovate da Jacobi, Frobenius, Hermite. La dimostrazione che ogni matrice complessa può essere ridotta, per mezzo di una trasformazione per similitudine unitaria, nella forma di Jacobi viene data da Schur nel 1909, ed è per questo che oggi la forma normale di Jacobi è più nota come *forma normale di Schur*. La dimostrazione della riduzione di una matrice nella sua forma normale di Jordan è assai più complicata (si veda ad esempio [3]) di quella relativa alla riduzione in forma di Schur.

Anche la localizzazione degli autovalori è un argomento che ha interessato assai i matematici: ad esempio il primo teorema di Gerschgorin, è stato preceduto da molti risultati simili. Nel 1881 Levy dimostra che una matrice reale i cui termini principali sono negativi, quelli non principali sono positivi

e le cui somme degli elementi sulle righe sono negative, ha determinante non nullo, ed in particolare positivo se l'ordine è pari e negativo se l'ordine è dispari. Hadamard nel 1898 estende questo risultato al caso di una matrice ad elementi complessi. Anche Minkowski nel 1900 dà un'estensione del risultato di Levy. Il teorema di Hadamard viene conosciuto nell'ambito della scuola tedesca, ma la sua origine resta ignota, per cui nel 1931 Rohrbach facendo una generalizzazione del teorema di Hadamard all'equazione caratteristica, lo attribuisce a Minkowski. Nello stesso anno Gerschgorin, sotto ipotesi più semplici e nel campo reale, dimostra un teorema analogo. La prima dimostrazione completa del teorema di Gerschgorin è del 1946, ed è fatta da Brauer, un allievo di Schur, per cui in un primo tempo, nel 1948, Olga Taussky attribuisce il teorema di Gerschgorin a Brauer e solo successivamente a Gerschgorin. Fra il 1946 e il 1948 Brauer pubblica una raccolta sistematica di teoremi di limitazione degli autovalori, fra i quali quelli noti come secondo e terzo teorema di Gerschgorin e anche una generalizzazione del primo teorema di Gerschgorin in cui al posto dei cerchi vengono considerati degli ovali di Cassini (si veda l'esercizio 2.44).

Il libro più completo e dettagliato sugli autovalori resta ancora quello di Wilkinson [7]; si veda anche il libro di Householder [4]. Una esposizione più elementare è riportata nei testi di Atkinson [1], Isaacson, Keller [5] e di Stoer, Bulirsch [6]. Nei testi di Faddeev, Faddeeva [2] e Halmos [3] è riportata una esauriente trattazione della teoria degli operatori lineari, con la dimostrazione della forma normale di Jordan. Tutti questi testi riportano anche una estesa bibliografia.

Bibliografia

[1] K. E. Atkinson, *An Introduction to Numerical Analysis*, John Wiley and Sons, New York, 1978.

[2] D. K. Faddeev, V. N. Faddeeva, *Computational Methods of Linear Algebra*, Freeman and Co., San Francisco, 1963

[3] P. R. Halmos, *Finite-Dimentional Vector Spaces*, Van Nostrand-Reinhold, Princeton, 1958.

[4] A. S. Householder, *The Theory of Matrices in Numerical Analysis*, Blaisdell, Boston, 1964.

[5] E. Isaacson, H. B. Keller, *Analysis of Numerical Methods*, John Wiley and Sons, New York, 1966.

[6] J. Stoer, R. Bulirsch, *Introduction to Numerical Analysis*, Springer-Verlag, New York, 1980.

[7] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.

Capitolo 3

NORME

1. Norme vettoriali

In questo capitolo vengono introdotti i concetti di norma di un vettore e di norma di una matrice insieme con alcune delle proprietà che le caratterizzano. Il concetto di norma è una generalizzazione del concetto di lunghezza di un vettore $\mathbf{x} \in \mathbf{R}^n$, data dall'espressione

$$\sqrt{x_1^2 + \cdots + x_n^2}.$$

3.1 Definizione. Una funzione di $\mathbf{C}^n$ in $\mathbf{R}$

$$\mathbf{x} \to \|\mathbf{x}\|$$

che verifica le seguenti proprietà

a) $\|\mathbf{x}\| \geq 0$ e $\|\mathbf{x}\| = 0$ se e solo se $\mathbf{x} = \mathbf{0}$,

b) $\|\alpha\mathbf{x}\| = |\alpha| \, \|\mathbf{x}\|$ per ogni $\alpha \in \mathbf{C}$,

c) $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ per ogni $\mathbf{y} \in \mathbf{C}^n$,

è detta *norma vettoriale*. ∎

La proprietà c) corrisponde alla ben nota *disuguaglianza triangolare*, per la quale in un triangolo la somma delle lunghezze di due lati è maggiore od uguale alla lunghezza del terzo lato.

Per indicare in generale una norma si utilizza la notazione $\| \, . \, \|$, specificando con un indice se ci si riferisce ad una norma particolare. Si introducono alcune delle norme vettoriali comunemente usate.

3.2 Definizione. Sia $\mathbf{x} \in \mathbf{C}^n$; si definiscono:

$$\begin{aligned}
\|\mathbf{x}\|_1 &= \sum_{i=1}^{n} |x_i| && \text{norma 1} \\
\|\mathbf{x}\|_2 &= \sqrt{\mathbf{x}^H \mathbf{x}} = \sqrt{\sum_{i=1}^{n} |x_i|^2} && \text{norma 2} \\
\|\mathbf{x}\|_\infty &= \max_{i=1,\ldots,n} |x_i| && \text{norma } \infty
\end{aligned}$$

La norma 2 è quella che corrisponde alla lunghezza euclidea del vettore $\mathbf{x}$.∎

Si dimostra, come esempio, che la norma ∞ verifica le proprietà a), b) e c) della definizione 3.1:

a) poiché $|x_i| \geq 0$ per $i = 1, \ldots, n$, ne segue che $\max\limits_{i=1,\ldots,n} |x_i| \geq 0$ e quindi $\|\mathbf{x}\|_\infty \geq 0$; inoltre se $\max\limits_{i=1,\ldots,n} |x_i| = 0$, deve essere $|x_i| = 0$ per $i = 1, \ldots, n$, e viceversa, quindi $\|\mathbf{x}\|_\infty = 0$ se e solo se $\mathbf{x} = \mathbf{0}$;

b) $\|\alpha\mathbf{x}\|_\infty = \max\limits_{i=1,\ldots,n} |\alpha x_i| = \max\limits_{i=1,\ldots,n} |\alpha|\,|x_i| = |\alpha| \max\limits_{i=1,\ldots,n} |x_i| = |\alpha|\,\|\mathbf{x}\|_\infty$;

c) $\|\mathbf{x}+\mathbf{y}\|_\infty = \max\limits_{i=1,\ldots,n} |x_i + y_i| \leq \max\limits_{i=1,\ldots,n} (|x_i| + |y_i|)$
$\leq \max\limits_{i=1,\ldots,n} |x_i| + \max\limits_{i=1,\ldots,n} |y_i| = \|\mathbf{x}\|_\infty + \|\mathbf{y}\|_\infty$.

Per le altre due norme la dimostrazione è analoga, in particolare la proprietà c) per la $\|\,.\,\|_2$ viene dimostrata facendo uso della disuguaglianza di Cauchy-Schwarz ((1), cap.1).

Gli insiemi:

$$C_1 = \{\mathbf{x} \in \mathbf{R}^2 : \|\mathbf{x}\|_1 \leq 1\},$$
$$C_2 = \{\mathbf{x} \in \mathbf{R}^2 : \|\mathbf{x}\|_2 \leq 1\},$$
$$C_\infty = \{\mathbf{x} \in \mathbf{R}^2 : \|\mathbf{x}\|_\infty \leq 1\},$$

rappresentano in $\mathbf{R}^2$ i cerchi unitari rispetto alle norme 1, 2 e ∞ (vedere fig.3.1).

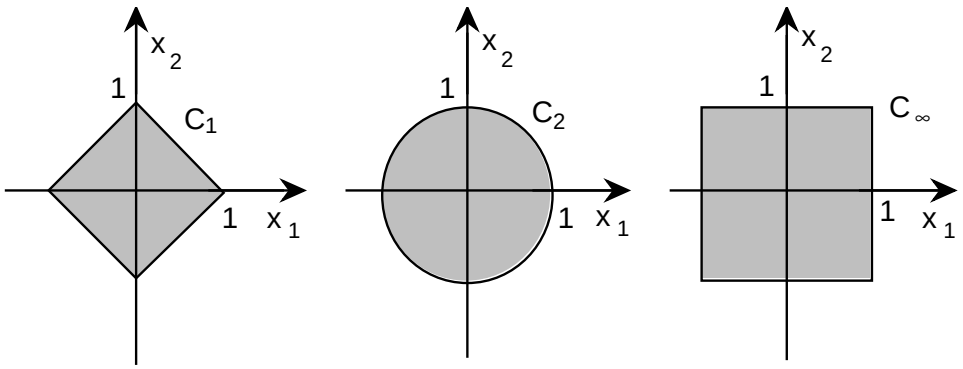


Fig. 3.1 - Cerchi unitari in $\mathbf{R}^2$ rispetto alle norme 1, 2 e ∞.

Alcune proprietà importanti delle norme vettoriali sono date nei seguenti teoremi.

3.3 Teorema. *La funzione* $\mathbf{x} \to \|\mathbf{x}\|$, $\mathbf{x} \in \mathbf{C}^n$, *è uniformemente continua.*

Dim. Siano $\mathbf{x}, \mathbf{y} \in \mathbf{C}^n$. Per la proprietà c) delle norme si ha:

$$\|\mathbf{x}\| = \|\mathbf{x} - \mathbf{y} + \mathbf{y}\| \leq \|\mathbf{x} - \mathbf{y}\| + \|\mathbf{y}\|,$$

da cui

$$\|\mathbf{x}\| - \|\mathbf{y}\| \leq \|\mathbf{x} - \mathbf{y}\|. \tag{1}$$

Inoltre

$$\|\mathbf{y}\| = \|\mathbf{x} + \mathbf{y} - \mathbf{x}\| \leq \|\mathbf{y} - \mathbf{x}\| + \|\mathbf{x}\| = \|\mathbf{x} - \mathbf{y}\| + \|\mathbf{x}\|,$$

da cui

$$-(\|\mathbf{x}\| - \|\mathbf{y}\|) \leq \|\mathbf{x} - \mathbf{y}\|. \tag{2}$$

Da (1) e (2) risulta

$$| \, \|\mathbf{x}\| - \|\mathbf{y}\| \, | \leq \|\mathbf{x} - \mathbf{y}\|. \tag{3}$$

Posto

$$\mathbf{x} - \mathbf{y} = \sum_{i=1}^{n} (x_i - y_i)\mathbf{e}_i,$$

dove $\mathbf{e}_i$ è l'i-esimo vettore della base canonica di $\mathbf{C}^n$, dalla (3) e dalle proprietà b) e c) si ha:

$$| \, \|\mathbf{x}\| - \|\mathbf{y}\| \, | \leq \sum_{i=1}^{n} |x_i - y_i| \, \|\mathbf{e}_i\| \leq \max_{i=1,\ldots,n} |x_i - y_i| \sum_{i=1}^{n} \|\mathbf{e}_i\|.$$

Poiché il numero $\alpha = \sum_{i=1}^{n} \|\mathbf{e}_i\|$ è diverso da zero e non dipende né da $\mathbf{x}$ né da $\mathbf{y}$, si ha che se

$$\max_{i=1,\ldots,n} |x_i - y_i| \leq \frac{\epsilon}{\alpha}, \qquad \text{risulta} \quad | \, \|\mathbf{x}\| - \|\mathbf{y}\| \, | \leq \epsilon. \qquad \blacksquare$$

Altre importanti proprietà sono date dai seguenti teoremi.

3.4 Teorema (di equivalenza delle norme). *Siano* $\|\,.\,\|'$ *e* $\|\,.\,\|''$ *due norme vettoriali. Allora le due norme sono topologicamente equivalenti, nel senso che esistono due costanti* α *e* $\beta \in \mathbf{R}$, $0 < \alpha \leq \beta$, *tali che per ogni* $\mathbf{x} \in \mathbf{C}^n$ *è*

$$\alpha\|\mathbf{x}\|'' \leq \|\mathbf{x}\|' \leq \beta\|\mathbf{x}\|''. \tag{4}$$

Dim. È sufficiente dimostrare la (4) nel caso in cui la norma $\|\,.\,\|''$ sia la norma ∞. Nel caso generale la (4) vale per confronto. Se $\mathbf{x} = \mathbf{0}$, la relazione è banalmente verificata. Se $\mathbf{x} \neq \mathbf{0}$, si considera l'insieme

$$S = \{\, \mathbf{y} \in \mathbf{C}^n \;:\; \|\mathbf{y}\|_\infty = 1 \,\},$$

che è chiuso e limitato perché è costituito dai vettori le cui componenti hanno modulo minore o uguale a 1, e almeno una componente ha modulo uguale a 1. Essendo $\|\,.\,\|'$ una funzione continua, essa assume su S massimo e minimo:

$$\alpha = \min_{\mathbf{y}\in S} \|\mathbf{y}\|' \qquad \text{e} \qquad \beta = \max_{\mathbf{y}\in S} \|\mathbf{y}\|', \qquad 0 < \alpha \leq \beta.$$

Poiché $\mathbf{y} \neq \mathbf{0}$, risulta che $\alpha \neq 0$, e quindi per ogni $\mathbf{y} \in S$ è

$$0 < \alpha \leq \|\mathbf{y}\|' \leq \beta. \tag{5}$$

Per ogni $\mathbf{x} \in \mathbf{C}^n, \mathbf{x} \neq \mathbf{0}$, si consideri il vettore

$$\mathbf{y} = \frac{\mathbf{x}}{\|\mathbf{x}\|_\infty};$$

si ha $\|\mathbf{y}\|_\infty = 1$ e quindi $\mathbf{y} \in S$ e

$$\|\mathbf{y}\|' = \left\| \frac{\mathbf{x}}{\|\mathbf{x}\|_\infty} \right\| = \frac{\|\mathbf{x}\|'}{\|\mathbf{x}\|_\infty},$$

per la proprietà b) delle norme vettoriali. Sostituendo nella (5), si ha:

$$\alpha \leq \frac{\|\mathbf{x}\|'}{\|\mathbf{x}\|_\infty} \leq \beta,$$

da cui

$$\alpha\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|' \leq \beta\|\mathbf{x}\|_\infty$$ ∎

Costanti α e β che verificano la (4), relative alle norme definite in 3.2, sono determinate nel seguente

3.5 Teorema. *Per ogni* $\mathbf{x} \in \mathbf{C}^n$ *si ha*

1. $\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_2 \leq \sqrt{n}\, \|\mathbf{x}\|_\infty;$
2. $\|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_1 \leq \sqrt{n}\, \|\mathbf{x}\|_2;$
3. $\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_1 \leq n\, \|\mathbf{x}\|_\infty.$

Dim. Per le disuguaglianze del punto 1, ci si può riferire direttamente alla dimostrazione del teorema 3.4; sia

$$S = \{\, \mathbf{x} \in \mathbf{C}^n \;:\; \|\mathbf{x}\|_\infty = 1 \,\}.$$

Su S la $\|\,.\,\|_2$ è minima per i vettori $\mathbf{x}$ che hanno una sola componente diversa da 0, in modulo uguale a 1, ed è massima per i vettori $\mathbf{x}$ che hanno tutte le componenti in modulo uguali a 1, cioè $|x_i| = 1,\ i = 1, \ldots, n$. Allora

$$\alpha = \min_{\mathbf{x}\in S} \|\mathbf{x}\|_2 = 1, \qquad \beta = \max_{\mathbf{x}\in S} \|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n} 1} = \sqrt{n}.$$

La prima disuguaglianza del punto 2 si ottiene notando che

$$\|\mathbf{x}\|_2^2 = \sum_{i=1}^{n} |x_i|^2 \le \sum_{i=1}^{n} |x_i|^2 + \sum_{\substack{i,j=1 \\ i\neq j}}^{n} |x_i|\,|x_j| = \left[\sum_{i=1}^{n} |x_i|\right]^2 = \|\mathbf{x}\|_1^2.$$

Per la seconda disuguaglianza del punto 2, si consideri il vettore $\mathbf{y}$ definito da

$$y_j = \begin{cases} \dfrac{|x_j|}{\overline{x}_j} & \text{se } x_j \neq 0, \\ 0 & \text{se } x_j = 0. \end{cases}$$

Allora per la disuguaglianza di Cauchy-Schwarz ((1), cap. 1), si ha:

$$|\mathbf{x}^H \mathbf{y}| \le \sqrt{\mathbf{x}^H \mathbf{x}}\ \sqrt{\mathbf{y}^H \mathbf{y}},$$

ed essendo

$$|\mathbf{x}^H \mathbf{y}| = \left|\sum_{j=1}^{n} \overline{x}_j\, y_j\right| = \sum_{j=1}^{n} |x_j| = \|\mathbf{x}\|_1 \quad \text{e} \quad \mathbf{y}^H \mathbf{y} \le n,$$

risulta $\|\mathbf{x}\|_1 \le \sqrt{n}\ \|\mathbf{x}\|_2$.

Le disuguaglianze del punto 3 si ottengono combinando fra loro quelle degli altri due punti. ∎

Si osservi che per una matrice unitaria A risulta

$$\|A\mathbf{x}\|_2 = \|\mathbf{x}\|_2 \quad \text{per ogni } \mathbf{x} \in \mathbf{C}^n, \tag{6}$$

essendo

$$\|A\mathbf{x}\|_2 = \sqrt{(A\mathbf{x})^H (A\mathbf{x})} = \sqrt{\mathbf{x}^H A^H A \mathbf{x}} = \sqrt{\mathbf{x}^H \mathbf{x}} = \|\mathbf{x}\|_2.$$

2. Norme matriciali

Il concetto di norma può essere esteso al caso delle matrici.

3.6 Definizione. Una funzione di $\mathbf{C}^{n\times n}$ in $\mathbf{R}$

$$A \to \|A\|$$

che verifica le seguenti proprietà

a) $\|A\| \geq 0$ e $\|A\| = 0$ se e solo se $A = O$,

b) $\|\alpha A\| = |\alpha|\ \|A\|$ per ogni $\alpha \in \mathbf{C}$,

c) $\|A + B\| \leq \|A\| + \|B\|$ per ogni $B \in \mathbf{C}^{n\times n}$,

d) $\|AB\| \leq \|A\|\ \|B\|$ per ogni $B \in \mathbf{C}^{n\times n}$,

è detta *norma matriciale*. ∎

Anche per le norme matriciali si utilizza la stessa notazione usata per le norme vettoriali. Poiché le proprietà a), b) e c) delle norme matriciali coincidono con quelle delle norme vettoriali, ne segue che anche le norme matriciali sono funzioni uniformemente continue e anche per esse vale un teorema di equivalenza analogo al 3.4.

Si mostra ora come sia possibile associare ad una norma vettoriale una corrisponedente norma matriciale. Si osservi che, poiché la norma vettoriale è una funzione continua, l'insieme

$$\{\, \mathbf{x} \in \mathbf{C}^n \ :\ \|\mathbf{x}\| = 1 \,\}$$

è chiuso; inoltre, poiché per il teorema 3.4 esiste α tale che $\|\mathbf{x}\|_\infty \leq \alpha\|\mathbf{x}\|$, ossia $\max\limits_{i=1,\ldots,n} |x_i| \leq \alpha$, l'insieme è anche limitato. Poiché una funzione continua assume su un sottoinsieme chiuso e limitato di $\mathbf{C}^n$ massimo e minimo, si ha che esiste

$$\max_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|.$$

3.7 Teorema. *Sia* $\|\ .\ \|$ *una norma vettoriale. La funzione*

$$A \to \max_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|, \quad A \in \mathbf{C}^{n\times n}, \quad \mathbf{x} \in \mathbf{C}^n$$

è una norma matriciale.

Dim. Si verificano le proprietà a), b), c), d), della definizione 3.6.

a) $\|A\mathbf{x}\| \geq 0$ e quindi $\max\limits_{\|\mathbf{x}\|=1} \|A\mathbf{x}\| \geq 0$. Inoltre se $A = O$, allora $\|A\mathbf{x}\| = 0$ per ogni $\mathbf{x}$; viceversa se $\max\limits_{\|\mathbf{x}\|=1} \|A\mathbf{x}\| = 0$, allora $\|A\mathbf{x}\| = 0$ per ogni $\mathbf{x}$ tale che $\|\mathbf{x}\| = 1$ e quindi $A = O$.

b) $\max\limits_{\|\mathbf{x}\|=1} \|\alpha A\mathbf{x}\| = \max\limits_{\|\mathbf{x}\|=1} |\alpha| \, \|A\mathbf{x}\| = |\alpha| \max\limits_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|$.

c) $\max\limits_{\|\mathbf{x}\|=1} \|(A+B)\mathbf{x}\| = \max\limits_{\|\mathbf{x}\|=1} \|A\mathbf{x} + B\mathbf{x}\| \leq \max\limits_{\|\mathbf{x}\|=1} (\|A\mathbf{x}\| + \|B\mathbf{x}\|)$
$\leq \max\limits_{\|\mathbf{x}\|=1} \|A\mathbf{x}\| + \max\limits_{\|\mathbf{x}\|=1} \|B\mathbf{x}\|$.

d) Se $AB = O$, allora $\|AB\| = 0$; quindi la disuguaglianza è verificata. Se $AB \neq O$, allora esiste un vettore $\mathbf{y}$, con $\|\mathbf{y}\| = 1$, tale che

$$\|AB\mathbf{y}\| = \max_{\|\mathbf{x}\|=1} \|AB\mathbf{x}\| \neq 0.$$

Posto $\mathbf{z} = B\mathbf{y}$, risulta $\mathbf{z} \neq \mathbf{0}$ (infatti se fosse $\mathbf{z} = \mathbf{0}$, sarebbe $(AB)\mathbf{y} = \mathbf{0}$ e quindi $AB = O$), si ha

$$\|(AB)\mathbf{y}\| = \|A(B\mathbf{y})\| = \|A\mathbf{z}\| = \|\mathbf{z}\| \frac{\|A\mathbf{z}\|}{\|\mathbf{z}\|} = \|B\mathbf{y}\| \left\| A \frac{\mathbf{z}}{\|\mathbf{z}\|} \right\|.$$

Poiché il vettore $\mathbf{u} = \dfrac{\mathbf{z}}{\|\mathbf{z}\|}$ è tale che $\|\mathbf{u}\| = 1$, risulta

$$\max_{\|\mathbf{x}\|=1} \|AB\mathbf{x}\| = \|A\mathbf{u}\| \, \|B\mathbf{y}\| \leq \max_{\|\mathbf{v}\|=1} \|A\mathbf{v}\| \max_{\|\mathbf{w}\|=1} \|B\mathbf{w}\|. \qquad \blacksquare$$

3.8 Definizione. La norma definita da

$$\|A\| = \max_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|,$$

viene detta *norma matriciale indotta* dalla norma vettoriale $\| \, . \, \|$. $\blacksquare$

3.9 Teorema. *Dalle tre norme vettoriali definite in 3.2, si ottengono le corrispondenti norme matriciali indotte*

$$\|A\|_1 = \max_{j=1,\ldots,n} \sum_{i=1}^{n} |a_{ij}| \qquad \textit{norma 1}$$

$$\|A\|_2 = \sqrt{\rho(A^H A)} \qquad \textit{norma 2}$$

$$\|A\|_\infty = \max_{i=1,\ldots,n} \sum_{j=1}^{n} |a_{ij}| \qquad \textit{norma } \infty$$

Dim. Norma 1 - Sia $\mathbf{x} \in \mathbf{C}^n$, tale che $\|\mathbf{x}\|_1 = 1$. Allora

$$\|A\mathbf{x}\|_1 = \sum_{i=1}^{n} \left| \sum_{j=1}^{n} a_{ij} x_j \right| \leq \sum_{i=1}^{n} \sum_{j=1}^{n} |a_{ij}||x_j| = \sum_{j=1}^{n} |x_j| \sum_{i=1}^{n} |a_{ij}|$$
$$\leq \left[\max_{j=1,\ldots,n} \sum_{i=1}^{n} |a_{ij}| \right] \sum_{j=1}^{n} |x_j| = \max_{j=1,\ldots,n} \sum_{i=1}^{n} |a_{ij}|,$$

e quindi

$$\max_{\|\mathbf{x}\|_1=1} \|A\mathbf{x}\|_1 \leq \max_{j=1,\ldots,n} \sum_{i=1}^{n} |a_{ij}|.$$

Si tratta ora di verificare che esiste un vettore $\mathbf{x}$, $\|\mathbf{x}\|_1 = 1$, per cui

$$\|A\mathbf{x}\|_1 = \max_{j=1,\ldots,n} \sum_{i=1}^{n} |a_{ij}|.$$

Questo vettore esiste in quanto, se k è l'indice della colonna di A in cui la somma dei moduli degli elementi è massima, cioè

$$\sum_{i=1}^{n} |a_{ik}| = \max_{j=1,\ldots,n} \sum_{i=1}^{n} |a_{ij}|,$$

il vettore $\mathbf{x} = \mathbf{e}_k$ è tale che $\|\mathbf{x}\|_1 = 1$ e

$$\|A\mathbf{x}\|_1 = \|A\mathbf{e}_k\|_1 = \|(a_{1k}, \ldots, a_{nk})^T\|_1 = \sum_{i=1}^{n} |a_{ik}|.$$

Norma 2 - Poiché la matrice $A^H A$ è hermitiana, per il teorema 2.26 risulta

$$A^H A = U D U^H,$$

dove U è unitaria e D diagonale con gli autovalori di $A^H A$ come elementi principali. Se $A = O$, allora $\rho(A^H A) = 0$, e inversamente, se $\rho(A^H A) = 0$, risulta $D = O$ e quindi $A = O$. Se $A \neq O$, si ha

$$\mathbf{x}^H A^H A \mathbf{x} \geq 0 \quad \text{per } \mathbf{x} \neq \mathbf{0}.$$

Procedendo in modo analogo a quanto fatto nella dimostrazione del teorema 2.31, risulta che gli autovalori di $A^H A$ sono non negativi e per almeno uno di essi, corrispondente al raggio spettrale di $A^H A$, si ha

$$\lambda_1 = \rho(A^H A) > 0.$$

Sia $\mathbf{x}$ tale che $\|\mathbf{x}\|_2 = 1$ e $\mathbf{y} = U^H\mathbf{x}$; poiché U è unitaria, per la (6) risulta $\|\mathbf{y}\|_2 = 1$ e quindi

$$\begin{aligned}\max_{\|\mathbf{x}\|_2=1} \|A\mathbf{x}\|_2 &= \max_{\|\mathbf{x}\|_2=1} \sqrt{\mathbf{x}^H A^H A\mathbf{x}} = \max_{\|\mathbf{y}\|_2=1} \sqrt{\mathbf{y}^H D\mathbf{y}} \\ &= \max_{\|\mathbf{y}\|_2=1} \sqrt{\sum_{i=1}^{n} \lambda_i |y_i|^2} \leq \max_{\|\mathbf{y}\|_2=1} \sqrt{\lambda_1 \sum_{i=1}^{n} |y_i|^2} \\ &= \sqrt{\lambda_1} = \sqrt{\rho(A^H A)}.\end{aligned}$$

Si tratta ora di verificare che esiste un vettore $\mathbf{x}$, $\|\mathbf{x}\|_2 = 1$, per cui

$$\|A\mathbf{x}\|_2 = \sqrt{\rho(A^H A)}.$$

Questo vettore è $\mathbf{x}_1$, autovettore di $A^H A$ relativo all'autovalore λ_1 normalizzato in modo che $\|\mathbf{x}_1\|_2 = 1$. Infatti risulta:

$$\mathbf{x}_1^H A^H A\mathbf{x}_1 = \lambda_1 \mathbf{x}_1^H \mathbf{x}_1 = \lambda_1 = \rho(A^H A).$$

Norma ∞ - Procedendo in modo analogo a quanto fatto per la norma 1, risulta

$$\max_{\|\mathbf{x}\|_\infty=1} \|A\mathbf{x}\|_\infty \leq \max_{i=1,\ldots,n} \sum_{j=1}^{n} |a_{ij}|.$$

Si tratta ora di verificare che esiste un vettore $\mathbf{x}, \|\mathbf{x}\|_\infty = 1$, per cui

$$\|A\mathbf{x}\|_\infty = \max_{i=1,\ldots,n} \sum_{j=1}^{n} |a_{ij}|.$$

Se $A = O$ basta scegliere $\mathbf{x} = \mathbf{e}_1$, se $A \neq O$ il vettore $\mathbf{x}$ è dato da

$$x_j = \begin{cases} \dfrac{|a_{kj}|}{a_{kj}} & \text{se } a_{kj} \neq 0 \\ 0 & \text{altrimenti,} \end{cases}$$

dove k è l'indice della riga di A in cui la somma dei moduli degli elementi è massima. ∎

Se A è una matrice hermitiana, risulta

$$\begin{aligned}&\|A\|_1 = \|A\|_\infty \\ &\|A\|_2 = \sqrt{\rho(A^H A)} = \sqrt{\rho(A^2)} = \sqrt{\rho^2(A)} = \rho(A),\end{aligned}$$

e se A è anche definita positiva risulta

$$\|A\|_2 = \lambda_{\max},$$

dove $\lambda_{\max}$ è il massimo degli autovalori di A.

Un'altra norma frequentemente usata, anche per la sua semplicità di calcolo, è quella così definita

$$\|A\|_F = \sqrt{\sum_{i,j=1}^{n} |a_{ij}|^2} = \sqrt{\mathrm{tr}(A^H A)}, \tag{7}$$

che è detta norma di *Frobenius (o di Schur)* di A.

La (7) verifica le proprietà a), b) e c) della definizione 3.6, in quanto gli elementi di A si possono pensare disposti come elementi di un vettore $\mathbf{a} \in \mathbf{C}^m$, con $m = n^2$, per cui $\|A\|_F = \|\mathbf{a}\|_2$. Per quanto riguarda la proprietà d), sia $C = AB$, ossia

$$c_{ij} = \mathbf{a}_i^T \mathbf{b}_j = \overline{\mathbf{a}}_i^H \mathbf{b}_j,$$

dove $\mathbf{a}_i^T \in \mathbf{C}^{1\times n}$ è la i-esima riga di A e $\mathbf{b}_j \in \mathbf{C}^n$ è la j-esima colonna di B. Per la disuguaglianza di Cauchy-Schwarz ((1), cap. 1) si ha

$$|c_{ij}|^2 \leq (\overline{\mathbf{a}}_i^H \overline{\mathbf{a}}_i)\ (\mathbf{b}_j^H \mathbf{b}_j) = (\mathbf{a}_i^H \mathbf{a}_i)\ (\mathbf{b}_j^H \mathbf{b}_j),$$

da cui

$$\|C\|_F^2 = \sum_{i,j=1}^{n} |c_{ij}|^2 \leq \sum_{i=1}^{n} \mathbf{a}_i^H \mathbf{a}_i \sum_{j=1}^{n} \mathbf{b}_j^H \mathbf{b}_j = \|A\|_F^2 \|B\|_F^2.$$

Sia $U \in \mathbf{C}^{n\times n}$ una matrice unitaria. Poiché $(UA)^H UA = A^H A$, risulta

$$\|A\|_2 = \|UA\|_2 \quad \text{e} \quad \|A\|_F = \|UA\|_F,$$

poiché $A^H A$ e $(AU)^H AU$ sono matrici simili, risulta

$$\|A\|_2 = \|AU\|_2 \quad \text{e} \quad \|A\|_F = \|AU\|_F,$$

e poiché anche $A^H A$ e $(UAU^H)^H UAU^H = UA^H AU^H$ sono matrici simili, risulta

$$\|A\|_2 = \|UAU^H\|_2 \quad \text{e} \quad \|A\|_F = \|UAU^H\|_F.$$

3. Alcune proprietà delle norme

– Siano $A \in \mathbf{C}^{n\times n}$ e $\mathbf{x} \in \mathbf{C}^n$. Per la norma matriciale $\|\ .\ \|$ indotta dalla norma vettoriale $\|\ .\ \|$ risulta

$$\|A\mathbf{x}\| \leq \|A\|\ \|\mathbf{x}\|.$$

Infatti, se $\mathbf{x} = \mathbf{0}$, la relazione è ovvia; se $\mathbf{x} \neq \mathbf{0}$, si ha

$$\|A\mathbf{x}\| = \|\mathbf{x}\|\ \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} = \|\mathbf{x}\|\ \|A\mathbf{y}\|,$$

dove il vettore $\mathbf{y} = \dfrac{\mathbf{x}}{\|\mathbf{x}\|}$ è tale che $\|\mathbf{y}\| = 1$ e quindi

$$\|A\mathbf{y}\| \leq \max_{\|\mathbf{z}\|=1} \|A\mathbf{z}\| = \|A\|.$$

– Poiché $\|AB\| \leq \|A\|\|B\|$, per ogni norma matriciale, risulta

$$\|A^m\| \leq \|A\|^m \qquad \text{per ogni intero } m \text{ positivo.}$$

– Poiché $\|I\| = \|I\,I\| \leq \|I\|\ \|I\|$, risulta $\|I\| \geq 1$ per ogni norma matriciale.

– Se $\|\ .\ \|$ è una norma matriciale indotta, allora risulta $\|I\| = 1$. Infatti per definizione di norma indotta si ha:

$$\|I\| = \max_{\|\mathbf{x}\|=1} \|I\mathbf{x}\| = \max_{\|\mathbf{x}\|=1} \|\mathbf{x}\| = 1.$$

Per questo si osservi che la norma di Frobenius non può essere una norma indotta in quanto

$$\|I\|_F = \sqrt{n} \neq 1 \quad \text{per } n > 1.$$

– Se $A \in \mathbf{C}^{n\times n}$ è non singolare, allora, poiché

$$\|I\| = \|A^{-1}A\| \leq \|A^{-1}\|\ \|A\|,$$

per ogni norma matriciale risulta:

$$\|A^{-1}\| \geq \frac{1}{\|A\|}.$$

3.10 Teorema. *Per ogni norma* $\| \, . \, \|$ *matriciale indotta vale*

$$\rho(A) \leq \|A\|.$$

Dim. Siano λ un autovalore e $\mathbf{x}$ il corrispondente autovettore normalizzato rispetto alla norma $\| \, . \, \|$:

$$A\mathbf{x} = \lambda\mathbf{x}, \qquad \|\mathbf{x}\| = 1.$$

Allora

$$|\lambda| = \|A\mathbf{x}\|,$$

da cui segue che

$$|\lambda| \leq \max_{\|\mathbf{v}\|=1} \|A\mathbf{v}\| = \|A\|.$$

Questa relazione vale per ogni autovalore λ di A e quindi anche per quello di modulo massimo. ∎

3.11 Teorema. *La funzione*

$$A \to \|S^{-1}AS\|_\infty,$$

dove S *è una matrice non singolare, è una norma matriciale indotta.*

Dim. La funzione

$$\mathbf{x} \to \|S^{-1}\mathbf{x}\|_\infty, \tag{8}$$

poiché S^{-1} è una matrice non singolare, verifica le proprietà a), b) e c) della definizione 3.1, e quindi è una norma vettoriale. La norma matriciale indotta dalla (8) è data da

$$\|A\| = \max_{\|S^{-1}\mathbf{x}\|_\infty=1} \|S^{-1}A\mathbf{x}\|_\infty = \max_{\|\mathbf{y}\|_\infty=1} \|S^{-1}AS\mathbf{y}\|_\infty,$$

avendo posto $\mathbf{y} = S^{-1}\mathbf{x}$. ∎

3.12 Teorema. *Sia* $A \in \mathbf{C}^{n\times n}$*; allora per ogni* $\epsilon > 0$ *esiste una norma matriciale indotta* $\| \, . \, \|$ *tale che*

$$\|A\| \leq \rho(A) + \epsilon.$$

Dim. Sia J la forma canonica di Jordan di A (si veda il teorema 2.18):

$$A = TJT^{-1},$$

in cui J è una matrice diagonale a blocchi e ciascun blocco è della forma

$$C_i^{(j)} = \begin{bmatrix} \lambda_i & 1 & & \\ & \ddots & \ddots & \\ & & \lambda_i & 1 \\ & & & \lambda_i \end{bmatrix}.$$

dove λ_i è un autovalore di A. Data la matrice

$$E = \begin{bmatrix} 1 & & & & \\ & \epsilon & & & \\ & & \epsilon^2 & & \\ & & & \ddots & \\ & & & & \epsilon^{n-1} \end{bmatrix},$$

risulta che la matrice

$$E^{-1}JE = E^{-1}T^{-1}ATE$$

è ancora una matrice diagonale a blocchi e ciascun blocco è della forma

$$D_i^{(j)} = \begin{bmatrix} \lambda_i & \epsilon & & \\ & \ddots & \ddots & \\ & & \lambda_i & \epsilon \\ & & & \lambda_i \end{bmatrix}.$$

Si ha:

$$\|E^{-1}JE\|_\infty = \|E^{-1}T^{-1}ATE\|_\infty = \max_{i,j} \|D_i^{(j)}\|_\infty \leq \rho(A) + \epsilon,$$

dove la disuguaglianza vale in senso stretto nel caso in cui i blocchi $D_i^{(j)}$ relativi agli autovalori di modulo massimo siano di ordine 1, ed ϵ sia abbastanza piccolo. Per il teorema 3.11, $\|E^{-1}T^{-1}ATE\|_\infty$ è una norma matriciale indotta di A. ∎

Si osservi che se gli autovalori corrispondenti a $\rho(A)$ hanno molteplicità algebrica uguale a quella geometrica, esiste una norma matriciale indotta $\| \cdot \|$ tale che

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

In particolare questo accade se la matrice A è diagonalizzabile.

3.13 Teorema. *Sia $\| \cdot \|$ una norma matriciale indotta e sia $A \in \mathbf{C}^{n\times n}$, tale che $\|A\| < 1$. Allora la matrice $I + A$ è non singolare e vale la disuguaglianza*

$$\|(I + A)^{-1}\| \leq \frac{1}{1 - \|A\|}.$$

Dim. Essendo $\|A\| < 1$, per il teorema 3.10 risulta $\rho(A) < 1$. Quindi la matrice $I + A$ non può avere autovalori nulli, ed è non singolare. Dalla relazione

$$(I + A)\,(I + A)^{-1} = I$$

segue che

$$(I + A)^{-1} = I - A(I + A)^{-1},$$

e poiché $\|I\| = 1$, per le proprietà c) e d) delle norme si ha:

$$\|(I + A)^{-1}\| \leq 1 + \|A\|\,\|(I + A)^{-1}\|,$$

e quindi

$$(1 - \|A\|)\,\|(I + A)^{-1}\| \leq 1,$$

da cui, essendo $\|A\| < 1$, segue la tesi. ∎

4. Principali relazioni fra le norme matriciali

Per le norme che sono state introdotte valgono le seguenti relazioni, che possono essere dimostrate usando il teorema 3.5 e la definizione di norma indotta:

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\,\|A\|_\infty \leq \|A\|_2 \leq \sqrt{n}\,\|A\|_\infty,\\
&\frac{1}{\sqrt{n}}\,\|A\|_1 \leq \|A\|_2 \leq \sqrt{n}\,\|A\|_1,\\
&\max_{i,j}|a_{ij}| \leq \|A\|_2 \leq n \max_{i,j}|a_{ij}|,\\
&\|A\|_2 \leq \sqrt{\|A\|_1\,\|A\|_\infty}
\end{aligned}$$

Ad esempio, l'ultima relazione si dimostra nel modo seguente: poiché gli autovalori della matrice semidefinita positiva $A^H A$ sono non negativi, l'autovalore massimo $\lambda_{\max}$ di $A^H A$ coincide con $\rho(A^H A)$, e dalla relazione

$$A^H A\mathbf{x} = \lambda_{\max}\,\mathbf{x},$$

passando alle norme, segue che

$$\begin{aligned}
\rho(A^H A)\,\|\mathbf{x}\|_\infty &= \lambda_{\max}\,\|\mathbf{x}\|_\infty = \|A^H A\mathbf{x}\|_\infty\\
&\leq \|A^H\|_\infty\,\|A\|_\infty\,\|\mathbf{x}\|_\infty = \|A\|_1\,\|A\|_\infty\,\|\mathbf{x}\|_\infty,
\end{aligned}$$

da cui

$$\|A\|_2^2 \leq \|A\|_1\,\|A\|_\infty.$$

Inoltre si ha

$$\rho(A^H A) = \lambda_{\max} \leq \sum_{i=1}^{n} \lambda_i = \mathrm{tr}(A^H A) \leq \sum_{i=1}^{n} \rho(A^H A) = n\,\rho(A^H A),$$

da cui si ricava la seguente relazione di equivalenza fra la norma 2 e la norma di Frobenius:

$$\|A\|_2 \leq \|A\|_F \leq \sqrt{n}\,\|A\|_2.$$

Esercizi proposti

3.1 Si determinino dei vettori di $\mathbf{C}^n$ tali che

(1) $\|\mathbf{x}\|_1 = \|\mathbf{x}\|_2 = \|\mathbf{x}\|_\infty$

(2) $\|\mathbf{x}\|_1 = \sqrt{n}\,\|\mathbf{x}\|_2 = n\,\|\mathbf{x}\|_\infty$.

3.2 Siano $\mathbf{x}$ e $\mathbf{y} \in \mathbf{C}^n$ due vettori ortogonali. Si dimostri che vale la seguente relazione

$$\|\mathbf{x} \pm \mathbf{y}\|_2^2 = \|\mathbf{x}\|_2^2 + \|\mathbf{y}\|_2^2$$

(se $\mathbf{x}$ e $\mathbf{y} \in \mathbf{R}^2$ la relazione esprime il *teorema di Pitagora*).

3.3 Sia $\|\ .\ \|$ una norma vettoriale e sia S la *sfera unitaria* rispetto a tale norma, cioè

$$S = \{\ \mathbf{x} \in \mathbf{C}^n \quad \text{tali che } \|\mathbf{x}\| \leq 1\ \}.$$

Si dimostri che S è un insieme convesso.

(Traccia: siano $\mathbf{x}$ e $\mathbf{y} \in S$; per ogni $\alpha \in [0,1]$ risulta

$$\|\alpha\mathbf{x} + (1-\alpha)\ \mathbf{y}\| \leq \alpha\|\mathbf{x}\| + (1-\alpha)\ \|\mathbf{y}\| \leq 1.)$$

3.4 Sia $p > 1$ e $q = \dfrac{p}{p-1}$. Si dimostri che

a) per ogni $\alpha,\ \beta > 0$ risulta

$$\alpha\beta \leq \frac{\alpha^p}{p} + \frac{\beta^q}{q};$$

b) se $\mathbf{x}, \mathbf{y} \in \mathbf{C}^n$, posto

$$f_p(\mathbf{x}) = \sqrt[p]{\sum_{i=1}^{n} |x_i|^p} \quad \text{e} \quad f_q(\mathbf{y}) = \sqrt[q]{\sum_{i=1}^{n} |y_i|^q},$$

valgono le seguenti disuguaglianze

$$\sum_{i=1}^{n} |x_i||y_i| \leq f_p(\mathbf{x}) f_q(\mathbf{y}), \qquad (\textit{disuguaglianza di Hölder})$$

$$\sqrt[p]{\sum_{i=1}^{n} (|x_i| + |y_i|)^p} \leq f_p(\mathbf{x}) + f_p(\mathbf{y}); \qquad (\textit{disuguaglianza di Minkowski})$$

c) la funzione

$$\mathbf{x} \ \rightarrow \ f_p(\mathbf{x}) = \sqrt[p]{\sum_{i=1}^{n} |x_i|^p}$$

è una norma vettoriale per $p \geq 1$. La norma così definita è chiamata norma *hölderiana* e viene indicata con il simbolo $\| \ . \ \|_p$. Per $p = 1$ e $p = 2$ si ottengono le norme $\| \ . \ \|_1$ e $\| \ . \ \|_2$.

d) $$\lim_{p \to \infty} \sqrt[p]{\sum_{i=1}^{n} |x_i|^p} = \max_{i=1,\ldots,n} |x_i|,$$

ciò che giustifica la notazione della $\| \ . \ \|_\infty$.

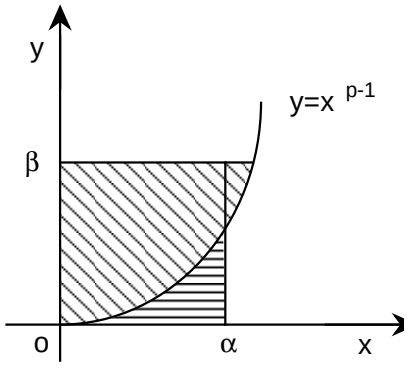
e) Si dica per quali valori di p vale la disuguaglianza

$$|\mathbf{x}^H \mathbf{y}| \leq \|\mathbf{x}\|_p \ \|\mathbf{y}\|_p$$

(per p=2 la disuguaglianza è detta di *Cauchy-Schwarz*). La disuguaglianza vale per la norma ∞?

f) Si dica che cosa accade se $p < 1$. Si esamini ad esempio il caso $p = 0.5$.

(Traccia: a) dal disegno (per il caso particolare $\beta \geq \alpha^{p-1}$, se $\beta < \alpha^{p-1}$ il disegno è analogo)



risulta che l'area del rettangolo di lati α e β è minore o uguale alla somma delle aree tratteggiate, date dagli integrali

$$\int_0^\alpha x^{p-1}dx = \frac{\alpha^p}{p}$$

(con tratteggio orizzontale) e

$$\int_0^\beta y^{\frac{1}{p-1}}dy = \frac{\beta^q}{q}$$

(con tratteggio obliquo). b) Indicate con

$$\alpha = \frac{|x_i|}{f_p(\mathbf{x})} \quad \text{e} \quad \beta = \frac{|y_i|}{f_q(\mathbf{y})} \ ,$$

per il punto a) si ha

$$\frac{|x_i|\ |y_i|}{f_p(\mathbf{x})f_q(\mathbf{y})} \leq \frac{|x_i|^p}{p\big[f_p(\mathbf{x})\big]^p} + \frac{|y_i|^q}{q\big[f_q(\mathbf{y})\big]^q} \ .$$

Sommando per $i = 1, \ldots, n$, si ha

$$\sum_{i=1}^n |x_i|\ |y_i| \leq (\frac{1}{p} + \frac{1}{q}) f_p(\mathbf{x}) f_q(\mathbf{y}),$$

da cui si ottiene la disuguaglianza di Hölder. Per l'altra disuguaglianza si ha

$$(|x_i| + |y_i|)^p = |x_i|(|x_i| + |y_i|)^{p-1} + |y_i|(|x_i| + |y_i|)^{p-1} = |x_i||z_i| + |y_i||z_i|,$$

dove $|z_i| = (|x_i| + |y_i|)^{p-1}$. Sommando e applicando la disuguaglianza di Hölder, si ha

$$\begin{aligned}\sum_{i=1}^n (|x_i| + |y_i|)^p &= \sum_{i=1}^n |x_i||z_i| + \sum_{i=1}^n |y_i||z_i| \\ &\leq f_p(\mathbf{x}) f_q(\mathbf{z}) + f_p(\mathbf{y}) f_q(\mathbf{z}) = \big[f_p(\mathbf{x}) + f_p(\mathbf{y})\big] f_q(\mathbf{z}),\end{aligned}$$

e poiché

$$f_q(\mathbf{z}) = \left[\sum_{i=1}^n |z_i|^q\right]^{\frac{1}{q}} = \left[\sum_{i=1}^n (|x_i| + |y_i|)^{(p-1)q}\right]^{\frac{1}{q}} = \left[\sum_{i=1}^n (|x_i| + |y_i|)^p\right]^{\frac{1}{q}},$$

la disuguaglianza di Minkowski segue dal fatto che

$$\sum_{i=1}^{n}(|x_i|+|y_i|)^p\left[f_q(\mathbf{z})\right]^{-1} = \left[\sum_{i=1}^{n}(|x_i|+|y_i|)^p\right]^{1-\frac{1}{q}} = \left[\sum_{i=1}^{n}(|x_i|+|y_i|)^p\right]^{\frac{1}{p}}.$$

c) le prime due proprietà della definizione 3.1 sono banalmente verificate, la terza discende dalla disuguaglianza di Minkowski; d) si noti che

$$\left[\max_{i=1,\ldots,n}|x_i|\right]^p \leq \sum_{i=1}^{n}|x_i|^p \leq n\left[\max_{i=1,\ldots,n}|x_i|\right]^p,$$

da cui si ha

$$\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_p \leq \sqrt[p]{n}\,\|\mathbf{x}\|_\infty,$$

e si faccia tendere $p \to \infty$. e) Per $p = 1$ la disuguaglianza vale e si dimostra per verifica diretta, per $p > 1$ dalla disuguaglianza di Hölder si ha

$$|\mathbf{x}^H\mathbf{y}| \leq \sum_{i=1}^{n}|x_i|\,|y_i| \leq f_p(\mathbf{x})f_q(\mathbf{y})$$

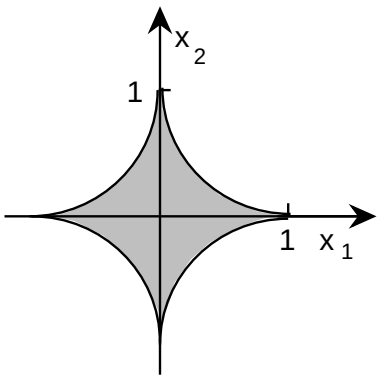
e $f_q(\mathbf{y}) \leq f_p(\mathbf{y})$ per $q \geq p$, cioè per $1 < p \leq 2$. Quindi la disuguaglianza in e) vale per $p \leq 2$. Non vale per $p > 2$, né per la norma ∞, come si può verificare per i vettori

$$\mathbf{x} = \mathbf{y} = [1,\ 1,\ldots,1]^T,$$

per cui è $\mathbf{x}^T\mathbf{y} = n$ e $\|\mathbf{x}\|_p = \|\mathbf{y}\|_p = \sqrt[p]{n}$. f) Per $p < 1$ la disuguaglianza di Minkowski non è verificata e quindi non vale la proprietà c) della definizione 3.1. Per $p = 0.5$ l'insieme

$$\{\ \mathbf{x} \in \mathbf{R}^2 \text{ tali che } \left[\sum_{i=1}^{2}\sqrt{|x_i|}\right]^2 \leq 1\ \}$$

non è convesso, come si vede dalla figura



3.5 Si dimostri che le seguenti funzioni da $\mathbf{C}^n$ in $\mathbf{R}$

(1) $\|\mathbf{x}\| = |x_1 - x_2| + \|\mathbf{x}\|_\infty,$

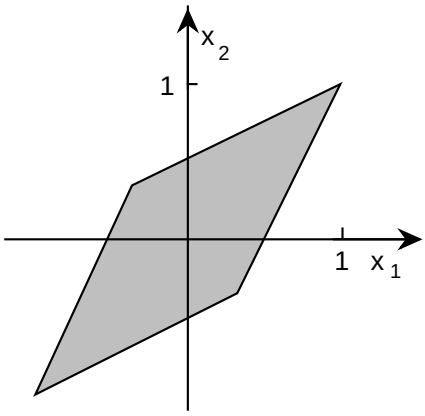
(2) $\|\mathbf{x}\| = |x_1 - x_2| + \|\mathbf{x}\|_1,$

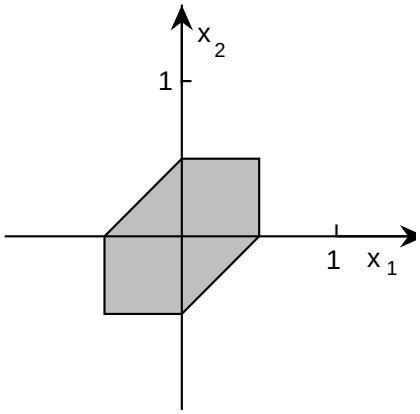
(3) $\|\mathbf{x}\| = \max\{\ \|\mathbf{x}\|_\infty, p\|\mathbf{x}\|_1\ \}, \quad p > 0,$

sono norme e si disegni per ciascuna di esse il cerchio unitario nel piano cartesiano

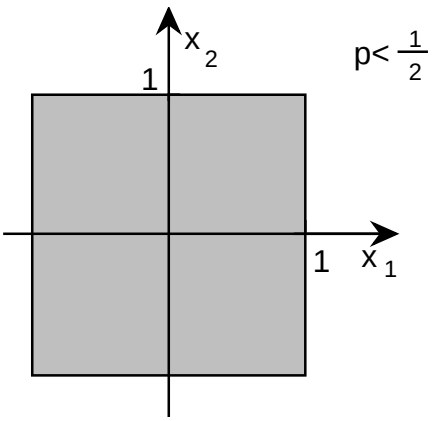
$$S = \{\ \mathbf{x} \in \mathbf{R}^2 \text{ tali che } \|\mathbf{x}\| \leq 1\ \}.$$

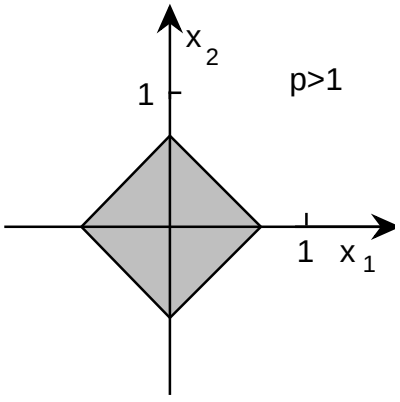
(Risposta: cerchi unitari per le norme (1) e (2)

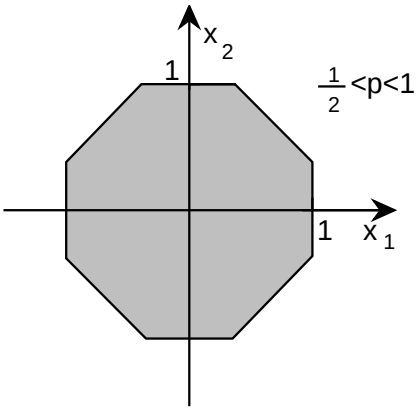




cerchi unitari per la norma (3)







3.6 Sia $A \in \mathbf{C}^{n\times n}$ una matrice definita positiva. Si dimostri che

$$\|\mathbf{x}\|_A = \sqrt{\mathbf{x}^H A\mathbf{x}}$$

è una norma (si vedano gli esercizi 1.26 e 1.27).

3.7 Si calcoli la norma ∞ delle matrici degli esercizi 1.49-1.54.

3.8 Si determinino delle matrici $A \in \mathbf{C}^{n\times n}$ tali che

(1) $\|A\|_2 = \dfrac{1}{\sqrt{n}}\,\|A\|_1 = \dfrac{1}{\sqrt{n}}\,\|A\|_\infty,$

(2) $\|A\|_2 = \sqrt{n}\,\|A\|_1 = \sqrt{n}\,\|A\|_\infty.$

3.9 Si determinino due matrici A e $B \in \mathbf{C}^{n\times n}$ tali che

$$\rho(A+B) > \rho(A) + \rho(B).$$

Questo dimostra che se A non è hermitiana $\rho(A)$ non può essere una norma matriciale.

3.10 Sia $A \in \mathbf{R}^{n\times n}$ la matrice

$$a_{ij} = \frac{1}{3^{|i-j|}}\ , \quad i,j = 1,\ldots,n.$$

Si verifichi che A è non singolare e si determini una maggiorazione di $\|A^{-1}\|_\infty$.
(Traccia: si applichi il teorema 3.13.)

3.11 Sia

$$A = \begin{bmatrix} 2 & -1 & -1 & 1 \\ -1 & 2 & 1 & -1 \\ -1 & 1 & 2 & -1 \\ 1 & -1 & -1 & 2 \end{bmatrix}.$$

a) Si determinino $\|A\|_1$, $\|A\|_2$ e $\|A\|_\infty$;
b) si determinino dei vettori $\mathbf{x}$ tali che

$$\|\mathbf{x}\|_1 = 1 \text{ e } \|A\mathbf{x}\|_1 = \|A\|_1,$$

$$\|\mathbf{x}\|_2 = 1 \text{ e } \|A\mathbf{x}\|_2 = \|A\|_2,$$

$$\|\mathbf{x}\|_\infty = 1 \text{ e } \|A\mathbf{x}\|_\infty = \|A\|_\infty.$$

(Risposta: a) $\|A\|_1 = \|A\|_2 = \|A\|_\infty = 5$; b) $\mathbf{x} = \mathbf{e}_1$, $\mathbf{x} = \left[\frac{1}{2}, -\frac{1}{2}, -\frac{1}{2}, \frac{1}{2}\right]^T$, $\mathbf{x} = [1, -1, -1, 1]^T$.)

3.12 Si determini una maggiorazione della norma 2 della matrice

$$A = \begin{bmatrix} -1 & 0 & 1+2\mathbf{i} \\ 0 & 2 & 1-\mathbf{i} \\ 1-2\mathbf{i} & 1+\mathbf{i} & 0 \end{bmatrix}.$$

(Risposta: $\|A\|_2 = \rho(A) \leq \|A\|_\infty = \sqrt{5} + \sqrt{2}$.)

3.13 Sia

$$A = \begin{bmatrix} \alpha & \beta & 0 \\ 0 & \alpha & \beta \\ \beta & 0 & \alpha \end{bmatrix}, \quad \alpha,\ \beta \in \mathbf{R}.$$

Si determinino α e β in modo che $\|A\|_\infty = 1$ e $\rho(A)$ sia minimo.
(Traccia: per gli autovalori di A si veda l'esercizio 2.37; risulta $\alpha = -\beta$, $|\alpha| = \dfrac{1}{2}$, $\rho(A) = \dfrac{\sqrt{3}}{2}$.)

3.14 Sia U una matrice unitaria. Si dimostri che

$$\|U\|_1 \geq 1, \quad \|U\|_2 = 1, \quad \|U\|_\infty \geq 1,$$

$$\|AU\|_2 = \|UA\|_2 = \|A\|_2, \quad \text{per ogni } A \in \mathbf{C}^{n\times n}.$$

Quest'ultima proprietà vale anche per le norme 1, ∞ e F?

3.15 Siano $\mathbf{x}, \mathbf{y} \in \mathbf{C}^n$. Si dimostri che

$$\|\mathbf{x}\mathbf{y}^H\|_2 = \|\mathbf{x}\|_2 \|\mathbf{y}\|_2.$$

(Traccia: $(\mathbf{x}\mathbf{y}^H)^H\mathbf{x}\mathbf{y}^H = (\mathbf{x}^H\mathbf{x})\mathbf{y}\mathbf{y}^H$ e $\rho(\mathbf{y}\mathbf{y}^H) = \mathbf{y}^H\mathbf{y}$. Si faccia prima l'esercizio 2.23.)

3.16 Sia $A \in \mathbf{R}^{n\times n}$ una matrice con elementi non negativi tali che

$$\sum_{j=1}^{n} a_{ij} = \sigma, \quad \text{per } i = 1, \ldots, n.$$

a) Si dimostri che A ha l'autovalore $\lambda = \sigma$ e si determini un autovettore corrispondente;

b) si dimostri che $\rho(A) = \sigma$.

(Traccia: a) $\mathbf{x} = [1,\ 1,\ \ldots,\ 1]^T$; b) si sfrutti la relazione $\rho(A) \leq \|A\|_\infty$.)

3.17 Sia $A \in \mathbf{C}^{n\times n}$ una matrice non singolare. Si dimostri che

a) $$\|A^{-1}\| = \frac{1}{\min\limits_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|} ;$$

b) per ogni autovalore λ di A vale la relazione

$$\frac{1}{\|A^{-1}\|} \leq |\lambda| .$$

(Traccia: a)

$$\|A^{-1}\| = \max_{\mathbf{y}\neq\mathbf{0}} \frac{\|A^{-1}\mathbf{y}\|}{\|\mathbf{y}\|} = \max_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{x}\|}{\|A\mathbf{x}\|} = \left(\min_{\mathbf{x}\neq\mathbf{0}} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|}\right)^{-1} ;$$

b) segue da a), essendo $\|A^{-1}\| \geq \big(\|A\mathbf{x}\|\big)^{-1}$, con $\mathbf{x}$ tale che $\|\mathbf{x}\| = 1$ e $A\mathbf{x} = \lambda\mathbf{x}$.)

3.18 Sia $A \in \mathbf{C}^{n\times n}$ e siano B e C le matrici

$$B = \begin{bmatrix} A & O \\ O & A \end{bmatrix}, \quad C = \begin{bmatrix} O & \mathbf{i}A \\ -\mathbf{i}A^H & O \end{bmatrix}.$$

Si dimostri che $\|B\|_2 = \|C\|_2 = \|A\|_2$.

3.19 Sia $A \in \mathbf{C}^{n\times n}$ e sia B la matrice

$$B = \begin{bmatrix} I & A \\ A^H & I \end{bmatrix}.$$

Si dimostri che B è definita positiva se e solo se $\|A\|_2 < 1$.

(Traccia: si sfrutti l'esercizio 2.29.)

3.20 Sia $A \in \mathbf{C}^{n\times n}$ e $B = \frac{1}{2}(A + A^H)$ e $C = \frac{1}{2}(A - A^H)$. Siano $\lambda_j = a_j + \mathbf{i}b_j$, per $j = 1, \ldots, n$ gli autovalori di A. Si dimostri il seguente *teorema di Schur*:

$$\sum_{i=1}^{n} |\lambda_i|^2 \leq \|A\|_F^2,$$

$$\sum_{i=1}^{n} |a_i|^2 \leq \|B\|_F^2,$$

$$\sum_{i=1}^{n} |b_i|^2 \leq \|C\|_F^2.$$

Il segno di uguaglianza vale nelle tre relazioni se e solo se la matrice A è normale.

(Traccia: sia $A = U(D + T)U^H$ la forma normale di Schur di A, in cui D è diagonale e T triangolare superiore in senso stretto, con $T = 0$ se e solo se la matrice A è normale. Per la prima disuguaglianza si ha

$$\|A\|_F^2 = \|D + T\|_F^2 = \sum_{i=1}^{n} |\lambda_i|^2 + \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} |a_{ij}|^2.$$

Per la seconda disuguaglianza si ha

$$\|B\|_F^2 = \Big\| \frac{D + D^H}{2} + \frac{T + T^H}{2} \Big\|_F^2 = \sum_{i=1}^{n} \Big| \frac{\lambda_i + \overline{\lambda}_i}{2} \Big|^2 + \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \Big| \frac{a_{ij} + \overline{a}_{ji}}{2} \Big|^2.$$

Per la terza disuguaglianza si proceda in modo analogo.

3.21 Sia $A \in \mathbf{C}^{n\times n}$ una matrice normale e $A = B + \mathbf{i}C$, con B e C hermitiane (si veda l'esercizio 2.19). Si dimostri che

$$\|A\|_F^2 = \|B\|_F^2 + \|C\|_F^2.$$

3.22 Sia $A \in \mathbf{C}^{n\times n}$. Si dimostri che per ogni $\epsilon > 0$ esiste una matrice $B \in \mathbf{C}^{n\times n}$ diagonalizzabile tale che

$$\|A - B\|_2 \leq \epsilon.$$

Da questa relazione segue che l'insieme delle matrici diagonalizzabili è denso in $\mathbf{C}^{n\times n}$.

(Traccia: si usi la forma normale di Schur.)

3.23 Si dimostri che per ogni ϵ esiste una costante α tale che

$$\|A^k\|_2 \leq \alpha(\rho(A) + \epsilon)^k$$

per ogni k intero positivo.

(Traccia: per il teorema 3.12 esiste una norma indotta $\| \cdot \|$ tale che $\|A\| < \rho(A) + \epsilon$. Per l'equivalenza delle norme matriciali esiste una costante α tale che $\|A^k\|_2 \leq \alpha\|A^k\|$.)

3.24 Si dimostri che per ogni norma e per ogni $A \in \mathbf{C}^{n\times n}$ risulta

$$\|\mathrm{e}^A\| \leq \mathrm{e}^{\|A\|}$$

(per la definizione dell'esponenziale di matrice si veda l'esercizio 2.43).

3.25 Sia $\| \cdot \|$ una norma su $\mathbf{C}^n$. Si dimostri che le seguenti relazioni sono equivalenti;

(1) $\| \cdot \|$ è *assoluta*, cioè $\|\mathbf{v}\| = \| \, |\mathbf{v}| \, \|$, dove $|\mathbf{v}|$ è il vettore le cui componenti sono i moduli delle componenti di $\mathbf{v}$;

(2) $\| \cdot \|$ è *monotona*, cioè $\|\mathbf{v}\| \geq \|\mathbf{u}\|$ se $|v_i| \geq |u_i|$ per $i = 1, \ldots, n$;

(3) per la norma matriciale indotta vale $\|D\| = \max\limits_{i=1,\ldots,n} |d_{ii}|$ per ogni matrice diagonale D.

Si dica quali fra le norme viste nel testo e le norme definite negli esercizi 3.4 e 3.5 sono norme assolute.

(Traccia: (1) $\Rightarrow$ (2) si verifichi prima che è sufficiente dimostrare che per $x, y \in \mathbf{R}$, se $x > y \geq 0$, allora

$$\| \, [x, \, v_2, \, \ldots, \, v_n]^T \, \| \geq \| \, [y, \, v_2, \, \ldots, \, v_n]^T \, \|.$$

Si osservi poi che, posto $\mathbf{v}_1 = [x, \, v_2, \, \ldots, \, v_n]^T$, $\mathbf{v}_2 = [y, \, v_2, \, \ldots, \, v_n]^T$, $\mathbf{v}_3 = [-x, \, v_2, \, \ldots, \, v_n]^T$, $\mathbf{v}_2$ giace sul segmento di estremi $\mathbf{v}_1$ e $\mathbf{v}_3$, e quindi la tesi segue dal fatto che $\|\mathbf{v}_1\| = \|\mathbf{v}_3\|$ e che l'insieme $\{ \, \mathbf{u} \in \mathbf{C}^n \, : \, \|\mathbf{u}\| \leq \|\mathbf{v}_1\| \, \}$ è convesso; (2) $\Rightarrow$ (3) per ogni $\mathbf{x}$ tale che $\|\mathbf{x}\| = 1$ si ha

$$\|D\mathbf{x}\| \leq \|d\mathbf{x}\| = d\|\mathbf{x}\| = d,$$

dove $d = \max\limits_{i=1,\ldots,n} |d_{ii}|$, da cui segue che $\|D\| \leq d$. Si verifichi poi che $\|D\mathbf{e}_j\| = |d_{jj}| \, \|\mathbf{e}_j\|$ e si applichi tale relazione con j tale che $|d_{jj}| = d$; (3) $\Rightarrow$ (1) siano $\mathbf{x} = [x_1, \, x_2, \, \ldots, \, x_n]^T$ e D la matrice diagonale tale che

$$d_{ii} = \begin{cases} 1 & \text{se } x_i = 0, \\ |x_i|/x_i & \text{se } x_i \neq 0. \end{cases}$$

Si verifichi che $|\mathbf{x}| = D\mathbf{x}$. Poiché $\|D\| = \|D^{-1}\| = 1$, prendendo le norme in entrambi i membri delle relazioni $|\mathbf{x}| = D\mathbf{x}$ e $\mathbf{x} = D^{-1}|\mathbf{x}|$, si ottiene la tesi. La norma dell'esercizio 3.4 e la norma (3) dell'esercizio 3.5 sono assolute, mentre le norme (1) e (2) dell'esercizio 3.5 non lo sono.)

3.26 Una funzione $\nu : \mathbf{C}^{n\times n} \to \mathbf{R}$ che verifica le proprietà a), b) e c) della definizione 3.6, ma non necessariamente la proprietà d), cioè tale che

$\nu(A) \geq 0$ e $\nu(A) = 0$ se e solo se $A = 0$,

$\nu(\alpha A) = |\alpha|\ \nu(A)$ per ogni $\alpha \in \mathbf{C}$,

$\nu(A + B) \leq \nu(A) + \nu(B)$ per ogni $B \in \mathbf{C}^{n\times n}$,

è detta norma matriciale *generalizzata*.

a) Si verifichi che la funzione

$$\nu(A) = \max_{i,j=1,\ldots,n} |a_{ij}|$$

è una norma generalizzata ma non una norma;

b) si dimostri che per ogni norma generalizzata $\nu(A)$ esiste uno scalare σ tale che $\sigma\nu(A)$ è una norma;

c) si determini σ per la norma generalizzata del punto a).

(Traccia: a) si consideri come controesempio il caso delle matrici $A = B$, con $a_{ij} = 1$, per $i, j = 1, \ldots, n$; b) il teorema di equivalenza delle norme matriciali la cui dimostrazione, come nel caso delle norme vettoriali, non fa uso della proprietà d) della definizione 3.6, vale anche per le norme generalizzate. Sia quindi $\|\ .\ \|$ una qualunque norma e α e β tali che

$$\alpha\|A\| \leq \nu(A) \leq \beta\|A\|$$

per ogni $A \in \mathbf{C}^{n\times n}$. Risulta

$$\nu(AB) \leq \beta\|AB\| \leq \beta\|A\|\|B\| \leq \frac{\beta}{\alpha^2}\nu(A)\nu(B).$$

Basta quindi porre $\sigma = \dfrac{\beta}{\alpha^2}$; c) $\sigma = n$.)

3.27 Siano $\|\ .\ \|'$ e $\|\ .\ \|''$ due norme su $\mathbf{C}^n$. Si dica se le seguenti applicazioni da $\mathbf{C}^n$ in $\mathbf{R}$ sono norme:

(1) $\mathbf{x} \to \alpha\|\mathbf{x}\|' + \beta\|\mathbf{x}\|'', \quad \alpha,\ \beta > 0$;

(2) $\mathbf{x} \to \max\{\ \|\mathbf{x}\|', \|\mathbf{x}\|''\ \}$;

(3) $\mathbf{x} \to \min\{\ \|\mathbf{x}\|', \|\mathbf{x}\|''\ \}$.

(Risposta: (1) e (2) sono norme, (3) non lo è.)

3.28 Sia $A \in \mathbf{C}^{n\times n}$ e siano $\| \cdot \|'$ e $\| \cdot \|''$ due norme vettoriali. Si dimostri che la funzione

$$A \to \max_{\|\mathbf{x}\|'=1} \|A\mathbf{x}\|''$$

è una norma generalizzata. Si dica sotto quale ipotesi tale funzione è una norma e si esamini il caso che $\| \cdot \|'$ e $\| \cdot \|''$ siano due diverse fra le norme 1, 2 e ∞.

(Traccia: poiché

$$\|AB\| = \max_{\mathbf{x}\neq\mathbf{0}} \frac{\|AB\mathbf{x}\|''}{\|\mathbf{x}\|'} = \max_{\mathbf{x}\neq\mathbf{0}} \frac{\|AB\mathbf{x}\|''}{\|B\mathbf{x}\|'} \frac{\|B\mathbf{x}\|''}{\|\mathbf{x}\|'} \frac{\|B\mathbf{x}\|'}{\|B\mathbf{x}\|''} ,$$

la funzione data è una norma se $\|\mathbf{x}\|' \leq \|\mathbf{x}\|''$ per ogni $\mathbf{x} \in \mathbf{C}^n$; è una norma nei casi (1,2), (1,∞) e (2,∞); nel caso (∞,1) si ottiene la norma generalizzata introdotta al punto a) dell'esercizio 3.26.)

3.29 Sia $A \in \mathbf{C}^{n\times n}$ e siano $\| \cdot \|'$ e $\| \cdot \|''$ due norme vettoriali. Detta $\mathbf{a}^{(i)}$ la i-esima colonna di A, si consideri il vettore

$$\mathbf{v} = \begin{bmatrix} \|\mathbf{a}^{(1)}\|' \\ \|\mathbf{a}^{(2)}\|' \\ \vdots \\ \|\mathbf{a}^{(n)}\|' \end{bmatrix} .$$

Si dica sotto quale ipotesi la funzione $A \to \|\mathbf{v}\|''$ è una norma generalizzata e per quali coppie fra le seguenti è una norma:

$$(1,1), \quad (1,\infty), \quad (\infty,1), \quad (\infty,\infty), \quad (2,2).$$

(Risposta: sotto l'ipotesi che la $\| \cdot \|''$ sia assoluta (si veda l'esercizio 3.25); è una norma nei casi $(1,1)$, $(1,\infty)$, $(\infty,1)$, $(2,2)$.)

3.30 Siano $A \in \mathbf{C}^{mn\times mn}$ una matrice $m \times m$ a blocchi $A_{ij} \in \mathbf{C}^{n\times n}$, per $i,j = 1,\ldots,m$, in cui i blocchi A_{ii} siano hermitiani.

a) Detti $\mu_k^{(i)}$, per $k = 1,\ldots,n$ gli autovalori di A_{ii}, si dimostri che gli autovalori di A sono contenuti nell'unione dei cerchi

$$K_{ij} = \{\ z \in \mathbf{C} \quad \text{tali che} \quad |z - \mu_j^{(i)}| \leq \sum_{\substack{j=1 \\ j\neq i}}^{m} \|A_{ij}\|_2 \ \}.$$

b) Per la matrice tridiagonale ad m blocchi $A \in \mathbf{R}^{2m\times 2m}$

$$A = \begin{bmatrix} I_2 & T & & \\ T & I_2 & \ddots & \\ & \ddots & \ddots & T \\ & & T & I_2 \end{bmatrix}, \quad \text{dove } T = \begin{bmatrix} 10 & 10 \\ 1 & 1 \end{bmatrix},$$

si dia una localizzazione degli autovalori di A con i cerchi del punto a) e si confronti con la localizzazione che si ottiene con i cerchi di Gerschgorin.

(Traccia: a) Sia λ un autovalore di A che non sia autovalore di alcun A_{ii} (se ciò non fosse la tesi sarebbe ovvia) e $A\mathbf{x} = \lambda\mathbf{x}$, con $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m]^T$, $\mathbf{x}_i \in \mathbf{C}^n$ per $i = 1, \ldots, m$. Si ha

$$(A_{ii} - \lambda I)\mathbf{x}_i = -\sum_{\substack{j=1 \\ j\neq i}}^{m} A_{ij}\mathbf{x}_j$$

da cui

$$\|\mathbf{x}_i\|_2 \leq \|(A_{ii} - \lambda I)^{-1}\|_2 \sum_{\substack{j=1 \\ j\neq i}}^{m} \|A_{ij}\|_2 \|\mathbf{x}_j\|_2.$$

Si scelga come indice i quello per cui $\|\mathbf{x}_i\|_2 = \max\limits_{j=1,\ldots,m} \|\mathbf{x}_j\|_2$ e si tenga conto del fatto che

$$\|(A_{ii} - \lambda I)^{-1}\|_2 = \left[\min_{j=1,\ldots,n} |\mu_j^{(i)} - \lambda| \right]^{-1}.$$

b) È $\mu_j^{(i)} = 1$ per $i = 1, \ldots, m,\ j = 1, 2$, $\|T\|_2 = \sqrt{202}$, per cui gli autovalori appartengono al cerchio di centro 1 e raggio $2\sqrt{202}$. Con il teorema di Gerschgorin gli autovalori di A vengono localizzati nel cerchio di centro 1 e raggio 40.)

3.31 Si dimostri che la funzione

$$A \to \sqrt[p]{\sum_{i=1}^{n}\sum_{j=1}^{n} |a_{ij}|^p}, \qquad p \leq 1,$$

definisce una norma matriciale generalizzata (detta norma *hölderiana*), che è una norma per $1 \leq p \leq 2$. È una norma indotta?

(Traccia: si usino le disuguaglianze di Hölder e di Minkowski dimostrate nell'esercizio 3.4; non è una norma indotta.)

3.32 Siano $A \in \mathbf{C}^{n\times n}$ e $B \in \mathbf{C}^{m\times m}$. Si dimostri che

$$\|A \otimes B\| = \|A\| \, \|B\|,$$

per le norme 1, 2 e ∞ (per la definizione di prodotto diretto $\otimes$ si veda l'esercizio 1.60, per le proprietà degli autovalori l'esercizio 2.41.)

Commento bibliografico

Per quanto il concetto di norma sia estesamente usato in analisi funzionale, e fin dal 1887 Peano abbia introdotto la definizione di norma di uno spazio vettoriale a dimensione finita, l'uso delle norme in analisi numerica è relativamente recente. Infatti, mentre il teorema 3.10 è stato formulato da Frobenius, il teorema 3.12, che viene utilizzato nello studio della convergenza dei metodi iterativi per la risoluzione dei sistemi lineari, è stato formulato da Ostrowski nel 1960 [4].

La tecnica attuale di definire le norme vettoriali e matriciali indipendentemente e in modo assiomatico, e poi collegare le due definizioni con le proprietà di consistenza, è stato suggerita da Faddeeva nel 1950 e ripreso da Householder nel 1958 [2].

Le tre norme definite nel paragrafo 1 possono essere considerate come caso particolare della *norma p* (si veda l'esercizio 3.4); la norma 2 viene anche detta norma euclidea (su $\mathbf{R}$) o norma unitaria (su $\mathbf{C}$), la norma ∞ è anche detta norma del massimo o di Chebyshev.

Una descrizione elementare delle proprietà più importanti delle norme 1 e ∞ si trova in [1]. Per una trattazione generale della norme, in particolare per i teoremi di equivalenza delle norme, si veda [3]. Un elenco di relazioni di equivalenza fra norme matriciali è riportata in [5].

Bibliografia

[1] V. N. Faddeeva, *Computational Methods of Linear Algebra*, Dover, New York, 1959.

[2] A. S. Householder, "The Approximate Solution of Matrix Problems", *J. Assoc. Comp. Mach.* 5, 1958, pp. 204-243.

[3] A. M. Ostrowski, "Über Normen von Matrizen", *Math. Z. Bd.* 63, 1955, pp. 2-18.

[4] A. M. Ostrowski, *Solution of Equations and Systems of Equations*, Academic Press, 1960.

[5] A. M. Turing, "Rounding-off Errors in Matrix Processes", *Quart. J. Mech. and Appl. Math.*, 1, 1948, pp. 287-308.

Capitolo 4

METODI DIRETTI PER LA RISOLUZIONE DI SISTEMI DI EQUAZIONI LINEARI

1. Analisi dell'errore

Siano $A \in \mathbf{C}^{n\times n}, \mathbf{x}, \mathbf{b} \in \mathbf{C}^n$, e si supponga che il sistema lineare

$$A\mathbf{x} = \mathbf{b} \tag{1}$$

sia consistente.

I metodi per la risoluzione numerica del sistema (1) possono essere divisi in due classi: *metodi diretti* e *metodi iterativi*. In un metodo diretto, se non ci fossero errori di rappresentazione dei dati e di arrotondamento nei calcoli, la soluzione del sistema verrebbe calcolata esattamente. Invece in un metodo iterativo, anche nell'ipotesi che non ci siano errori di rappresentazione dei dati e di arrotondamento nei calcoli, si deve comunque operare un troncamento del procedimento, commettendo un errore (*errore analitico o di troncamento*).

In ogni caso però, qualunque metodo si usi, non si può prescindere dagli errori di rappresentazione dei dati e di arrotondamento nei calcoli. Lo studio dell'errore che viene fatto si basa su un'ipotesi generalmente verificata: che i termini contenenti espressioni quadratiche degli errori siano trascurabili rispetto ai termini contenenti espressioni lineari negli errori. Una maggiorazione dell'errore da cui è affetta la soluzione effettivamente calcolata può essere rappresentata, a meno di termini di ordine superiore, da due termini distinti, uno dovuto agli errori di rappresentazione dei dati, che non dipendono dal particolare metodo usato e che è detto *errore inerente*, e l'altro dovuto agli errori di arrotondamento nei calcoli, che dipende dal metodo usato, ma non dagli errori sui dati A e $\mathbf{b}$, e che viene detto *errore algoritmico*.

L'errore inerente misura la sensibilità della soluzione agli errori sui dati: un sistema lineare, per cui a "piccoli" errori nei dati corrispondono "grandi" errori nella soluzione, è un problema difficile da risolvere e viene detto *mal condizionato* o *mal posto*; un sistema lineare per cui a piccoli errori sui dati corrispondono piccoli errori sulla soluzione è detto *ben condizionato* o *ben posto*. Lo studio dell'errore inerente può essere fatto *perturbando* i dati ed esaminando gli effetti indotti da queste perturbazioni sulla soluzione.

4.1 Teorema. *Siano $\delta A \in \mathbf{C}^{n\times n}$ e $\delta\mathbf{b} \in \mathbf{C}^n$ rispettivamente la matrice e il vettore delle perturbazioni sui dati del sistema (1) dove $\mathbf{b} \neq \mathbf{0}$ e sia $\|\ .\ \|$ una qualunque norma matriciale indotta. Se A è non singolare e se $\|A^{-1}\|\ \|\delta A\| < 1$, allora anche la matrice $A + \delta A$ è non singolare. Indicata con $\mathbf{x} + \delta\mathbf{x}$ la soluzione del sistema perturbato*

$$(A + \delta A)\ (\mathbf{x} + \delta\mathbf{x}) = \mathbf{b} + \delta\mathbf{b}, \tag{2}$$

risulta

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \ \leq\ \mu(A)\ \frac{\|\delta A\|/\|A\| + \|\delta\mathbf{b}\|/\|\mathbf{b}\|}{1 - \mu(A)\,\|\delta A\|/\|A\|}$$

in cui $\mu(A) = \|A\|\|A^{-1}\|$ è il numero di condizionamento della matrice A.

Dim. Poiché $A+\delta A = A(I+A^{-1}\delta A)$ e, per ipotesi, $\|A^{-1}\delta A\| \leq \|A^{-1}\|\|\delta A\| < 1$, dal teorema 3.13 si ha che la matrice $I + A^{-1}\delta A$, e quindi la matrice $A + \delta A$, è non singolare e risulta

$$\|(I + A^{-1}\delta A)^{-1}\| \leq \frac{1}{1 - \|A^{-1}\|\ \|\delta A\|}\ .$$

Sottraendo membro a membro la (1) dalla (2) si ottiene

$$(A + \delta A)\delta\mathbf{x} = -\delta A\mathbf{x} + \delta\mathbf{b},$$

moltiplicando entrambi i membri per A^{-1} si ha

$$(I + A^{-1}\delta A)\delta\mathbf{x} = A^{-1}(-\delta A\mathbf{x} + \delta\mathbf{b}),$$

da cui

$$\delta\mathbf{x} = (I + A^{-1}\delta A)^{-1}A^{-1}(-\delta A\mathbf{x} + \delta\mathbf{b}),$$

e

$$\|\delta\mathbf{x}\| \ \leq\ \frac{\|A^{-1}\|\,(\|\delta A\|\,\|\mathbf{x}\| + \|\delta\mathbf{b}\|)}{1 - \|A^{-1}\|\|\delta A\|}. \tag{3}$$

Poiché per ipotesi è $\mathbf{b} \neq \mathbf{0}$ e A è non singolare, risulta $\|\mathbf{x}\| > 0$, per cui dividendo entrambi i membri della (3) per $\|\mathbf{x}\|$ e tenendo conto che per la (1) $\|\mathbf{b}\| \leq \|A\|\,\|\mathbf{x}\|$, si ha:

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \frac{\|A^{-1}\|\,(\|\delta A\| + \|\delta\mathbf{b}\|\|A\|/\|\mathbf{b}\|)}{1 - \|A^{-1}\|\|\delta A\|} = \mu(A)\ \frac{\|\delta A\|/\|A\| + \|\delta\mathbf{b}\|/\|\mathbf{b}\|}{1 - \mu(A)\,\|\delta A\|/\|A\|}\ .$$

∎

Indicando con $\epsilon_A = \|\delta A\|/\|A\|$ e $\epsilon_b = \|\delta\mathbf{b}\|/\|\mathbf{b}\|$ le perturbazioni relative della matrice A e del vettore $\mathbf{b}$ e con $\epsilon_x = \|\delta\mathbf{x}\|/\|\mathbf{x}\|$ la perturbazione

relativa indotta sul vettore $\mathbf{x}$, il teorema precedente può essere così riformulato: la *perturbazione relativa* ϵ_x della soluzione, indotta dalle perturbazioni relative dei dati ϵ_A e ϵ_b, è maggiorata dall'espressione

$$\epsilon_x \leq \mu(A)\ \frac{\epsilon_A + \epsilon_b}{1 - \mu(A)\epsilon_A}\ . \tag{4}$$

Si osservi che il numero di condizionamento è sempre maggiore o uguale a 1; infatti:

$$\mu(A) = \|A\|\ \|A^{-1}\| \geq \|AA^{-1}\| = 1.$$

Dalla (4) risulta che se $\mu(A)$ assume valori piccoli, allora piccole perturbazioni sui dati inducono piccole perturbazioni sulla soluzione e quindi il problema è ben posto: in questo caso la matrice del sistema si dice *ben condizionata*; se $\mu(A)$ assume valori grandi, allora piccole variazioni sui dati possono indurre grandi perturbazioni nella soluzione e quindi il problema può essere mal posto: in questo caso la matrice del sistema si dice *mal condizionata*. Se ad esempio $\mu(A) = 1000$, l'errore ϵ_x può essere 1000 volte quello presente nei dati.

Un esempio classico di matrice mal condizionata è la matrice di Hilbert.

4.2 Esempio. La matrice $A^{(n)}$ di ordine n, definita da

$$a_{ij}^{(n)} = \frac{1}{i+j-1}, \quad i,j = 1,\ldots,n,$$

è detta matrice di *Hilbert*. Per $n = 5$ si ha

$$A^{(5)} = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} \\ \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} & \frac{1}{8} \\ \frac{1}{5} & \frac{1}{6} & \frac{1}{7} & \frac{1}{8} & \frac{1}{9} \end{bmatrix}$$

$$[A^{(5)}]^{-1} = \begin{bmatrix} 25 & -300 & 1050 & -1400 & 630 \\ -300 & 4800 & -18900 & 26880 & -12600 \\ 1050 & -18900 & 79380 & -117600 & 56700 \\ -1400 & 26880 & -117600 & 179200 & -88200 \\ 630 & -12600 & 56700 & -88200 & 44100 \end{bmatrix}.$$

Per ogni valore di n la matrice $B^{(n)} = [A^{(n)}]^{-1}$ ha elementi $b_{ij}^{(n)}$ interi, tali che $|b_{ij}^{(n)}|$ è, per ogni i e j, una funzione crescente di n, $n \geq \max\{i, j\}$. Nella tabella che segue vengono riportati i valori del numero di condizionamento $\mu_2(A) = \|A\|_2\|A^{-1}\|_2$, in norma 2, e $\mu_\infty(A) = \|A\|_\infty\|A^{-1}\|_\infty$, in norma ∞, della matrice di Hilbert per valori di n da 2 a 10.

n	$\mu_2(A^{(n)})$	$\mu_\infty(A^{(n)})$
2	1.505	27
3	$5.241\ 10^2$	$7.480\ 10^2$
4	$1.551\ 10^4$	$2.837\ 10^4$
5	$4.766\ 10^5$	$9.436\ 10^5$
6	$1.495\ 10^7$	$2.907\ 10^7$
7	$4.754\ 10^8$	$9.852\ 10^8$
8	$1.526\ 10^{10}$	$3.387\ 10^{10}$
9	$4.932\ 10^{11}$	$1.099\ 10^{12}$
10	$1.603\ 10^{13}$	$3.535\ 10^{13}$

Asintoticamente il numero di condizionamento in norma 2 di $A^{(n)}$ risulta essere una funzione crescente di n dell'ordine di $e^{3.5n}$ [13]. ∎

Si osservi che la (4), essendo una maggiorazione, può fornire una stima eccessiva dell'errore della soluzione indotto dall'errore nei dati, tenuto anche conto che tale maggiorazione vale per qualunque vettore **b**.

4.3 Esempio. Data la matrice

$$A = \begin{bmatrix} 1 & 1 \\ 0.99 & 1 \end{bmatrix},$$

si ha

$$A^{-1} = \frac{1}{0.01}\begin{bmatrix} 1 & -1 \\ -0.99 & 1 \end{bmatrix}$$

e quindi

$$\mu_\infty(A) = \|A\|_\infty\ \|A^{-1}\|_\infty = 400.$$

Perturbando A nel modo seguente

$$A + \delta A = A + 0.002\ \begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix} = \begin{bmatrix} 1.002 & 1.002 \\ 0.988 & 0.998 \end{bmatrix},$$

ed essendo

$$\|A^{-1}\|_\infty = 200 \quad \text{e} \quad \|\delta A\|_\infty = 0.004,$$

risulta

$$\|A^{-1}\|_\infty \, \|\delta A\|_\infty = 0.8 < 1.$$

Il sistema lineare $A\mathbf{x} = \mathbf{b}$, con $\mathbf{b} = [2, 1.99]^T$, ha come soluzione il vettore $\mathbf{x} = [1, 1]^T$. Il sistema con matrice perturbata $(A + \delta A)\ (\mathbf{x} + \delta\mathbf{x}) = \mathbf{b}$ ha come soluzione il vettore $\mathbf{x} + \delta\mathbf{x} = [0.2016\ldots, 1.794\ldots]^T$, per cui $\epsilon_x = \|\delta\mathbf{x}\|_\infty / \|\mathbf{x}\|_\infty = 0.3992\ldots$ Si osservi che per la (4) risulta

$$\epsilon_x \leq \mu(A)\ \frac{\epsilon_A + \epsilon_b}{1 - \mu(A)\epsilon_A} = 4.$$

È opportuno rilevare che, pur rimanendo inalterata la maggiorazione dell'errore, per il vettore dei termini noti $\mathbf{b} = [0, -0.01]^T$, sia il sistema $A\mathbf{x} = \mathbf{b}$, che il sistema $(A + \delta A)\mathbf{x} = \mathbf{b}$ hanno la stessa soluzione $\mathbf{x} = [1, -1]^T$, e quindi risulta $\epsilon_x = 0$. ∎

Se A è una matrice hermitiana, utilizzando la norma 2, si ha che

$$\mu_2(A) = \|A\|_2 \|A^{-1}\|_2 = \frac{\sigma_{\max}}{\sigma_{\min}},$$

in cui $\sigma_{\max}$ e $\sigma_{\min}$ sono rispettivamente il modulo massimo e il modulo minimo degli autovalori di A. Cioè una matrice A è tanto meglio condizionata quanto più vicini sono fra loro i suoi autovalori. La matrice A è mal condizionata se un autovalore è in modulo molto piccolo rispetto agli altri. Si osservi che nel caso in cui la matrice A, oltre ad essere hermitiana, è anche definita positiva, risulta in norma 2

$$\mu_2(A) = \frac{\lambda_{\max}}{\lambda_{\min}}, \tag{5}$$

in cui $\lambda_{\max}$ e $\lambda_{\min}$ sono rispettivamente il massimo e il minimo degli autovalori di A.

Per analizzare l'errore algoritmico verrà usata la tecnica cosiddetta di *analisi all'indietro (backward analysis)*, in cui la soluzione effettivamente calcolata $\mathbf{y}$ viene considerata come soluzione esatta di un problema perturbato del tipo

$$(A + \Delta A)\ \mathbf{y} = \mathbf{b} + \Delta\mathbf{b}.$$

A differenza dell'analisi fatta prima per l'errore inerente, adesso la matrice A e il vettore $\mathbf{b}$ sono formati da numeri di macchina e ΔA e $\Delta\mathbf{b}$ non sono perturbazioni introdotte sui dati iniziali, ma sono legate agli errori commessi durante i calcoli e quindi alla precisione con cui vengono eseguite le operazioni.

Nell'analisi della propagazione degli errori di arrotondamento generati da un metodo di risoluzione si deve determinare da quali fattori, oltre alla

precisione con cui vengono eseguite le operazioni, dipendono ΔA e $\Delta \mathbf{b}$. Un metodo risulta più *stabile* di un altro se è meno sensibile agli errori indotti dai calcoli. Si tenga però presente che lo studio della *stabilità* di un metodo può perdere di significatività quando il problema è fortemente mal condizionato, poiché in questo caso l'errore inerente prevale sull'errore algoritmico.

L'efficienza di un metodo dipende, oltre che dalla sua stabilità numerica, anche dal suo *costo computazionale*, cioè dal numero di operazioni aritmetiche richieste. In pratica come misura di questo costo si considera solo il numero delle operazioni moltiplicative (moltiplicazioni e divisioni) richieste, in quanto il numero delle operazioni additive (addizioni e sottrazioni) è generalmente dello stesso ordine del numero delle operazioni moltiplicative. Il costo computazionale di un metodo è quindi legato al tempo richiesto da un calcolatore per l'esecuzione del relativo algoritmo. Una valutazione più accurata dovrebbe prendere in considerazione, oltre alle operazioni additive, anche le operazioni necessarie alla gestione dei dati del problema nella memoria del calcolatore (calcolo degli indici, permutazioni, ecc.).

Il costo computazionale è dato come funzione della dimensione n della matrice A. Di tale funzione si riportano solo i termini di ordine più elevato in n, usando il simbolo $\simeq$, che si legge appunto *uguale a meno di termini di ordine inferiore*. Ad esempio:

$$(2n+1)^2 \simeq 4n^2.$$

Per il calcolo del costo computazionale sono utili le seguenti formule

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2} \simeq \frac{n^2}{2},$$

$$\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6} \simeq \frac{n^3}{3},$$

che possono essere facilmente dimostrate per induzione.

2. Sistemi lineari con matrice triangolare

La risoluzione del sistema (1) è particolarmente semplice se la matrice A è triangolare. Se, ad esempio, A è triangolare superiore, risulta

$$\begin{cases} a_{ii}x_i + \displaystyle\sum_{j=i+1}^{n} a_{ij}x_j = b_i, \qquad i = 1, \ldots, n-1, \\ a_{nn}x_n = b_n. \end{cases}$$

Se A è non singolare, cioè $a_{ii} \neq 0$, per $i = 1, \ldots, n$, si ha:

$$\begin{cases} x_n = \dfrac{b_n}{a_{nn}} \\ x_i = \dfrac{1}{a_{ii}} \Big[b_i - \displaystyle\sum_{j=i+1}^{n} a_{ij} x_j \Big], \qquad i = n-1, \ldots, 1, \end{cases} \tag{6}$$

quindi la risoluzione procede calcolando nell'ordine $x_n, x_{n-1}, \ldots, x_1$: all'$i$-esimo passo, per calcolare x_i, vengono utilizzate le componenti di indice maggiore di i, già calcolate. Si tratta cioè di una *sostituzione all'indietro*.

Se A è singolare, allora esiste almeno un indice k per cui $a_{kk} = 0$, cioè la k-esima equazione del sistema risulta

$$\sum_{j=k+1}^{n} a_{kj} x_j = b_k. \tag{7}$$

Perciò se il sistema è consistente, la k-esima equazione è verificata per qualsiasi valore di x_k. La soluzione si calcola usando la (6) per ogni indice k per cui $a_{kk} \neq 0$; se $a_{kk} = 0$, si controlla la consistenza del sistema, verificando che le x_i, con $i > k$, già calcolate soddisfino la (7). Se così è, si assegna ad x_k un valore arbitrario e si prosegue la sostituzione all'indietro. Si osservi che, poiché si usa un'aritmetica finita, anche se il sistema è consistente, è possibile che la (7) sia verificata solo a meno di una quantità che dipende dalla precisione di macchina u e dalla grandezza degli elementi che intervengono nella (7).

Se la matrice del sistema fosse triangolare inferiore, la risoluzione avverrebbe in modo analogo, con il semplice scambio dell'ordine in cui svolgere i calcoli: dalla prima componente di $\mathbf{x}$ verso l'ultima (*sostituzione in avanti*).

Lo stesso procedimento può essere utilizzato per calcolare la matrice inversa di una matrice $A \in \mathbf{C}^{n \times n}$ non singolare e triangolare. Infatti la matrice X, inversa di A, è tale che

$$AX = I,$$

e quindi la k-esima colonna di X è un vettore $\mathbf{x}_k$ che verifica la relazione

$$A\mathbf{x}_k = \mathbf{e}_k, \quad k = 1, 2, \ldots, n, \tag{8}$$

dove $\mathbf{e}_k$ è la k-esima colonna di I. Poiché la matrice A è triangolare, anche la matrice X risulta triangolare: in particolare, se A è triangolare superiore (inferiore), il vettore $\mathbf{x}_k$ ha le ultime $n - k$ componenti (le prime $k - 1$ componenti) nulle.

Il costo computazionale della risoluzione con il procedimento (6) di un sistema con matrice triangolare superiore è determinato tenendo conto del fatto che la componente x_i viene calcolata con $n-i$ moltiplicazioni e 1 divisione, per cui risulta

$$\sum_{i=1}^{n}(n-i+1)=\sum_{i=1}^{n} i \simeq \frac{n^2}{2}.$$

L'inversa di una matrice triangolare si ottiene risolvendo gli n sistemi (8) in cui si determinano solo le prime k componenti di $\mathbf{x}_k$ se A è triangolare superiore (solo le ultime $n-k$ componenti di $\mathbf{x}_k$ se A è triangolare inferiore). Quindi la risoluzione del k-esimo sistema lineare (8) richiede $k^2/2$ (rispettivamente $(n-k+1)^2/2$) operazioni moltiplicative, e il costo computazionale del calcolo della matrice inversa è dato da

$$\sum_{k=1}^{n}\frac{(n-k+1)^2}{2}=\sum_{k=1}^{n}\frac{k^2}{2}\simeq\frac{n^3}{6}.$$

3. Fattorizzazioni

Molti dei metodi numerici diretti utilizzano per la risoluzione di (1) una fattorizzazione della matrice A nel prodotto di due matrici B e C

$$A=BC,$$

dove le matrici B e C sono facilmente invertibili. Il sistema (1) risulta allora

$$BC\mathbf{x}=\mathbf{b}$$

e la soluzione di (1) viene calcolata risolvendo successivamente i due sistemi lineari

$$\begin{aligned} B\mathbf{y}&=\mathbf{b},\\ C\mathbf{x}&=\mathbf{y}. \end{aligned} \tag{9}$$

Fattorizzazioni diverse della matrice A sono associate a metodi di risoluzione del sistema (1) diversi. Tre fattorizzazioni classiche sono le seguenti.

1. La *fattorizzazione LU*: L è una matrice triangolare inferiore con elementi principali uguali ad 1 ed U è una matrice triangolare superiore. Tale fattorizzazione è associata al metodo di Gauss.
2. La *fattorizzazione* LL^H: L è una matrice triangolare inferiore con elementi principali positivi. Tale fattorizzazione è associata al metodo di Cholesky.

3. La *fattorizzazione QR*: Q è una matrice unitaria ed R è una matrice triangolare superiore. Tale fattorizzazione è associata al metodo di Householder.

Se la matrice A è reale, le matrici delle tre fattorizzazioni, quando esistono, sono reali.

Il costo computazionale della fattorizzazione è dato, come si vedrà, da un numero di operazioni dell'ordine di n^3, mentre il costo computazionale della risoluzione dei sistemi (9) è dato da un numero di operazioni dell'ordine di n^2.

La fattorizzazione QR esiste per ogni matrice A, mentre non sempre è possibile ottenere le fattorizzazioni LU e LL^H. Valgono infatti i seguenti teoremi.

4.4 Teorema. *Sia A una matrice di ordine n e siano A_k le sue sottomatrici principali di testa di ordine k. Se A_k è non singolare per $k = 1, \ldots, n-1$, allora esiste ed è unica la fattorizzazione LU di A.*

Dim. Si procede per induzione.

Se $n = 1$, $A_1 = [a_{11}]$ e quindi si ha $L = [1]$ e $U = [a_{11}]$, univocamente.

Se $n = k > 1$, la matrice A_k può essere partizionata nel modo seguente

$$A_k = \begin{bmatrix} A_{k-1} & \mathbf{d} \\ \mathbf{c}^H & \alpha \end{bmatrix},$$

in cui $A_{k-1} = L_{k-1}U_{k-1}$, con L_{k-1} matrice triangolare inferiore con elementi principali uguali ad 1 e U_{k-1} matrice triangolare superiore. Posto

$$L_k = \begin{bmatrix} L_{k-1} & \mathbf{0} \\ \mathbf{u}^H & 1 \end{bmatrix}, \qquad U_k = \begin{bmatrix} U_{k-1} & \mathbf{v} \\ \mathbf{0}^H & \beta \end{bmatrix},$$

occorre determinare $\mathbf{u}$, $\mathbf{v}$ e β in modo che $A_k = L_kU_k$. Poiché risulta

$$L_kU_k = \begin{bmatrix} L_{k-1}U_{k-1} & L_{k-1}\mathbf{v} \\ \mathbf{u}^HU_{k-1} & \mathbf{u}^H\mathbf{v} + \beta \end{bmatrix},$$

si ha che la relazione $A_k = L_kU_k$ è verificata se e solo se

$$\begin{aligned} L_{k-1}\mathbf{v} &= \mathbf{d}, \\ U_{k-1}^H\mathbf{u} &= \mathbf{c}, \\ \mathbf{u}^H\mathbf{v} + \beta &= \alpha. \end{aligned}$$

I vettori $\mathbf{u}$ e $\mathbf{v}$ risultano determinati univocamente dalle prime due relazioni, poiché $\det L_{k-1} = 1$ e $\det U_{k-1} = \det A_{k-1} \neq 0$ in quanto A_{k-1} è non singolare. Dalla terza relazione si ricava univocamente $\beta = \alpha - \mathbf{u}^H\mathbf{v}$. ∎

4.5 Teorema. *Sia A una matrice di ordine n. Allora esiste una matrice di permutazione Π per cui si può ottenere la fattorizzazione LU di ΠA, cioè*

$$\Pi A = LU.$$

Dim. Si procede per induzione su n.

Se $n = 1$, $L = [1]$ e $U = [a_{11}]$ e quindi $\Pi = [1]$.

Se $n = k > 1$, possono presentarsi questi due casi:

a) tutti gli elementi della prima colonna di A sono nulli, e quindi la matrice A è della forma

$$A = \begin{bmatrix} 0 & \mathbf{c}^H \\ \mathbf{0} & A_{k-1} \end{bmatrix},$$

dove $A_{k-1} \in \mathbf{C}^{(n-1)\times(n-1)}$. Per l'ipotesi induttiva esiste una matrice Π_{k-1} per cui si può scrivere la fattorizzazione LU della matrice $\Pi_{k-1}A_{k-1}$, cioè

$$\Pi_{k-1}A_{k-1} = L_{k-1}U_{k-1},$$

per cui si ha

$$\begin{bmatrix} 1 & \mathbf{0}^H \\ \mathbf{0} & \Pi_{k-1} \end{bmatrix} A = \begin{bmatrix} 1 & \mathbf{0}^H \\ \mathbf{0} & L_{k-1} \end{bmatrix} \begin{bmatrix} 0 & \mathbf{c}^H \\ \mathbf{0} & U_{k-1} \end{bmatrix},$$

che rappresenta la fattorizzazione LU di ΠA, dove

$$\Pi = \begin{bmatrix} 1 & \mathbf{0}^H \\ \mathbf{0} & \Pi_{k-1} \end{bmatrix}.$$

b) non tutti gli elementi della prima colonna sono nulli, allora se $a_{11} \neq 0$ si pone $\Pi' = I$, altrimenti, se $a_{11} = 0$ e se i è un indice tale che $a_{i1} \neq 0$, allora si sceglie come Π' la matrice di permutazione ottenuta scambiando la prima con la i-esima riga di I.

La matrice $\Pi' A$ è della forma

$$\Pi' A = \begin{bmatrix} \alpha & \mathbf{c}^H \\ \mathbf{d} & A_{k-1} \end{bmatrix},$$

dove $A_{k-1} \in \mathbf{C}^{(n-1)\times(n-1)}$ e $\alpha \neq 0$. La matrice $\Pi' A$ si può allora scrivere come prodotto

$$\Pi' A = \begin{bmatrix} 1 & \mathbf{0}^H \\ \frac{1}{\alpha}\,\mathbf{d} & I \end{bmatrix} \begin{bmatrix} \alpha & \mathbf{c}^H \\ \mathbf{0} & B_{k-1} \end{bmatrix}$$

dove

$$B_{k-1} = A_{k-1} - \frac{1}{\alpha}\,\mathbf{d}\mathbf{c}^H.$$

Per l'ipotesi induttiva esiste una matrice di permutazione Π_{k-1} tale che esiste la fattorizzazione LU della matrice $\Pi_{k-1}B_{k-1}$, cioè $\Pi_{k-1}B_{k-1} = L_{k-1}U_{k-1}$. Si ha allora

$$\begin{aligned}\Pi' A &= \begin{bmatrix} 1 & \mathbf{0}^H \\ \frac{1}{\alpha}\,\mathbf{d} & I \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0}^H \\ \mathbf{0} & \Pi_{k-1}^T L_{k-1} \end{bmatrix} \begin{bmatrix} \alpha & \mathbf{c}^H \\ \mathbf{0} & U_{k-1} \end{bmatrix} \\ &= \begin{bmatrix} 1 & \mathbf{0}^H \\ \frac{1}{\alpha}\mathbf{d} & \Pi_{k-1}^T L_{k-1} \end{bmatrix} \begin{bmatrix} \alpha & \mathbf{c}^H \\ \mathbf{0} & U_{k-1} \end{bmatrix} \\ &= \begin{bmatrix} 1 & \mathbf{0}^H \\ \mathbf{0} & \Pi_{k-1}^T \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0}^H \\ \frac{1}{\alpha}\Pi_{k-1}\mathbf{d} & L_{k-1} \end{bmatrix} \begin{bmatrix} \alpha & \mathbf{c}^H \\ \mathbf{0} & U_{k-1} \end{bmatrix},\end{aligned}$$

da cui

$$\begin{bmatrix} 1 & \mathbf{0}^H \\ \mathbf{0} & \Pi_{k-1} \end{bmatrix} \Pi' A = \begin{bmatrix} 1 & \mathbf{0}^H \\ \frac{1}{\alpha}\Pi_{k-1}\mathbf{d} & L_{k-1} \end{bmatrix} \begin{bmatrix} \alpha & \mathbf{c}^H \\ \mathbf{0} & U_{k-1} \end{bmatrix},$$

che rappresenta la fattorizzazione LU della matrice ΠA, dove

$$\Pi = \begin{bmatrix} 1 & \mathbf{0}^H \\ \mathbf{0} & \Pi_{k-1} \end{bmatrix} \Pi'.$$

∎

Si osservi che, data una matrice A, vi possono essere diverse matrici di permutazione Π tali che le matrici ΠA soddisfano alle ipotesi del teorema 4.4.

4.6 Esempio. La matrice

$$A = \begin{bmatrix} 1 & 2 & -1 \\ -1 & -1 & 2 \\ 1 & 1 & 2 \end{bmatrix}$$

soddisfa alle ipotesi del teorema 4.4. La sua fattorizzazione LU è

$$A = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 1 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & -1 \\ 0 & 1 & 1 \\ 0 & 0 & 4 \end{bmatrix}.$$

La matrice

$$A = \begin{bmatrix} 1 & 2 & -1 \\ -1 & -2 & 0 \\ 1 & 1 & 2 \end{bmatrix}$$

non soddisfa alle ipotesi del teorema 4.4, poiché la sottomatrice principale di testa di ordine 2 è singolare. Ponendo

$$\Pi = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix},$$

risulta

$$\Pi A = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & -1 \\ 0 & -1 & 3 \\ 0 & 0 & -1 \end{bmatrix},$$

e ponendo

$$\Pi = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix},$$

risulta

$$\Pi A = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 1 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 2 \\ 0 & -1 & 2 \\ 0 & 0 & -1 \end{bmatrix}.$$

La matrice singolare

$$A = \begin{bmatrix} 1 & 2 & -1 \\ -1 & -2 & 1 \\ 1 & 1 & 2 \end{bmatrix},$$

non soddisfa alle ipotesi del teorema 4.4, poiché la sottomatrice principale di testa di ordine 2 è singolare. Ponendo

$$\Pi = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix},$$

risulta

$$\Pi A = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & -1 \\ 0 & -1 & 3 \\ 0 & 0 & 0 \end{bmatrix}.$$

■

4.7 Teorema. *Sia A una matrice hermitiana di ordine n. Se A è definita positiva, allora esiste ed è unica la fattorizzazione LL^H di A.*

Dim. Per il teorema 2.33 tutte le sottomatrici principali di testa di A sono non singolari. Quindi per il teorema 4.4 risulta in modo univoco che

$$A = MU, \tag{10}$$

in cui M è una matrice triangolare inferiore con elementi principali uguali ad 1 e U è una matrice triangolare superiore. Se D è la matrice diagonale i cui elementi principali sono quelli di U, risulta

$$A = MDR,$$

in cui R, matrice triangolare superiore con elementi principali uguali ad 1, è tale che $DR = U$. Poiché A è hermitiana, si ha:

$$A = A^H = R^H D^H M^H,$$

e per l'unicità della decomposizione (10) segue

$$R^H = M \quad \text{e} \quad D^H M^H = U = DR,$$

da cui $R = M^H$ e $D = D^H$. Risulta allora univocamente

$$A = MDM^H,$$

in cui D è una matrice diagonale reale. Se $\mathbf{x} = M^H\mathbf{y}$ si ha

$$\mathbf{x}^H D\mathbf{x} = \mathbf{y}^H MDM^H\mathbf{y},$$

e poiché M è non singolare, se $\mathbf{x} \neq \mathbf{0}$ si ha $\mathbf{y} \neq \mathbf{0}$ e

$$\mathbf{x}^H D\mathbf{x} = \mathbf{y}^H A\mathbf{y} > 0,$$

essendo A definita positiva. Ne segue che anche D è definita positiva e quindi i suoi elementi principali sono reali e positivi. Esiste allora un'unica matrice diagonale F ad elementi principali reali e positivi, tale che $F^2 = D$, e posto $L = MF$ si ha

$$A = MF^2M^H = LL^H. \qquad \blacksquare$$

A differenza della fattorizzazione LU e LL^H, la fattorizzazione QR di una matrice A non è unica. Infatti per ogni matrice S diagonale e unitaria (e quindi con elementi principali di modulo 1), detta *matrice di fase*,

$$S = \begin{bmatrix} \theta_1 & & & \\ & \theta_2 & & \\ & & \ddots & \\ & & & \theta_n \end{bmatrix}, \quad |\theta_i| = 1,$$

risulta

$$QR = QSS^H R = Q'R',$$

in cui $Q' = QS$ è unitaria e $R' = S^H R$ è triangolare superiore. Però se A è non singolare esiste un'unica matrice di fase S tale che gli elementi principali di R' siano reali e positivi, e si può dimostrare che se QR e $Q'R'$ sono due fattorizzazioni di A, esiste una matrice di fase S tale che $Q' = QS$ e $R' = S^H R$ (si veda l'esercizio 4.37).

La determinazione delle matrici della fattorizzazione di A viene generalmente effettuata nei due modi seguenti:

a) applicando alla matrice A una successione di matrici elementari (metodo di Gauss, metodo di Householder);

b) con "tecniche compatte" (metodo di Cholesky, metodo di Crout per la fattorizzazione LU).

4. Matrici elementari

In questo paragrafo si introduce la classe delle matrici elementari e se ne analizzano le principali proprietà.

4.8 Definizione. Siano $\sigma \in \mathbf{C}$ e $\mathbf{u}, \mathbf{v} \in \mathbf{C}^n$, $\mathbf{u}, \ \mathbf{v} \neq \mathbf{0}$. Si definisce *matrice elementare* una matrice di ordine n della forma

$$E(\sigma, \mathbf{u}, \mathbf{v}) = I - \sigma \mathbf{u}\mathbf{v}^H. \qquad \blacksquare$$

La classe delle matrici elementari non singolari è chiusa rispetto all'operazione di inversione. Vale infatti il seguente teorema.

4.9 Teorema. *Ogni matrice elementare $E(\sigma, \mathbf{u}, \mathbf{v})$ per cui $\sigma \mathbf{v}^H \mathbf{u} \neq 1$ è invertibile e la sua inversa è ancora una matrice elementare della forma $E(\tau, \mathbf{u}, \mathbf{v})$, $\tau \in \mathbf{C}$.*

Dim. Se $\sigma = 0$, la tesi è ovvia. Se $\sigma \neq 0$, si dimostra che esiste τ tale che $E(\tau, \mathbf{u}, \mathbf{v})$ è la matrice inversa di $E(\sigma, \mathbf{u}, \mathbf{v})$, ossia

$$(I - \sigma \mathbf{u}\mathbf{v}^H)\,(I - \tau \mathbf{u}\mathbf{v}^H) = I.$$

Sviluppando si ha

$$(\sigma + \tau - \sigma\tau \mathbf{v}^H \mathbf{u})\, \mathbf{u}\mathbf{v}^H = O,$$

da cui si ottiene che il parametro τ deve verificare la relazione

$$\mathbf{v}^H\mathbf{u} = \frac{1}{\sigma} + \frac{1}{\tau}, \tag{11}$$

e quindi la matrice $E(\sigma, \mathbf{u}, \mathbf{v})$ è invertibile se $\mathbf{v}^H\mathbf{u} \neq \dfrac{1}{\sigma}$. ∎

Assegnati comunque due vettori non nulli, esiste sempre una matrice elementare non singolare che trasforma il primo vettore nel secondo. Vale infatti il seguente

4.10 Teorema. *Siano* $\mathbf{x}, \mathbf{y} \in \mathbf{C}^n, \mathbf{x} \neq \mathbf{0}, \mathbf{y} \neq \mathbf{0}$. *Esistono matrici elementari non singolari* $E(\sigma, \mathbf{u}, \mathbf{v})$ *tali che*

$$E(\sigma, \mathbf{u}, \mathbf{v})\mathbf{x} = \mathbf{y}.$$

Dim. La condizione

$$(I - \sigma\mathbf{u}\mathbf{v}^H)\mathbf{x} = \mathbf{y}$$

è verificata se $\mathbf{v}$ è un vettore tale che $\mathbf{v}^H\mathbf{x} \neq 0$, e il vettore $\mathbf{u}$ e il numero σ sono tali che

$$\sigma\mathbf{u} = \frac{(\mathbf{x} - \mathbf{y})}{\mathbf{v}^H\mathbf{x}}.$$

Se inoltre $\mathbf{v}^H\mathbf{y} \neq 0$, poiché

$$1 - \sigma\mathbf{v}^H\mathbf{u} = \frac{\mathbf{v}^H\mathbf{y}}{\mathbf{v}^H\mathbf{x}},$$

la matrice $E(\sigma, \mathbf{u}, \mathbf{v})$ è non singolare. ∎

Due classi importanti di matrici elementari sono le matrici di Gauss e le matrici di Householder.

a) Matrici elementari di Gauss

Sia $\mathbf{x} \in \mathbf{C}^n$, con $x_1 \neq 0$. Si vuole determinare una matrice

$$M = E(\sigma, \mathbf{u}, \mathbf{e}_1) = I - \sigma\mathbf{u}\mathbf{e}_1^H$$

per cui

$$M\mathbf{x} = x_1\mathbf{e}_1,$$

cioè tale che trasformi il vettore $\mathbf{x}$ in un vettore con tutte le componenti nulle, eccetto la prima che resta invariata. Per il teorema 4.10 si ha che ciò è possibile in quanto

$$\mathbf{e}_1^H\mathbf{x} = x_1 \neq 0$$

e

$$\sigma\mathbf{u} = \left[0, \frac{x_2}{x_1}, \ldots, \frac{x_n}{x_1}\right]^T.$$

La matrice M è perciò

$$M = \begin{bmatrix} 1 & & & \\ -m_{21} & 1 & & \\ \vdots & & \ddots & \\ -m_{n1} & & & 1 \end{bmatrix},$$

in cui $m_{i1} = \dfrac{x_i}{x_1}$ per $i = 2, \ldots, n$. Poiché $\sigma\mathbf{e}_1^H\mathbf{u} = 0$, la matrice M è invertibile e la sua inversa è

$$M^{-1} = \begin{bmatrix} 1 & & & \\ m_{21} & 1 & & \\ \vdots & & \ddots & \\ m_{n1} & & & 1 \end{bmatrix}. \tag{12}$$

b) Matrici elementari di Householder

Una matrice elementare hermitiana

$$P = I - \beta\mathbf{v}\mathbf{v}^H,$$

con $\beta \in \mathbf{R}$ e $\mathbf{v} \in \mathbf{C}^n$, $\mathbf{v} \neq \mathbf{0}$, è detta *matrice di Householder* se è unitaria, cioè se $P^HP = PP^H = I$. Imponendo la condizione che P sia unitaria, si ottiene dalla (11) che se $\beta \neq 0$ allora

$$\beta = \frac{2}{\|\mathbf{v}\|_2^2}. \tag{13}$$

Le matrici di Householder vengono anche chiamate *matrici di riflessione*. Infatti se S è il sottospazio generato dal vettore $\mathbf{w} = \mathbf{v}/\|\mathbf{v}\|_2$, ed $\mathbf{x}$ è un vettore di $\mathbf{C}^n$, decomposto $\mathbf{x}$ come

$$\mathbf{x} = \gamma\mathbf{w} + \mathbf{y}, \quad \text{dove } \gamma \in \mathbf{C},\ \mathbf{y} \in S^\perp,$$

cioè $\mathbf{w}^H\mathbf{y} = 0$, si ha

$$\begin{aligned} P\mathbf{x} &= (I - 2\mathbf{w}\mathbf{w}^H)\,\mathbf{x} = (I - 2\mathbf{w}\mathbf{w}^H)\,(\gamma\mathbf{w} + \mathbf{y}) \\ &= \gamma\mathbf{w} + \mathbf{y} - 2\gamma\mathbf{w}\mathbf{w}^H\mathbf{w} - 2\mathbf{w}\mathbf{w}^H\mathbf{y} = -\gamma\mathbf{w} + \mathbf{y}. \end{aligned}$$

Il caso $\mathbf{x} \in \mathbf{R}^3$ è illustrato nella figura 4.1.

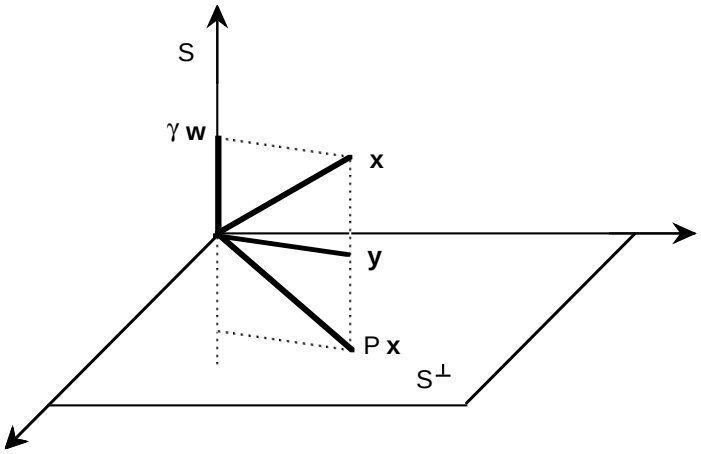


Fig. 4.1 - Riflessione di un vettore $\mathbf{x}$.

Per ogni vettore $\mathbf{x} \in \mathbf{C}^n$, con $\mathbf{x} \neq \mathbf{0}$, si può determinare una matrice elementare di Householder P tale che

$$P\mathbf{x} = \alpha \mathbf{e}_1, \tag{14}$$

dove α è un'opportuna costante.

Poiché P è unitaria, dalla (14) risulta

$$\|\mathbf{x}\|_2 = \|P\mathbf{x}\|_2 = |\alpha|, \tag{15}$$

e poiché P è hermitiana, risulta $\mathbf{x}^H P\mathbf{x} \in \mathbf{R}$, ossia $\mathbf{x}^H \alpha \mathbf{e}_1 \in \mathbf{R}$, cioè il prodotto di α per $\overline{x}_1$ è reale. Quindi, posto

$$\theta = \begin{cases} \dfrac{x_1}{|x_1|} & \text{se } x_1 \neq 0, \\ 1 & \text{se } x_1 = 0, \end{cases}$$

per la (15) è

$$\alpha = \pm \|\mathbf{x}\|_2 \, \theta. \tag{16}$$

Inoltre si ha:

$$(I - \beta \mathbf{v}\mathbf{v}^H)\, \mathbf{x} = \alpha \mathbf{e}_1,$$

da cui

$$(\beta \mathbf{v}^H \mathbf{x})\, \mathbf{v} = \mathbf{x} - \alpha \mathbf{e}_1.$$

Questa relazione è verificata scegliendo $\mathbf{v} = \mathbf{x} - \alpha \mathbf{e}_1$, e, per la (13)

$$\beta = \frac{2}{\|\mathbf{x} - \alpha \mathbf{e}_1\|_2^2}.$$

La prima componente del vettore $\mathbf{v}$ è

$$v_1 = x_1 - \alpha = -[-|x_1| \pm \|\mathbf{x}\|_2]\ \theta, \tag{17}$$

per cui nella (16) conviene scegliere il segno negativo per evitare il rischio che nella (17) si produca un errore di cancellazione connesso con l'operazione di sottrazione di due numeri positivi. Si pone quindi

$$\alpha = -\|\mathbf{x}\|_2\ \theta,$$

e si ha:

$$\begin{aligned}\|\mathbf{v}\|_2^2 = \|\mathbf{x} - \alpha\mathbf{e}_1\|_2^2 = (\mathbf{x} - \alpha\mathbf{e}_1)^H(\mathbf{x} - \alpha\mathbf{e}_1) = 2\,(\|\mathbf{x}\|_2^2 - \alpha\overline{x}_1)\\ = 2\|\mathbf{x}\|_2\ (\|\mathbf{x}\|_2 + \overline{x}_1\theta) = 2\|\mathbf{x}\|_2\ (\|\mathbf{x}\|_2 + |x_1|),\end{aligned}$$

$$\beta = \frac{1}{\|\mathbf{x}\|_2(\|\mathbf{x}\|_2 + |x_1|)}$$

e

$$v_1 = x_1 - \alpha = x_1 + \|\mathbf{x}\|_2\ \theta = \theta\ (|x_1| + \|\mathbf{x}\|_2).$$

Quindi

$$\mathbf{v} = \begin{bmatrix} \theta(|x_1| + \|\mathbf{x}\|_2) \\ x_2 \\ \vdots \\ x_n \end{bmatrix}.$$

Se il vettore $\mathbf{x}$ è reale, anche il vettore $\mathbf{v}$ è reale:

$$\mathbf{v} = \begin{bmatrix} \mathrm{sgn}(x_1)\ (|x_1| + \|\mathbf{x}\|_2) \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

e quindi la matrice P risulta una matrice reale, simmetrica e ortogonale.

4.11 Esempio. Si consideri il vettore $\mathbf{x} = [4, 7, 4]^T$. La matrice elementare di Gauss che trasforma il vettore $\mathbf{x}$ nel vettore $x_1\mathbf{e}_1$ è data da:

$$M = \begin{bmatrix} 1 & 0 & 0 \\ -\frac{7}{4} & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}$$

e si ha $M\mathbf{x} = [4, 0, 0]^T$. La matrice elementare di Householder che trasforma $\mathbf{x}$ nel vettore $\alpha\mathbf{e}_1$ è

$$P = I - \beta\mathbf{v}\mathbf{v}^T,$$

dove $\beta = \dfrac{1}{117}$ e $\mathbf{v} = [13, 7, 4]^T$. Quindi

$$P = I - \frac{1}{117}\begin{bmatrix} 169 & 91 & 52 \\ 91 & 49 & 28 \\ 52 & 28 & 16 \end{bmatrix} = \frac{1}{117}\begin{bmatrix} -52 & -91 & -52 \\ -91 & 68 & -28 \\ -52 & -28 & 101 \end{bmatrix}$$

e si ha $P\mathbf{x} = [9, 0, 0]^T$. ∎

5. Fattorizzazione mediante matrici elementari

Sfruttando le proprietà delle matrici elementari di trasformare un qualunque vettore non nullo in un vettore con al più una componente diversa da zero, è possibile trasformare in forma triangolare superiore una matrice, moltiplicandola successivamente per opportune matrici elementari.

Sia A una matrice di ordine n; posto $A^{(1)} = A$, $A^{(1)}$ può essere così partizionata

$$A^{(1)} = [\,\mathbf{a}_1 \,|\, B\,],$$

in cui $\mathbf{a}_1$ è il vettore formato dagli elementi della prima colonna di A. Sia $E^{(1)}$ una matrice elementare di ordine n non singolare che trasforma il vettore $\mathbf{a}_1$ nel vettore

$$\mathbf{b}_1 = E^{(1)}\mathbf{a}_1$$

che ha nulle tutte le componenti di indice maggiore di 1. Moltiplicando $E^{(1)}$ per $A^{(1)}$, si ottiene una matrice $A^{(2)}$ della forma

$$A^{(2)} = E^{(1)}A^{(1)} = [\,\mathbf{b}_1 \,|\, E^{(1)}B\,],$$

cioè una matrice la cui prima colonna ha nulli tutti gli elementi con indice di riga maggiore di 1. Si può allora rappresentare la matrice $A^{(2)}$ nella forma

$$A^{(2)} = \begin{bmatrix} \alpha & \mathbf{c}^H \\ \mathbf{0} & B^{(2)} \end{bmatrix} \begin{matrix} \} & 1 \text{ riga} \\ \} & n-1 \text{ righe} \end{matrix}$$

dove $\alpha \in \mathbf{C}$ e $B^{(2)} \in \mathbf{C}^{(n-1)\times(n-1)}$.

Applicando in modo analogo $n-1$ volte il procedimento descritto, si ottiene una successione di matrici $A^{(k)}$, $k = 2, \ldots, n$, tali che la matrice $A^{(k)}$ ha nulli gli elementi delle prime $k-1$ colonne che si trovano al di sotto della diagonale principale.

Al k-esimo passo si opera nel modo seguente: la matrice $A^{(k)}$ è della forma:

$$A^{(k)} = \begin{bmatrix} C^{(k)} & D^{(k)} \\ O & B^{(k)} \end{bmatrix} \begin{matrix} \} & k-1 \text{ righe} \\ \} & n-k+1 \text{ righe,} \end{matrix} \tag{18}$$

in cui $B^{(k)} \in \mathbf{C}^{(n-k+1)\times(n-k+1)}$ e $C^{(k)}$ è triangolare superiore. Applicando il procedimento sopra descritto alla matrice $B^{(k)}$, si determina una matrice elementare non singolare $F^{(k)} \in \mathbf{C}^{(n-k+1)\times(n-k+1)}$, tale che la matrice $F^{(k)}\ B^{(k)}$ sia della forma:

$$F^{(k)}\ B^{(k)} = \begin{bmatrix} \beta & \mathbf{d}^H \\ \mathbf{0} & B^{(k+1)} \end{bmatrix} \begin{matrix} \} & 1 \text{ riga} \\ \} & n-k \text{ righe,} \end{matrix}$$

dove $\beta \in \mathbf{C}$ e $B^{(k+1)} \in \mathbf{C}^{(n-k)\times(n-k)}$. La matrice

$$E^{(k)} = \begin{bmatrix} I_{(k-1)} & O \\ O & F^{(k)} \end{bmatrix} \tag{19}$$

è ancora una matrice elementare: infatti se

$$F^{(k)} = I - \sigma \mathbf{u}\mathbf{v}^H, \quad \text{con} \quad \mathbf{u}, \mathbf{v} \in \mathbf{C}^{(n-k+1)},$$

si ha

$$E^{(k)} = I - \sigma \mathbf{t}\mathbf{z}^H,$$

con

$$\mathbf{t} = \begin{bmatrix} \mathbf{0} \\ \mathbf{u} \end{bmatrix} \begin{matrix} \} & k-1 \text{ componenti} \\ \} & n-k+1 \text{ componenti,} \end{matrix} \qquad \mathbf{z} = \begin{bmatrix} \mathbf{0} \\ \mathbf{v} \end{bmatrix} \begin{matrix} \} & k-1 \text{ componenti} \\ \} & n-k+1 \text{ componenti.} \end{matrix}$$

Moltiplicando $E^{(k)}$ per $A^{(k)}$ si ottiene

$$A^{(k+1)} = E^{(k)}A^{(k)} = \begin{bmatrix} C^{(k)} & D^{(k)} \\ O & \begin{bmatrix} \beta & \mathbf{d}^H \\ \mathbf{0} & B^{(k+1)} \end{bmatrix} \end{bmatrix}$$

$$= \begin{bmatrix} C^{(k+1)} & D^{(k+1)} \\ O & B^{(k+1)} \end{bmatrix} \begin{matrix} \} & k \text{ righe} \\ \} & n-k \text{ righe,} \end{matrix}$$

in cui $C^{(k+1)}$ è ancora triangolare superiore. All'$(n-1)$-esimo passo si ottiene una matrice $A^{(n)}$ della forma:

$$A^{(n)} = \begin{bmatrix} C^{(n)} & \mathbf{g} \\ \mathbf{0}^H & \gamma \end{bmatrix} \begin{matrix} \} \; n-1 \text{ righe} \\ \} \quad 1 \text{ riga}, \end{matrix}$$

in cui $\mathbf{g} \in \mathbf{C}^{n-1}$ e $\gamma \in \mathbf{C}$. Quindi $A^{(n)}$ è triangolare superiore. Le matrici $A = A^{(1)}, A^{(2)}, \ldots, A^{(n)}$, risultano così legate dalla relazione

$$A^{(k+1)} = E^{(k)} A^{(k)}, \quad k = 1, \ldots, n-1. \tag{20}$$

Dalla (20), poiche $E^{(k)}$ è non singolare, si ha:

$$A^{(k)} = [E^{(k)}]^{-1} A^{(k+1)}, \quad k = 1, \ldots, n-1,$$

e

$$A = A^{(1)} = [E^{(1)}]^{-1} A^{(2)} = \ldots = [E^{(1)}]^{-1} \ldots [E^{(n-1)}]^{-1} A^{(n)} = E A^{(n)}, \tag{21}$$

dove $E = [E^{(1)}]^{-1} \ldots [E^{(n-1)}]^{-1}$.

Si è così ottenuta una fattorizzazione di A nel prodotto di una matrice E per una matrice $A^{(n)}$ triangolare superiore, dove la forma della matrice E dipende dalle particolari matrici elementari $E^{(k)}$ usate.

Il procedimento descritto può essere applicato anche se la matrice A non è quadrata. Se $A \in \mathbf{C}^{m \times n}$, $m > n$, le matrici elementari $E^{(k)}$, $k = 1, \ldots, n$, sono di ordine m, e dopo n passi risulta

$$A = E A^{(n+1)},$$

in cui $E = [E^{(1)}]^{-1} \ldots [E^{(n)}]^{-1} \in \mathbf{C}^{m \times m}$ e $A^{(n+1)} \in \mathbf{C}^{m \times n}$ può essere così rappresentata

$$A^{(n+1)} = \begin{bmatrix} T \\ O \end{bmatrix} \begin{matrix} \} \quad n \text{ righe} \\ \} \; m-n \text{ righe}, \end{matrix}$$

dove $T \in \mathbf{C}^{n \times n}$ è una matrice triangolare superiore.

La fattorizzazione di A nel prodotto $EA^{(n)}$ può essere ottenuta solo se tutte le $E^{(k)}$ sono non singolari. Questo è sempre vero se le $E^{(k)}$ sono matrici elementari di Householder, mentre nel caso delle matrici elementari di Gauss le $E^{(k)}$ possono non esistere, per cui non sempre è possibile completare la fattorizzazione.

6. Il metodo di Gauss per la fattorizzazione *LU*

Il procedimento descritto nel paragrafo precedente, quando si utilizzano le matrici elementari di Gauss è detto *metodo (di eliminazione) di Gauss*. Gli elementi delle matrici $A^{(k)}$ vengono indicati con la notazione consueta $a_{rs}^{(k)},\ r,s=1,\dots,n$.

Al primo passo, posto $A^{(1)}=A$, se $a_{11}^{(1)}\neq 0$, si considera il vettore

$$\mathbf{m}^{(1)}=[0,m_{21},\dots,m_{n1}]^T,$$

dove $m_{r1}=a_{r1}^{(1)}\ /\ a_{11}^{(1)},\ r=2,\dots,n$. La prima matrice elementare è data da

$$E^{(1)}=E(1,\mathbf{m}^{(1)},\mathbf{e}_1)=M^{(1)}=\begin{bmatrix}1 & & & \\ -m_{21} & 1 & & \\ \vdots & & \ddots & \\ -m_{n1} & & & 1\end{bmatrix}.$$

Al k-esimo passo, se $a_{kk}^{(k)}\neq 0$, indicato con $\mathbf{m}^{(k)}, k=1,\dots,n$, il vettore

$$\mathbf{m}^{(k)}=[\ \underbrace{0,\dots,0}_{k\text{ componenti}},m_{k+1,k},\dots,m_{nk}]^T,$$

dove $m_{rk}=a_{rk}^{(k)}\ /\ a_{kk}^{(k)},\ r=k+1,\dots,n$, la k-esima matrice elementare per la (19) è data da

$$E^{(k)}=E(1,\mathbf{m}^{(k)},\mathbf{e}_k)=M^{(k)}=\begin{bmatrix}1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & -m_{k+1,k} & & \\ & & \vdots & \ddots & \\ & & -m_{nk} & & 1\end{bmatrix}.$$

Risulta, come nella (12), che

$$[M^{(k)}]^{-1}=I+\mathbf{m}^{(k)}\mathbf{e}_k^T=\begin{bmatrix}1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & m_{k+1,k} & & \\ & & \vdots & \ddots & \\ & & m_{nk} & & 1\end{bmatrix}.$$

Perciò la matrice $E=[M^{(1)}]^{-1}\dots[M^{(n-1)}]^{-1}$, prodotto di matrici triangolari inferiori con elementi principali uguali ad 1 è ancora una matrice

triangolare inferiore con elementi principali uguali ad 1. Per l'unicità della fattorizzazione LU, dalla (21) si ha allora che

$$L = E \quad \text{e} \quad U = A^{(n)}.$$

Poiché $\mathbf{m}^{(r)}\mathbf{e}_r^T\mathbf{m}^{(s)}\mathbf{e}_s^T = 0$ per $r < s$, ne segue che la matrice L ha la forma:

$$L = \begin{bmatrix} 1 & & & & & \\ m_{21} & \ddots & & & & \\ \vdots & \ddots & 1 & & & \\ \vdots & & m_{k+1,k} & \ddots & & \\ \vdots & & \vdots & \ddots & \ddots & \\ m_{n1} & \cdots & m_{nk} & \cdots & m_{n,n-1} & 1 \end{bmatrix}.$$

4.12 Teorema. *Se la matrice A soddisfa alle ipotesi del teorema 4.4 (quindi esiste la fattorizzazione LU di A), allora il metodo di Gauss è applicabile.*

Dim. Si dimostra che $a_{kk}^{(k)} \neq 0$. Per $k = 1$ è $a_{11}^{(1)} = a_{11} \neq 0$ per ipotesi. Per $k > 1$, la sottomatrice principale di testa di ordine k della matrice $A^{(k)}$ è triangolare superiore e per la (18) il suo determinante è dato da $a_{kk}^{(k)} \det C^{(k)}$. Poiché questa sottomatrice è uguale al prodotto delle corrispondenti sottomatrici principali di testa delle matrici $M^{(k-1)}, \ldots, M^{(1)}, A$, il determinante della sottomatrice principale di testa di ordine k di A e il corrispondente di $A^{(k)}$ coincidono (infatti il determinante di una matrice elementare di Gauss, come di ogni sua sottomatrice principale, è uguale a 1). Ne segue che $a_{kk}^{(k)} \neq 0$. ∎

4.13 Esempio. Sia A la matrice tridiagonale di ordine n:

$$A = \begin{bmatrix} a_1 & c_1 & & & \\ b_1 & a_2 & c_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & c_{n-1} \\ & & & b_{n-1} & a_n \end{bmatrix}.$$

Posto

$$\left.\begin{array}{l} \alpha_1 = a_1 \\ \beta_i = b_i/\alpha_i \\ \alpha_{i+1} = a_{i+1} - \beta_i c_i \end{array}\right\} \quad \text{per } i = 1, \ldots, n-1,$$

se $\alpha_i \neq 0$ per $i = 1, \ldots, n-1$, si ha:

$$LU = \begin{bmatrix} 1 & & & & \\ \beta_1 & 1 & & & \\ & \beta_2 & \ddots & & \\ & & \ddots & \ddots & \\ & & & \beta_{n-1} & 1 \end{bmatrix} \begin{bmatrix} \alpha_1 & c_1 & & & \\ & \alpha_2 & c_2 & & \\ & & \ddots & \ddots & \\ & & & \ddots & c_{n-1} \\ & & & & \alpha_n \end{bmatrix}.$$

■

Il numero delle operazioni moltiplicative richieste al k-esimo passo del metodo di Gauss è dato dal numero di operazioni moltiplicative occorrenti per costruire la matrice $M^{(k)}$ e per moltiplicare le due matrici $M^{(k)}$ e $A^{(k)}$: la costruzione di $M^{(k)}$ richiede il calcolo degli $n-k$ elementi m_{rk}, mentre per moltiplicare le due matrici occorrono $(n-k)^2$ operazioni. Quindi, a meno di termini di ordine inferiore, al k-esimo passo occorrono $(n-k)^2$ operazioni e allora, per gli $n-1$ passi richiesti dalla fattorizzazione, il costo computazionale del metodo di Gauss è dato da

$$\sum_{k=1}^{n-1}(n-k)^2 = \sum_{k=1}^{n-1} k^2 \simeq \frac{n^3}{3}.$$

La fattorizzazione LU di una matrice tridiagonale, come nell'esempio 4.13, richiede $2n-2$ operazioni moltiplicative.

7. Il metodo di Gauss per la risoluzione del sistema lineare

Si utilizza il metodo di Gauss per la risoluzione del sistema lineare $A\mathbf{x} = \mathbf{b}$, se A soddisfa alle ipotesi del teorema 4.4. La soluzione del sistema

$$L\mathbf{y} = \mathbf{b}$$

viene calcolata durante i passi del procedimento della fattorizzazione LU, in quanto il vettore $\mathbf{y}$ viene costruito moltiplicando successivamente per le matrici $M^{(k)}$ il vettore $\mathbf{b}$ così come si fa con la matrice A. Per questo si considera la matrice

$$[\, A^{(1)} |\, \mathbf{b}^{(1)} \,] = [\, A \mid \mathbf{b} \,]$$

e si costruisce la successione

$$[\, A^{(1)} \mid \mathbf{b}^{(1)} \,], [\, A^{(2)} |\, \mathbf{b}^{(2)} \,],\ \ldots,\ [\, A^{(n)} |\, \mathbf{b}^{(n)} \,] = [\, U \mid \mathbf{y} \,]$$

tale che

$$[\, A^{(k+1)} \mid \mathbf{b}^{(k+1)} \,] = M^{(k)} [\, A^{(k)} \mid \mathbf{b}^{(k)} \,], \quad k = 1, \ldots, n-1.$$

Data la struttura della matrice $M^{(k)}$, la moltiplicazione per $M^{(k)}$ corrisponde a operare sulla matrice $[\, A^{(k)} \mid \mathbf{b}^{(k)} \,]$ delle combinazioni lineari di righe: esattamente la i-esima riga, $i > k$, viene sostituita dalla differenza della stessa riga con la k-esima riga moltiplicata per il fattore $m_{ik} = a_{ik}^{(k)}/a_{kk}^{(k)}$:

$$\left.\begin{aligned} a_{ij}^{(k+1)} &= a_{ij}^{(k)} - m_{ik} a_{kj}^{(k)},\ j = k, \ldots, n, \\ b_i^{(k+1)} &= b_i^{(k)} - m_{ik} b_k^{(k)}, \end{aligned}\right\} \quad i = k+1, \ldots, n. \tag{22}$$

Il sistema lineare che si ottiene al k-esimo passo

$$A^{(k)}\mathbf{x} = \mathbf{b}^{(k)} \tag{23}$$

è equivalente a quello iniziale $A\mathbf{x} = \mathbf{b}$, e per la forma della matrice $A^{(k)}$ la componente x_j, $j < k$, del vettore delle incognite $\mathbf{x}$ è presente solo nelle prime j equazioni e non nelle successive. Il metodo di Gauss consiste quindi nell'eliminare passo per passo le incognite $x_1, x_2, \ldots, x_{n-1}$ dalle equazioni successive rispettivamente alla prima, seconda, ..., $(n-1)$-esima. Per questo il metodo di Gauss è detto anche *metodo di eliminazione*.

4.14 Esempio. È dato il sistema $A\mathbf{x} = \mathbf{b}$, dove

$$A = \begin{bmatrix} \text{-}2 & 4 & \text{-}1 & \text{-}1 \\ 4 & \text{-}9 & 0 & 5 \\ \text{-}4 & 5 & \text{-}5 & 5 \\ \text{-}8 & 8 & \text{-}23 & 20 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 12 \\ \text{-}32 \\ 3 \\ \text{-}13 \end{bmatrix}.$$

La risoluzione con il metodo di Gauss procede nel modo seguente: si pone

$$[\, A^{(1)} \mid \mathbf{b}^{(1)} \,] = [\, A \mid \mathbf{b} \,] = \left[\begin{array}{cccc|c} \text{-}2 & 4 & \text{-}1 & \text{-}1 & 12 \\ 4 & \text{-}9 & 0 & 5 & \text{-}32 \\ \text{-}4 & 5 & \text{-}5 & 5 & 3 \\ \text{-}8 & 8 & \text{-}23 & 20 & \text{-}13 \end{array}\right].$$

Al primo passo l'elemento $a_{11}^{(1)} = -2$ è diverso da zero e i fattori per le combinazioni lineari sono $m_{21} = a_{21}^{(1)}/a_{11}^{(1)} = -2$, $m_{31} = a_{31}^{(1)}/a_{11}^{(1)} = 2$, $m_{41} = a_{41}^{(1)}/a_{11}^{(1)} = 4$. Quindi alla seconda riga viene sommata la prima riga moltiplicata per 2, alla terza riga viene sommata la prima riga moltiplicata per -2, alla quarta riga viene sommata la prima riga moltiplicata per -4 e si ottiene

$$[\, A^{(2)} \mid \mathbf{b}^{(2)} \,] = \left[\begin{array}{cccc|c} \text{-}2 & 4 & \text{-}1 & \text{-}1 & 12 \\ 0 & \text{-}1 & \text{-}2 & 3 & \text{-}8 \\ 0 & \text{-}3 & \text{-}3 & 7 & \text{-}21 \\ 0 & \text{-}8 & \text{-}19 & 24 & \text{-}61 \end{array}\right].$$

Al secondo passo l'elemento $a_{22}^{(2)} = -1$ è ancora diverso da zero e i fattori per le combinazioni lineari sono $m_{32} = a_{32}^{(2)}/a_{22}^{(2)} = 3$, $m_{42} = a_{42}^{(2)}/a_{22}^{(2)} = 8$. Quindi alla terza riga viene sommata la seconda riga moltiplicata per -3 e alla quarta riga viene sommata la seconda riga moltiplicata per -8 e si ottiene

$$[\, A^{(3)} \mid \mathbf{b}^{(3)} \,] = \left[\begin{array}{cccc|c} \text{-2} & 4 & \text{-1} & \text{-1} & 12 \\ 0 & \text{-1} & \text{-2} & 3 & \text{-8} \\ 0 & 0 & 3 & \text{-2} & 3 \\ 0 & 0 & \text{-3} & 0 & 3 \end{array}\right].$$

Al terzo passo l'elemento $a_{33}^{(3)} = 3$ è ancora diverso da zero e vi è una sola combinazione lineare da fare, con $m_{43} = a_{43}^{(3)}/a_{33}^{(3)} = -1$. Quindi alla quarta riga viene sommata la terza riga e si ottiene

$$[\, A^{(4)} \mid \mathbf{b}^{(4)} \,] = \left[\begin{array}{cccc|c} \text{-2} & 4 & \text{-1} & \text{-1} & 12 \\ 0 & \text{-1} & \text{-2} & 3 & \text{-8} \\ 0 & 0 & 3 & \text{-2} & 3 \\ 0 & 0 & 0 & \text{-2} & 6 \end{array}\right] = [\, U \mid \mathbf{y} \,].$$

Risolvendo il sistema lineare $U\mathbf{x} = \mathbf{y}$ con il procedimento di sostituzione all'indietro si ottiene la soluzione $\mathbf{x} = [-2, 1, -1, -3]^T$. ∎

L'elemento $a_{kk}^{(k)}$ della matrice $A^{(k)}$, detto *pivot* al k-esimopasso, per l'ipotesi fatta che la sottomatrice principale di testa di ordine k sia non singolare, è diverso da zero. Il metodo di Gauss però è applicabile anche nel caso in cui la matrice non singolare A non verifica le ipotesi del teorema 4.4, se si utilizza la *variante del pivot*. Infatti, se al k-esimo passo risulta $a_{kk}^{(k)} = 0$, per l'ipotesi della non singolarità di A esiste almeno una riga di indice $j > k$, con l'elemento $a_{jk}^{(k)} \neq 0$; basta allora scambiare la k-esima riga della matrice $[\, A \mid \mathbf{b} \,]$ con la j-esima, in modo da portare nella posizione del pivot un elemento non nullo. Si osservi che l'operazione di scambio di due righe della matrice $[\, A \mid \mathbf{b} \,]$ può essere anche descritta mediante la moltiplicazione per una matrice di permutazione Π, trasformando il sistema lineare nel sistema equivalente $\Pi A\mathbf{x} = \Pi\mathbf{b}$. La fattorizzazione della matrice A che corrisponde alla variante del pivot è del tipo $\Pi A = LU$, la cui esistenza è stata provata nel teorema 4.5.

Se la matrice A è singolare e il sistema è consistente, allora il metodo di Gauss con la variante del pivot è ancora applicabile. Infatti se al k-esimo passo risulta $a_{kk}^{(k)} = 0$ e tutti gli elementi della k-esima colonna di $A^{(k)}$, al di sotto di quello principale sono nulli, si assume

$$[\, A^{(k+1)} \mid \mathbf{b}^{(k+1)} \,] = [\, A^{(k)} \mid \mathbf{b}^{(k)} \,],$$

ossia il k-esimo passo non comporta alcuna operazione, e si continua con la colonna successiva. La matrice $A^{(n)}$, ottenuta al termine del procedimento, ha l'elemento $a_{kk}^{(n)}$ nullo, ma per l'ipotesi di consistenza del sistema si può procedere ugualmente al calcolo della soluzione mediante sostituzione all'indietro.

Anche nel caso che $A \in \mathbf{C}^{m\times n}$, $m > n$, cioè quando la matrice A non è quadrata, se il sistema è consistente il metodo di Gauss con la variante del pivot è ancora applicabile. In questo caso dopo n passi si ottiene il sistema equivalente

$$A^{(n+1)}\mathbf{x} = \mathbf{b}^{(n+1)},$$

dove

$$A^{(n+1)} = \begin{bmatrix} T \\ \\ O \end{bmatrix} \begin{matrix} \} & n \text{ righe} \\ \\ \} & m-n \text{ righe,} \end{matrix} \qquad \mathbf{b}^{(n+1)} = \begin{bmatrix} \mathbf{c} \\ \\ \mathbf{0} \end{bmatrix} \begin{matrix} \} & n \text{ componenti} \\ \\ \} & m-n \text{ componenti,} \end{matrix}$$

e T è triangolare superiore. La soluzione $\mathbf{x}$ viene calcolata risolvendo il sistema $T\mathbf{x} = \mathbf{c}$.

Il costo computazionale del metodo di Gauss per la risoluzione di un sistema lineare è uguale, a meno di termini di ordine inferiore, al costo della fattorizzazione LU di A, cioè $n^3/3$ operazioni. Infatti l'aggiunta della colonna $\mathbf{b}$ alla matrice A e la successiva risoluzione del sistema triangolare, comportano un numero di operazioni dell'ordine di n^2. Naturalmente il costo computazionale può essere molto più basso se la matrice del sistema ha qualche struttura particolare: ad esempio, nel caso di un sistema con matrice tridiagonale, al costo della fattorizzazione che, come si è visto, richiede $2n-2$ operazioni moltiplicative, vanno aggiunte altre n operazioni per le combinazioni lineari sugli elementi di $\mathbf{b}$ e $2n$ operazioni per la risoluzione del sistema $U\mathbf{x} = \mathbf{y}$. In totale il costo computazionale è di $5n$ operazioni.

Il metodo di Gauss può essere utilizzato per la risoluzione contemporanea di più sistemi lineari con la stessa matrice dei coefficienti A e diverse colonne di termini noti. Sia infatti $B \in \mathbf{C}^{n\times r}$ la matrice formata da r colonne di termini noti. Allora la matrice $X \in \mathbf{C}^{n\times r}$, soluzione del sistema

$$AX = B,$$

si ottiene costruendo la successione

$$[\,A^{(1)} \mid B^{(1)}] = [\,A \mid B\,],\ [\,A^{(2)} \mid B^{(2)}\,],\ \ldots\ ,\ [\,A^{(n)} \mid B^{(n)}\,],$$

tale che

$$[\,A^{(k+1)} \mid B^{(k+1)}\,] = M^{(k)}[\,A^{(k)} \mid B^{(k)}\,],$$

e risolvendo al termine i sistemi

$$A^{(n)}X = B^{(n)},$$

dove $A^{(n)}$ è una matrice triangolare superiore, con formule analoghe a quelle usate nel caso di una sola colonna di termini noti. Complessivamente il costo computazionale è dato da:

a) al k-esimo passo, per il calcolo di $[\, A^{(k+1)} \,|\, B^{(k+1)} \,]$ occorrono $(n-k)(n-k+r)$ operazioni moltiplicative (ad $A^{(k)}$ sono state infatti affiancate r colonne, mentre il numero di righe è rimasto invariato); per gli $n-1$ passi richiesti il costo computazionale, a meno di termini di ordine inferiore, è dato da

$$\sum_{k=1}^{n-1}(n-k)(n-k+r) = \sum_{k=1}^{n-1}k^2 + r\sum_{k=1}^{n-1}k \simeq \frac{n^3}{3} + r\frac{n^2}{2};$$

b) per la risoluzione degli r sistemi lineari la cui matrice è triangolare superiore, il costo computazionale è dato da $rn^2/2$.

Complessivamente quindi il costo computazionale è

$$\frac{n^3}{3} + rn^2. \tag{24}$$

Un caso particolarmente importante è quello relativo al calcolo dell'inversa di una matrice A non singolare, in cui la matrice B è la matrice identica di ordine n

$$AX = I.$$

Il costo computazionale del calcolo dell'inversa di una matrice di ordine n con il metodo di Gauss è però inferiore a $4n^3/3$ come risulterebbe dalla (24) per $n = r$. Infatti in questo caso si ha $B^{(n)} = L^{-1}$, cioè $B^{(n)}$ è triangolare inferiore: ne segue che al k-esimo passo per la costruzione di $[\, A^{(k+1)} \,|\, B^{(k+1)} \,]$ occorrono $(n-k)n$ operazioni moltiplicative. Quindi per gli $n-1$ passi richiesti il costo computazionale è

$$\sum_{k=1}^{n-1} n(n-k) \simeq \frac{n^3}{2}.$$

Infine la risoluzione dei sistemi $UX = L^{-1}$ ha lo stesso costo computazionale. In totale il costo computazionale dell'inversione di una matrice di ordine n con il metodo di Gauss è di n^3.

8. Analisi dell'errore del metodo di Gauss

La maggior parte dei calcolatori opera secondo un modello di rappresentazione dei dati e di calcolo che viene definito *sistema in virgola mobile* (in inglese *floating point system*), e che è caratterizzato da quattro parametri (numeri interi):

$\beta \geq 2$	la base della rappresentazione,
$t \geq 1$	il numero di cifre della rappresentazione,
$-m,\ M$	il minimo e il massimo esponente rappresentabili.

Un numero $x \neq 0$ è così rappresentato in questo sistema

$$x = \pm\ \beta^p\ (d_1 d_2 \dots d_t),$$

dove	$\pm$	indica il segno,
	p	è l'*esponente*, tale che $-m \leq p \leq M$,
	$d_1 d_2 \dots d_t$	sono le cifre della *mantissa*, tali che

$$0 \leq d_i \leq \beta - 1, \quad \text{con} \quad d_1 \neq 0,$$

ed è detto *numero di macchina*.

Alla rappresentazione $x = \pm\ \beta^p\ (d_1 d_2 \dots d_t)$ corrisponde il valore

$$\pm\ \beta^p \sum_{i=1}^{t} d_i \beta^{-i}.$$

Dato un numero reale x, con $\beta^{-m} \leq |x| \leq \beta^M$, il numero $\widetilde{x}$ rappresentato troncando o arrotondando alla t-esima le cifre di x, è tale che

$$\widetilde{x} = x(1 + \epsilon), \tag{25}$$

in cui ϵ è *l'errore (relativo) di arrotondamento*. Si può dimostrare che, indicata con

$$u = \begin{cases} \beta^{1-t} & \text{se le cifre di } x \text{ sono state troncate,} \\ \frac{1}{2}\beta^{1-t} & \text{se le cifre di } x \text{ sono state arrotondate,} \end{cases}$$

la *precisione di macchina*, risulta

$$|\epsilon| \leq \frac{u}{u+1}, \tag{26}$$

e quindi

$$|\epsilon| < u.$$

Se $|x| < \beta^{-m}$ o $|x| > \beta^M$, x non è rappresentabile come numero in virgola mobile e in questo caso il sistema segnala un errore, detto di *underflow* (nel primo caso) o di *overflow* (nel secondo caso).

Dalla (25) risulta anche

$$\widetilde{x} = \frac{x}{1+\eta},$$

dove per la (26) è

$$|\eta| = \frac{|\epsilon|}{|1+\epsilon|} \leq \frac{|\epsilon|}{1-|\epsilon|} \leq u.$$

In un sistema in virgola mobile sono implementate le *operazioni di macchina*. Se *op* è un'operazione aritmetica e x e y sono due numeri di macchina tali che la rappresentazione di x *op* y non dia luogo ad errori di underflow o di overflow, il risultato dell'applicazione ad x e y dell'operazione di macchina corrispondente ad *op* viene indicato con $fl(x\ op\ y)$ ed è tale che

$$fl(x\ op\ y) = (x\ op\ y)(1+\epsilon), \quad |\epsilon| \leq \frac{u}{1+u} < u, \tag{27}$$

$$fl(x\ op\ y) = \frac{x\ op\ y}{1+\eta}, \quad |\eta| \leq u. \tag{28}$$

Lo studio della propagazione dell'errore che si genera nell'applicazione di un algoritmo costituito da più operazioni aritmetiche fa uso delle relazioni (27) e (28). Se il numero delle operazioni è elevato, conviene, quando è possibile, usare la tecnica dell'*analisi dell'errore all'indietro* (*backward analysis*) introdotta da Wilkinson.

In questo paragrafo è riportata un'analisi dell'errore del metodo di Gauss nel caso che la matrice A e il vettore $\mathbf{b}$ abbiano elementi reali. Le considerazioni sulla stabilità che se ne deducono valgono però anche nel caso che A e $\mathbf{b}$ abbiano elementi complessi. Nelle maggiorazioni dei teoremi che seguono si usano le seguenti notazioni: con $|A|$ si intende la matrice che ha per elementi i moduli dei corrispondenti elementi di A e la relazione $A < B$ (risp. $A \leq B$) significa $a_{ij} < b_{ij}$ (risp. $a_{ij} \leq b_{ij}$) per $i, j = 1, \ldots, n$.

Si esamina dapprima l'errore generato dall'applicazione del metodo di sostituzione alla risoluzione di un sistema lineare con matrice triangolare.

4.15 Teorema. *Siano A una matrice triangolare inferiore di ordine n e $\mathbf{b}$ un vettore di ordine n aventi per elementi dei numeri di macchina e sia $\widetilde{\mathbf{x}}$ la soluzione effettivamente calcolata del sistema $A\mathbf{x} = \mathbf{b}$ con l'algoritmo*

$$x_1 = \frac{b_1}{a_{11}},$$
$$\text{per } i = 2, \ldots, n,$$

$$
\begin{aligned}
&s_{i0} = 0,\\
&\text{per } j = 1, \dots, i-1, \quad y_{ij} = a_{ij}x_j, \quad s_{ij} = s_{i,j-1} + y_{ij},\\
&x_i = \frac{b_i - s_{i,i-1}}{a_{ii}}\,.
\end{aligned}
$$

Allora esiste una matrice E di ordine n tale che

$$(A + E)\widetilde{\mathbf{x}} = \mathbf{b}, \quad |E| \leq nu|A| + O(u^2),$$

in cui $O(u^2)$ è una matrice i cui elementi sono funzioni di u di ordine maggiore o uguale al secondo.

Dim. Tenendo conto degli errori di arrotondamento che si producono ad ogni passo dell'algoritmo, applicando le relazioni (27) e (28), si ha per i valori effettivamente calcolati

$$
\left.
\begin{aligned}
\widetilde{x}_1 &= \frac{b_1}{a_{11}(1+\delta_1)}\\
\widetilde{y}_{ij} &= a_{ij}\widetilde{x}_j(1+\epsilon_{ij})\\
\widetilde{s}_{ij} &= (\widetilde{s}_{i,j-1} + \widetilde{y}_{ij})(1+\zeta_{ij})\\
\widetilde{x}_i &= \frac{b_i - \widetilde{s}_{i,i-1}}{a_{ii}(1+\eta_i)(1+\delta_i)}
\end{aligned}
\right\}
\quad
\begin{aligned}
&|\delta_i|, |\epsilon_{ij}|, |\zeta_{ij}|, |\eta_i| \leq u\\
&\text{per} \quad i = 2, \dots, n,\\
&\quad j = 1, \dots, i-1,
\end{aligned}
$$

in cui $\zeta_{i1} = 0$. Vale quindi

$$
\begin{aligned}
&\widetilde{s}_{21} = a_{21}\widetilde{x}_1(1+\epsilon_{21}),\\
&\widetilde{s}_{32} = a_{31}\widetilde{x}_1(1+\epsilon_{31}+\zeta_{32}+\epsilon_{31}\zeta_{32}) + a_{32}\widetilde{x}_2(1+\epsilon_{32}+\zeta_{32}+\epsilon_{32}\zeta_{32})
\end{aligned}
$$

e, in generale, per $i = 3, \dots, n$, si ha

$$\widetilde{s}_{i,i-1} = \sum_{k=1}^{i-1}\left[a_{ik}\widetilde{x}_k(1+\epsilon_{ik})\prod_{r=k}^{i-1}(1+\zeta_{ir})\right] = \sum_{k=1}^{i-1} a_{ik}\widetilde{x}_k(1+\gamma_{ik}),$$

dove

$$\gamma_{ik} = \epsilon_{ik} + \sum_{r=k}^{i-1}\zeta_{ir} + f(\epsilon_{ik}, \zeta_{ik}, \dots, \zeta_{i,i-1}), \quad \text{per} \quad k = 1, \dots, i-1,$$

in cui la funzione $f(\epsilon_{ik}, \zeta_{ik}, \dots, \zeta_{i,i-1})$ è tale che

$$|f(\epsilon_{ik}, \zeta_{ik}, \dots, \zeta_{i,i-1})| \leq O(u^2),$$

e

$$|\gamma_{ik}| \leq (i-k+1)u + O(u^2) \leq nu + O(u^2).$$

Si ha poi

$$\widetilde{x}_i = \frac{b_i - \widetilde{s}_{i,i-1}}{a_{ii}(1+\gamma_{ii})}, \quad \text{per} \quad i = 2, \ldots, n,$$

con $\gamma_{ii} = \eta_i + \delta_i + \eta_i\delta_i$, e quindi

$$|\gamma_{ii}| \leq 2u + O(u^2).$$

In conclusione, posto $\gamma_{11} = \delta_1$ e $\gamma_{21} = \epsilon_{21}$, risulta che il vettore $\widetilde{\mathbf{x}}$ effettivamente calcolato è quello che si otterrebbe applicando l'algoritmo in aritmetica esatta al sistema

$$(A+E)\mathbf{x} = \mathbf{b},$$

dove gli elementi della matrice $A+E$ sono

$$a_{ij}(1+\gamma_{ij}), \quad j = 1, \ldots, i, \ i = 1, \ldots, n,$$

e quindi per gli elementi $e_{ij} = a_{ij}\gamma_{ij}$ di E risulta

$$|e_{ij}| \leq nu|a_{ij}| + O(u^2). \qquad \blacksquare$$

Un teorema analogo al 4.15, con la stessa limitazione per $|E|$, vale anche nel caso che la matrice A sia triangolare superiore. Si esamina ora l'errore generato nella fattorizzazione LU di una matrice A.

4.16 Teorema. *Sia A una matrice di ordine n i cui elementi sono numeri di macchina e siano $\widetilde{L}$ e $\widetilde{U}$ le matrici della fattorizzazione LU di A effettivamente calcolate con il metodo di Gauss. Allora si ha che*

$$\widetilde{L}\widetilde{U} = A + H,$$

dove

$$|H| \leq 2nu\ (|A| + |\widetilde{L}|\ |\widetilde{U}|) + O(u^2).$$

Dim. Si procede per induzione.

Per $n = 1$ la relazione è ovviamente verificata.

Se $k = n > 1$, partizionata A nella forma

$$A^{(1)} = A = \begin{bmatrix} \alpha & \mathbf{c}^T \\ \mathbf{d} & B \end{bmatrix} \begin{matrix} \} & 1 \text{ riga} \\ \} & k-1 \text{ righe,} \end{matrix}$$

applicando il metodo di Gauss si ha

$$M^{(1)} = \begin{bmatrix} 1 & \mathbf{0}^T \\ \\ \text{-}\mathbf{v} & I_{n-1} \end{bmatrix}, \quad A^{(2)} = M^{(1)}A^{(1)} = \begin{bmatrix} \alpha & \mathbf{c}^T \\ \\ \mathbf{0} & A_1 \end{bmatrix},$$

dove

$$\mathbf{v} = \frac{1}{\alpha}\,\mathbf{d}, \quad A_1 = B - \mathbf{v}\mathbf{c}^T.$$

Per la (27) le componenti $\widetilde{v}_i$ effettivamente calcolate invece delle v_i sono date da

$$\widetilde{v}_i = fl\Big(\frac{1}{\alpha}\,d_i\Big) = \frac{1}{\alpha}\,d_i(1+\epsilon_i), \quad \text{dove} \quad |\epsilon_i| < u, \text{ per } i = 1, \ldots, k-1.$$

Sia $\widetilde{\mathbf{v}}$ il vettore le cui componenti sono le $\widetilde{v}_i$, risulta

$$\widetilde{\mathbf{v}} = \mathbf{v} + \mathbf{g}, \tag{29}$$

in cui la i-esima componente di $\mathbf{g}$ è data da $\dfrac{1}{\alpha}\,d_i\epsilon_i$, e quindi si ha

$$|\mathbf{g}| \leq u|\mathbf{v}|. \tag{30}$$

Invece della matrice $\mathbf{v}\mathbf{c}^T$ viene effettivamente calcolata una matrice $\widetilde{Z}$ eseguendo il prodotto esterno fra il vettore $\widetilde{\mathbf{v}}$ e il vettore $\mathbf{c}$, e poiché ogni prodotto è affetto dall'errore dovuto alla moltiplicazione, gli elementi di $\widetilde{Z}$ sono dati da

$$\widetilde{z}_{ij} = \widetilde{v}_i c_j(1+\eta_{ij}), \quad \text{dove} \quad |\eta_{ij}| < u, \text{ per } i, j = 1, \ldots, k-1.$$

Quindi si ha

$$\widetilde{Z} = \widetilde{\mathbf{v}}\mathbf{c}^T + E, \tag{31}$$

dove gli elementi $e_{ij} = \widetilde{v}_i c_j \eta_{ij}$ della matrice E sono tali che

$$|e_{ij}| \leq u|\widetilde{v}_i|\,|c_j|, \text{ per } i, j = 1, \ldots, k-1,$$

cioè

$$|E| \leq u|\widetilde{\mathbf{v}}|\,|\mathbf{c}|^T.$$

Analogamente la matrice $\widetilde{A}_1$, effettivamente calcolata invece della A_1, è data da

$$\widetilde{A}_1 = B - \widetilde{Z} + E' \tag{32}$$

dove la matrice E' ha come elementi e'_{ij}, gli errori generati dalla sottrazione dei termini di indici i, j di B e di $\widetilde{Z}$, e quindi

$$|e'_{ij}| \leq u|b_{ij} - \widetilde{z}_{ij}|, \quad i,\ j = 1, \ldots, k-1,$$

cioè

$$|E'| \leq u|B - \widetilde{Z}| \leq u\ (|B| + |\widetilde{Z}|).$$

Da (31) e (32) segue che

$$\widetilde{A}_1 = B - \widetilde{\mathbf{v}}\mathbf{c}^T + F, \quad |F| = |E' - E| \leq u\ (|B| + 2|\widetilde{\mathbf{v}}|\ |\mathbf{c}|^T) + O(u^2), \quad (33)$$

e quindi

$$|\widetilde{A}_1| \leq |B| + |\widetilde{\mathbf{v}}|\ |\mathbf{c}|^T + |F| \leq (1+u)|B| + (1+2u)|\widetilde{\mathbf{v}}|\ |\mathbf{c}|^T + O(u^2). \quad (34)$$

Si consideri adesso la fattorizzazione LU della matrice $\widetilde{A}_1$, cioè

$$\widetilde{A}_1 = L_1 U_1.$$

La matrice $\widetilde{A}_1$ è di ordine $k-1$. Allora, per l'ipotesi induttiva, per le matrici effettivamente calcolate $\widetilde{L}_1$ e $\widetilde{U}_1$ risulta

$$\widetilde{L}_1\widetilde{U}_1 = \widetilde{A}_1 + H_1, \quad |H_1| \leq 2(k-1)\,u\ (|\widetilde{A}_1| + |\widetilde{L}_1|\ |\widetilde{U}_1|) + O(u^2). \quad (35)$$

Le matrici $\widetilde{L}$ e $\widetilde{U}$ effettivamente calcolate nella fattorizzazione della matrice A possono essere così rappresentate

$$\widetilde{L} = \begin{bmatrix} 1 & \mathbf{0}^T \\ \widetilde{\mathbf{v}} & \widetilde{L}_1 \end{bmatrix}, \quad \widetilde{U} = \begin{bmatrix} \alpha & \mathbf{c}^T \\ \mathbf{0} & \widetilde{U}_1 \end{bmatrix},$$

e quindi

$$\widetilde{L}\widetilde{U} = \begin{bmatrix} \alpha & \mathbf{c}^T \\ \alpha\widetilde{\mathbf{v}} & (\widetilde{L}_1\widetilde{U}_1 + \widetilde{\mathbf{v}}\mathbf{c}^T) \end{bmatrix} = A + H,$$

dove per la (29), la (33) e la (35) è

$$H = \begin{bmatrix} 0 & \mathbf{0}^T \\ \alpha\mathbf{g} & (H_1 + F) \end{bmatrix}.$$

Da (33), (34) e (35), si ottiene

$$\begin{aligned}|H_1| + |F| &\leq 2(k-1)\,u\,(|B| + |\widetilde{\mathbf{v}}|\,|\mathbf{c}|^T + |\widetilde{L}_1|\,|\widetilde{U}_1|) + u\,(|B| + 2|\widetilde{\mathbf{v}}|\,|\mathbf{c}|^T) \\ &\qquad + O(u^2) \\ &\leq 2ku\,(|B| + |\widetilde{\mathbf{v}}|\,|\mathbf{c}|^T + |\widetilde{L}_1|\,|\widetilde{U}_1|) + O(u^2).\end{aligned}$$

Inoltre per la (30) è

$$|\alpha\mathbf{g}| \leq u|\alpha|\,|\mathbf{v}| = u|\mathbf{d}|,$$

e quindi

$$\begin{aligned}|H| &\leq 2ku \begin{bmatrix} 0 & \mathbf{0}^T \\ |\mathbf{d}| & (|B| + |\widetilde{\mathbf{v}}|\,|\mathbf{c}|^T + |\widetilde{L}_1|\,|\widetilde{U}_1|) \end{bmatrix} + O(u^2) \\ &\leq 2ku \left\{ \begin{bmatrix} |\alpha| & |\mathbf{c}|^T \\ |\mathbf{d}| & |B| \end{bmatrix} + \begin{bmatrix} 1 & \mathbf{0}^T \\ |\widetilde{\mathbf{v}}| & |\widetilde{L}_1| \end{bmatrix} \begin{bmatrix} |\alpha| & |\mathbf{c}|^T \\ \mathbf{0} & |\widetilde{U}_1| \end{bmatrix} \right\} + O(u^2) \\ &= 2ku(|A| + |\widetilde{L}|\,|\widetilde{U}|) + O(u^2). \end{aligned}$$

∎

Si utilizzano ora i risultati dei teoremi 4.15 e 4.16 per studiare la stabilità del metodo di Gauss nella risoluzione di un sistema lineare $A\mathbf{x} = \mathbf{b}$.

4.17 Teorema. *Siano A e $\mathbf{b}$ una matrice e un vettore di ordine n aventi per elementi dei numeri di macchina e sia $\widetilde{\mathbf{x}}$ il vettore effettivamente calcolato nella risoluzione del sistema $A\mathbf{x} = \mathbf{b}$ mediante i seguenti passi:*

a) determinazione delle matrici $\widetilde{L}$ e $\widetilde{U}$ effettivamente calcolate nella fattorizzazione LU della matrice A,

b) determinazione del vettore $\widetilde{\mathbf{y}}$ effettivamente calcolato nella risoluzione del sistema $\widetilde{L}\mathbf{y} = \mathbf{b}$,

c) determinazione del vettore $\widetilde{\mathbf{x}}$ effettivamente calcolato nella risoluzione del sistema $\widetilde{U}\mathbf{x} = \widetilde{\mathbf{y}}$.

Allora vale

$$(A + \Delta A)\,\widetilde{\mathbf{x}} = \mathbf{b}, \quad |\Delta A| \leq 4nu\,(|A| + |\widetilde{L}|\,|\widetilde{U}|) + O(u^2).$$

Dim. Poiché i sistemi di b) e c) hanno matrici triangolari, allora, per il teorema 4.15, i vettori effettivamente calcolati $\widetilde{\mathbf{x}}$ e $\widetilde{\mathbf{y}}$ sono tali che

$$(\widetilde{L} + F)\,\widetilde{\mathbf{y}} = \mathbf{b}, \quad |F| \leq nu|\widetilde{L}| + O(u^2), \tag{36}$$

$$(\widetilde{U} + G)\,\widetilde{\mathbf{x}} = \widetilde{\mathbf{y}}, \quad |G| \leq nu|\widetilde{U}| + O(u^2). \tag{37}$$

Sostituendo allora $\widetilde{\mathbf{y}}$ dalla (37) nella (36), si ha

$$(\widetilde{L}\widetilde{U} + F\widetilde{U} + \widetilde{L}G + FG)\,\widetilde{\mathbf{x}} = \mathbf{b}.$$

Per il teorema 4.16, posto

$$\Delta A = H + F\widetilde{U} + \widetilde{L}G + FG,$$

si ha che

$$(A + \Delta A)\,\widetilde{\mathbf{x}} = \mathbf{b}.$$

Per il teorema 4.16 e per le (36) e (37) si ha:

$$|\Delta A| \leq 2nu\,(|A| + |\widetilde{L}|\,|\widetilde{U}|) + 2nu\,|\widetilde{L}|\,|\widetilde{U}| + O(u^2),$$

da cui segue la tesi. ∎

Dal teorema 4.17 risulta che se le matrici $|\widetilde{L}|$ e $|\widetilde{U}|$ hanno elementi molto grandi, si possono produrre elevati errori algoritmici nel calcolo della soluzione.

4.18 Esempio. La soluzione del sistema lineare $A\mathbf{x} = \mathbf{b}$, dove

$$A = \begin{bmatrix} \epsilon & 1 \\ 1 & 0 \end{bmatrix} \quad \text{e} \quad \mathbf{b} = \begin{bmatrix} 1+\epsilon \\ 1 \end{bmatrix}, \quad \epsilon > 0,$$

è data da $\mathbf{x} = [1, 1]^T$. Questo problema è ben posto; si ha infatti:

$$A^{-1} = \begin{bmatrix} 0 & 1 \\ 1 & \text{-}\epsilon \end{bmatrix},$$

per cui il numero di condizionamento di A, $\mu_\infty(A) = (1+\epsilon)^2$, è di poco superiore all'unità se ϵ è piccolo. La fattorizzazione LU di A è

$$A = \begin{bmatrix} 1 & 0 \\ 1/\epsilon & 1 \end{bmatrix} \quad \begin{bmatrix} \epsilon & 1 \\ 0 & \text{-}1/\epsilon \end{bmatrix},$$

da cui risulta che gli elementi di $|L|$ e di $|U|$ possono diventare comunque grandi al diminuire di ϵ. Sia ad esempio $\epsilon = 0.3\ 10^{-3}$. Con il metodo di Gauss operando in virgola mobile in base 10 con 3 cifre significative e arrotondamento, si hanno le seguenti approssimazioni di $M^{(1)}$ e $[A^{(2)}|\,\mathbf{b}^{(2)}]$:

$$M^{(1)} = \begin{bmatrix} 1 & 0 \\ \text{-}0.333\ 10^4 & 1 \end{bmatrix}$$

e

$$[A^{(2)}|\,\mathbf{b}^{(2)}] = M^{(1)}[A^{(1)}|\,\mathbf{b}^{(1)}] = \left[\begin{array}{cc|c} 0.3\ 10^{-3} & 1 & 0.100\ 10^{1} \\ 0 & \text{-}0.333\ 10^{4} & \text{-}0.333\ 10^{4} \end{array}\right]$$

da cui si ottiene

$$\widetilde{x}_2 = \frac{0.333\ 10^4}{0.333\ 10^4} = 1, \quad \widetilde{x}_1 = \frac{0.100\ 10^1 - 0.100\ 10^1}{0.3\ 10^{-3}} = 0. \qquad \blacksquare$$

L'errore elevato da cui è affetta la soluzione calcolata dell'esempio 4.18 è causato dal fatto che l'elemento di massimo modulo di $\widetilde{L}$ e di $\widetilde{U}$ è più di 3000 volte l'elemento di massimo modulo di A. In generale per il metodo di Gauss la crescita degli elementi delle matrici $A^{(k)}$, e quindi delle matrici L e U, rispetto alla matrice A non è limitabile a priori con una espressione che dipende solo da n, quindi il metodo di Gauss può essere instabile anche quando è applicato a problemi ben posti.

9. Massimo pivot

Una strategia per il metodo di Gauss che consente di contenere la crescita degli elementi di $A^{(k)}$, e quindi di L e di U rispetto agli elementi di A, è quella che utilizza la tecnica del *massimo pivot*.

Al k-esimo passo con la tecnica del *massimo pivot parziale*, si determina l'indice di riga r per cui

$$|a_{rk}^{(k)}| = \max_{k \le i \le n} |a_{ik}^{(k)}|,$$

e si scambiano la r-esima riga con la k-esima prima di calcolare $A^{(k+1)}$. In tal modo gli elementi della matrice $M^{(k)}$ hanno modulo minore o uguale a 1. Per gli elementi di $A^{(k+1)}$ si ha dalla (22)

$$|a_{ij}^{(k+1)}| = |a_{ij}^{(k)} - m_{ik}a_{kj}^{(k)}| \le |a_{ij}^{(k)}| + |m_{ik}|\,|a_{kj}^{(k)}| \le |a_{ij}^{(k)}| + |a_{kj}^{(k)}|; \qquad (38)$$

indicando con $a_M^{(k)}$ il massimo modulo degli elementi di $A^{(k)}$, dalla (38) si ha:

$$a_M^{(k+1)} \le 2a_M^{(k)}.$$

Risulta quindi

$$a_M^{(k)} \le f(k)a_M^{(1)}, \qquad (39)$$

in cui $f(k) = 2^{(k-1)}$, e all'ultimo passo, cioè per $k = n$, si ha:

$$a_M^{(n)} \le 2^{n-1}a_M^{(1)}.$$

Perciò con il metodo di Gauss con la variante del massimo pivot parziale gli elementi della matrice L hanno modulo minore o uguale a 1 e gli elementi della matrice U hanno modulo minore o uguale a $2^{n-1}a_M^{(1)}$. Con questa variante il metodo di Gauss risulta in generale assai più stabile.

La maggiorazione (39) viene raramente raggiunta: esistono comunque delle matrici A per cui la (39) vale con il segno di uguaglianza.

4.19 Esempio. Si consideri la matrice A di ordine n i cui elementi sono

$$a_{ij} = \begin{cases} 1 & \text{se } i = j \text{ e se } j = n, \\ \text{-1} & \text{se } i > j, \\ 0 & \text{altrimenti.} \end{cases}$$

La matrice $A^{(n)}$ ottenuta con il metodo di Gauss con la variante del massimo pivot parziale ha gli elementi

$$a_{ij}^{(n)} = \begin{cases} 1 & \text{se } i = j \neq n, \\ 2^{i-1} & \text{se } j = n, \\ 0 & \text{altrimenti;} \end{cases}$$

e quindi $a_M^{(n)} = 2^{n-1}a_M^{(1)}$. Per $n = 4$ si ha:

$$A = \begin{bmatrix} 1 & 0 & 0 & 1 \\ \text{-1} & 1 & 0 & 1 \\ \text{-1} & \text{-1} & 1 & 1 \\ \text{-1} & \text{-1} & \text{-1} & 1 \end{bmatrix},$$

$$A^{(4)} = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 4 \\ 0 & 0 & 0 & 8 \end{bmatrix}.$$

Quindi in questo caso la maggiorazione (39) vale con il segno di uguaglianza.

∎

4.20 Esempio. Si consideri il sistema $A\mathbf{x} = \mathbf{b}$ dell'esempio 4.18 con $\epsilon = 0.3\ 10^{-3}$. Applicando il metodo di Gauss con la tecnica del massimo pivot parziale, si scambiano fra loro le righe della matrice A e del vettore $\mathbf{b}$. Operando con una aritmetica in virgola mobile in base 10 con tre cifre significative, si ottiene

$$\widetilde{M}^{(1)} = \widetilde{A}^{(1)}$$

e

$$[\widetilde{A}^{(2)} |\ \widetilde{\mathbf{b}}^{(2)}] = \widetilde{M}^{(1)}[A^{(1)} |\ \mathbf{b}^{(1)}] = \left[\begin{array}{cc|c} 1 & 0 & 1 \\ 0 & 1 & 1 \end{array}\right],$$

da cui si ottiene il vettore $\widetilde{\mathbf{x}} = [1, 1]^T$. Il risultato ottenuto, in questo caso, non è affetto da errore. ∎

Il metodo di Gauss con la tecnica del massimo pivot parziale corrisponde alla fattorizzazione LU della matrice ΠA, dove Π è la matrice di permutazione che opera un riordinamento delle righe di A tale che ad ogni passo k l'elemento di massimo modulo della k-esima colonna sia nella posizione (k, k) del pivot.

Un'altra strategia per il metodo di Gauss che consente in generale di ridurre ancora di più la crescita degli elementi delle matrici $A^{(k)}$ è quella che utilizza la tecnica del *massimo pivot totale*. Al k-esimo passo con la tecnica del massimo pivot totale si determina l'elemento di massimo modulo di tutta la sottomatrice $B^{(k)}$ e si utilizza tale elemento come pivot, cioè si determinano l'indice di riga r e l'indice di colonna s per cui

$$|a_{rs}^{(k)}| = \max_{k \le i,j \le n} |a_{ij}^{(k)}|.$$

Per portare l'elemento $a_{rs}^{(k)}$ nella posizione (k, k) del pivot, è necessario uno scambio fra le righe di indice r e k e uno scambio fra le colonne di indice s e k. Lo scambio di righe non modifica la soluzione del sistema lineare, che rimane equivalente a quello iniziale, mentre lo scambio di colonne modifica l'ordinamento delle componenti del vettore soluzione. Infatti, se Π'_k è la matrice di permutazione che scambia fra loro le colonne di indice s e k, allora al k-esimo passo si ha

$$A^{(k)} \Pi'_k \mathbf{y}^{(k)} = \mathbf{b}^{(k)},$$

ed essendo $[\Pi'_k]^{-1} = [\Pi'_k]^T$, risulta dalla (23)

$$\mathbf{y}^{(k)} = [\Pi'_k]^T \mathbf{x}^{(k)}.$$

Cioè il vettore $\mathbf{y}^{(k)}$ non è altro che il vettore $\mathbf{x}^{(k)}$ in cui sono state scambiate le componenti di indice s e k. Quindi, calcolato il vettore $\mathbf{y}^{(n)}$, si ha

$$\mathbf{x} = [\Pi'_{n-1} \Pi'_{n-2} \dots \Pi'_1]^T \mathbf{y}^{(n)}.$$

La variante del massimo pivot totale richiede un maggior tempo di elaborazione di quello richiesto dalla variante del massimo pivot parziale: al k-esimo passo per la ricerca dell'elemento di massimo modulo sono necessari $(n-k+1)^2$ confronti fra gli elementi della sottomatrice $B^{(k)}$. Globalmente sono richiesti $n^3/3$ confronti, e queste operazioni, che non modificano il costo computazionale del metodo di Gauss, richiedono un tempo di esecuzione confrontabile con quello richiesto dall'esecuzione delle operazioni aritmetiche.

Il metodo di Gauss con la tecnica del massimo pivot totale è in generale molto più stabile del metodo di Gauss con la tecnica del massimo pivot parziale: infatti la crescita degli elementi delle matrici $A^{(k)}$ risulta in questo caso limitata dalla relazione

$$a_M^{(k)} \leq g(k)\, a_M^{(1)}, \tag{40}$$

in cui

$$g(k) = \sqrt{k \prod_{j=2}^{k} j^{1/(j-1)}}, \quad k \geq 2$$

(si veda l'esercizio 4.27) La funzione $g(k)$ della maggiorazione (40) cresce con k assai più lentamente della funzione $f(k)$ della maggiorazione (39), come risulta anche dalla figura 4.2.

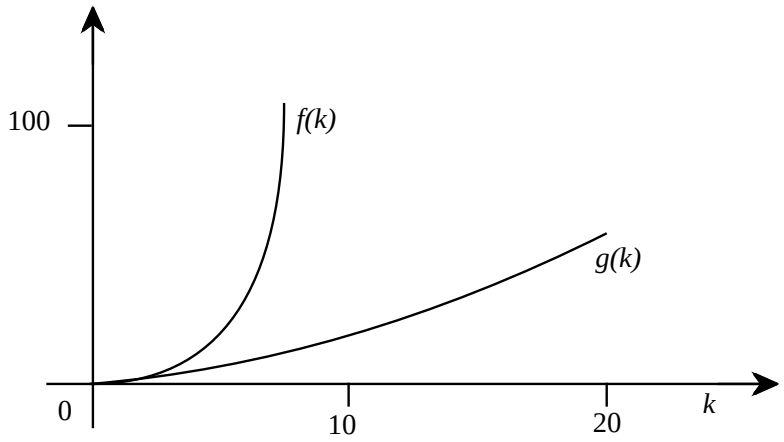


Fig. 4.2 - Grafici delle funzioni delle maggiorazioni (39) e (40).

Comunque la maggiorazione (40) risulta generalmente essere una forte sovrastima della crescita effettiva degli elementi della matrice $A^{(k)}$. Non si conoscono matrici per cui la maggiorazione (40) vale con il segno di uguaglianza: una congettura di Wilkinson, dimostrata per $n \leq 4$ [8], ipotizza che la maggiorazione (40) nel caso di matrici ad elementi reali debba valere con la funzione $g(k) = k$.

4.21 Esempio. Si consideri la seguente matrice (di *Hankel*) A_n di ordine n i cui elementi sono

$$a_{i,n+k-i}^{(n)} = \begin{cases} 2^k & \text{se } k > 0 \\ \\ 2^{1/(2-k)} & \text{se } k \leq 0 \end{cases}, \quad i = 1, \ldots, n, \quad k = i+1-n, \ldots, i.$$

Ad esempio, per $n = 4$ è

$$A_4 = \begin{bmatrix} \sqrt[4]{2} & \sqrt[3]{2} & \sqrt{2} & 2 \\ \sqrt[3]{2} & \sqrt{2} & 2 & 2^2 \\ \sqrt{2} & 2 & 2^2 & 2^3 \\ 2 & 2^2 & 2^3 & 2^4 \end{bmatrix}.$$

Nella seguente tabella sono riportati, per alcuni valori dell'ordine n, i corrispondenti valori del numero di condizionamento $\mu_2(A_n)$ e della funzione

$$h(n) = \frac{2\mu_2(A_n)\, u}{1 - \mu_2(A_n)\, u}$$

ottenuta dal secondo membro della (4) quando al posto di ϵ_A e ϵ_b si sostituisce la precisione di macchina $u = 16^{-5}$ di un calcolatore IBM serie 370.

n	$\mu_2(A_n)$	$h(n)$
4	$2.31\ 10^2$	$4.41\ 10^{-4}$
6	$1.13\ 10^3$	$2.15\ 10^{-3}$
8	$4.82\ 10^3$	$9.23\ 10^{-3}$
10	$1.99\ 10^4$	$3.86\ 10^{-2}$
12	$8.09\ 10^4$	$1.67\ 10^{-1}$
14	$3.28\ 10^5$	$9.09\ 10^{-1}$
16	$1.32\ 10^6$	

Per $n = 16$ non è riportato il valore di $h(n)$ perché $\mu_2(A_n)\, u > 1$. La funzione $h(n)$ è una maggiorazione dell'errore inerente del problema (1) quando le perturbazioni ϵ_A e ϵ_b sono minori della precisione di macchina, cioè quando gli elementi di A e di $\mathbf{b}$ sono affetti solo dagli errori di rappresentazione.

Costruito il vettore $\mathbf{b}$ in modo che il sistema $A\mathbf{x} = \mathbf{b}$ abbia come soluzione $\mathbf{x} = [1, 1, \ldots, 1]^T$, si risolve questo sistema lineare con i tre metodi:

a) Gauss,

b) Gauss con massimo pivot parziale,

c) Gauss con massimo pivot totale.

La tabella riporta gli errori relativi

$$\epsilon_x = \frac{\|\delta\mathbf{x}\|_2}{\|\mathbf{x}\|_2}$$

da cui sono affette le soluzioni calcolate con i tre metodi.

n	Gauss	pivot parziale	pivot totale
4	$1.61\ 10^{-5}$	$1.30\ 10^{-5}$	$1.60\ 10^{-6}$
6	$7.84\ 10^{-6}$	$1.83\ 10^{-5}$	$2.09\ 10^{-5}$
8	$4.96\ 10^{-4}$	$3.54\ 10^{-4}$	$1.93\ 10^{-5}$
10	$1.38\ 10^{-2}$	$3.05\ 10^{-4}$	$7.52\ 10^{-5}$
12	$7.84\ 10^{-3}$	$3.08\ 10^{-3}$	$1.07\ 10^{-4}$
14	$7.48\ 10^{-1}$	$3.01\ 10^{-3}$	$8.75\ 10^{-4}$
16	$2.26\ 10^{-1}$	$2.49\ 10^{-2}$	$9.15\ 10^{-4}$

Nella figura 4.3 è riportato in scala logaritmica, al variare di n, il grafico della funzione $h(n)$ (i valori sono indicati con dei quadratini) e gli errori effettivi ϵ_x generati con

il metodo di Gauss (rappresentati con $*$),

il metodo di Gauss con massimo pivot parziale (rappresentati con $+$),

il metodo di Gauss con massimo pivot totale (rappresentati con o).

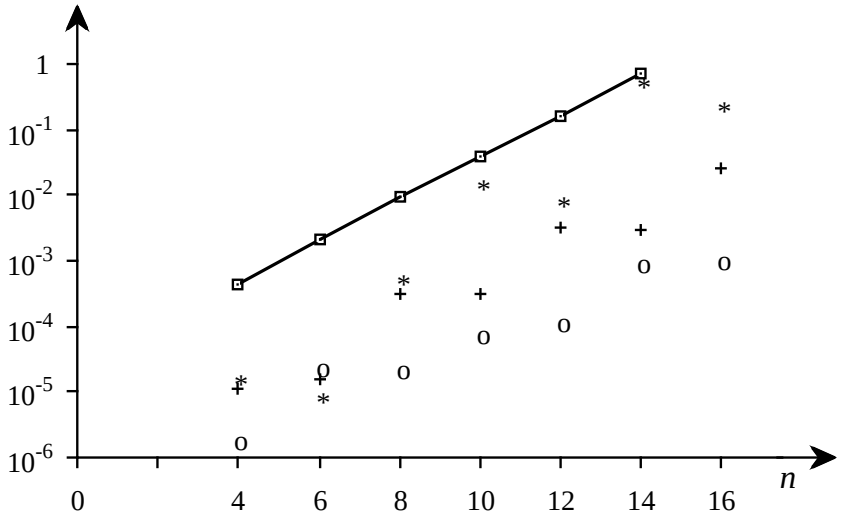


Fig. 4.3 - Errori relativi del metodo di Gauss e delle sue varianti.

Poiché gli errori effettivamente generati sono dati dalla somma degli errori inerenti e degli errori algoritmici, dalla figura 4.3 risulta che in questo caso la maggiorazione (4) fornisce una stima assai pessimistica dell'errore inerente. Per valori piccoli di n e quindi del condizionamento di A, i tre metodi generano errori confrontabili, mentre per valori più grandi di n, il metodo di Gauss con massimo pivot parziale produce risultati affetti da un

errore minore di quelli prodotti dal metodo di Gauss, e migliori risultati si ottengono con la variante del massimo pivot totale. ∎

10. Implementazione del metodo di Gauss

Nella implementazione su calcolatore della fattorizzazione LU di una matrice A con il metodo di Gauss l'area di memoria riservata per contenere inizialmente la matrice A può essere utilizzata per memorizzare le due matrici L ed U: i moltiplicatori m_{jk}, $j = k+1, \ldots, n$, del k-esimo passo sono memorizzati nella stessa area di memoria occupata dagli elementi $a_{jk}^{(k)}$, che non vengono più utilizzati nei passi successivi; gli elementi $a_{rs}^{(k)}, r, s = k+1, \ldots, n$, sono memorizzati nella stessa area di memoria occupata dagli $a_{rs}^{(k-1)}$. Al termine del procedimento la matrice L, esclusa la diagonale principale, i cui elementi sono uguali a 1, è memorizzata nella stessa area di memoria inizialmente occupata dalla parte strettamente triangolare inferiore di A e la matrice U è memorizzata nella stessa area di memoria inizialmente occupata dalla parte triangolare superiore di A.

Nel caso del metodo con la variante del massimo pivot parziale, gli scambi fra righe della matrice A possono non essere effettivamente eseguiti. La posizione della i-esima riga della matrice A può essere individuata utilizzando un vettore $\mathbf{v}$ i cui elementi inizialmente sono $v_i = i$, $i = 1, \ldots, n$. L'elemento a_{ij} della matrice A risulta allora memorizzato nella posizione di indice di riga v_i e colonna j. Quando è richiesto lo scambio delle righe di indice k e r della matrice A, questo scambio non viene eseguito sulla matrice ma sul vettore $\mathbf{v}$, scambiando fra loro gli elementi v_k e v_j. Dopo questo scambio la riga di A contenente il pivot è quella di indice v_k. In modo analogo si procede nel caso della variante del massimo pivot totale usando due vettori di indici $\mathbf{u}$ e $\mathbf{v}$, il primo per l'indice di riga e il secondo per l'indice di colonna.

Il metodo di Gauss può essere utilizzato, oltre che per risolvere sistemi, anche per calcolare il determinante o il rango di una matrice A. Infatti, il determinante di A è dato dal prodotto degli elementi principali di U (a meno del segno se il metodo è applicato con una strategia di pivot e sono richiesti scambi di righe). Se si usa la variante del massimo pivot totale, il rango di A è dato dal numero r degli elementi non nulli sulla diagonale di U.

L'implementazione del metodo di Gauss deve prevedere anche un controllo ad ogni passo sulla grandezza del pivot. Infatti se ad un certo passo il pivot risulta in modulo troppo piccolo, è possibile che l'esecuzione del programma si interrompa in modo anomalo (ad esempio per il verificarsi di un errore di *underflow* o di *overflow* o per una divisione per zero). Per questo, se il modulo del pivot assume valori più piccoli di una quantità prefissata

σ, esso viene considerato nullo. È opportuno rilevare che i pivot non nulli, ma in modulo minori di σ, possono corrispondere a elementi che in teoria dovrebbero essere nulli, ma che in pratica non lo sono, perché affetti dagli errori di arrotondamento: in questo caso la sostituzione di tali elementi con zero è appropriata. Ma è anche possibile che tali pivot corrispondano a elementi che in teoria non sono nulli: in tal caso la sostituzione con lo zero può non alterare eccessivamente la soluzione del sistema, mentre è critica per il calcolo del rango della matrice, in quanto il numero degli elementi principali di U che si assumono nulli viene a dipendere dal valore di σ. La determinazione di un valore adeguato di σ è difficile, in quanto una piccola variazione del valore di σ può generare una grande variazione del numero degli elementi principali di U che si assumono nulli. Una più efficiente determinazione del rango di una matrice si ottiene utilizzando il metodo dei valori singolari (si veda il capitolo 7). È possibile determinare esattamente il rango di una matrice di elementi interi o razionali usando il metodo di Gauss con una aritmetica modulo p, dove p è un numero primo opportuno (si veda l'esercizio 4.43).

4.22 Esempio. Si calcoli con il metodo di Gauss il rango della matrice

$$A = \begin{bmatrix} 0.58 & -1.1 & -0.52 \\ -0.56 & 1.12 & 0.56 \\ 0.02 & 0.02 & 0.04 \end{bmatrix}$$

operando in virgola mobile in base 10 con 3 cifre significative. Si ha:

$$M^{(1)} = \begin{bmatrix} 1 & 0 & 0 \\ 0.966 & 1 & 0 \\ 0.0345 & 0 & 1 \end{bmatrix}, \quad A^{(2)} = \begin{bmatrix} 0.58 & -1.1 & -0.52 \\ 0 & 0.06 & 0.058 \\ 0 & 0.0579 & -0.0579 \end{bmatrix},$$

$$M^{(2)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0.965 & 1 \end{bmatrix}, \quad A^{(3)} = \begin{bmatrix} 0.58 & -1.1 & -0.52 \\ 0 & 0.06 & 0.058 \\ 0 & 0 & 0.0019 \end{bmatrix}.$$

Se si pone $\sigma = 10^{-3}$, gli elementi diagonali di $A^{(3)}$ risultano tutti maggiori di σ e quindi A risulta di rango 3. Se si pone invece $\sigma = 2.\ 10^{-3}$, risulta che A ha rango 2 e se si pone $\sigma = 10^{-1}$, risulta che A ha rango 1 (da notare che la matrice A ha effettivamente rango 2). ∎

11. Metodo di Gauss-Jordan

Un'altra variante del metodo di Gauss è il metodo di *Gauss-Jordan*, che si basa ancora su una successione di moltiplicazioni della matrice A per matrici elementari. Con questa variante al k-esimo passo si annullano gli elementi della k-esima colonna con indice di riga diverso da k.

La k-esima matrice elementare è data da

$$E^{(k)} = J^{(k)} = \begin{bmatrix} 1 & & -m_{1k} & & \\ & \ddots & \vdots & & \\ & & -m_{k-1,k} & & \\ & & 1 & & \\ & & -m_{k+1,k} & & \\ & & \vdots & \ddots & \\ & & -m_{nk} & & 1 \end{bmatrix}.$$

Posto $A^{(1)} = A$, la successione delle matrici $A^{(k)}$ è così definita

$$A^{(k+1)} = J^{(k)} A^{(k)}, \quad k = 1, \ldots, n.$$

La matrice $A^{(k+1)}$ ha allora la forma:

$$A^{(k+1)} = \begin{bmatrix} C^{(k+1)} & D^{(k+1)} \\ O & B^{(k+1)} \end{bmatrix} \begin{matrix} \} & k \text{ righe} \\ \} & n-k \text{ righe}, \end{matrix}$$

in cui $C^{(k+1)}$ è una matrice diagonale. La matrice $A^{(n+1)}$ risulta quindi diagonale.

Se il metodo è applicato alla risoluzione del sistema lineare $A\mathbf{x} = \mathbf{b}$, la matrice che si ottiene al termine del procedimento è $[\, A^{(n+1)} \,|\, \mathbf{b}^{(n+1)} \,]$, dove $A^{(n+1)}$ è una matrice diagonale, e la soluzione è quindi data da

$$x_i = \frac{b_i^{(n+1)}}{a_{ii}^{(n+1)}}, \quad i = 1, \ldots, n.$$

Il costo computazionale del metodo di Gauss-Jordan per risolvere un sistema lineare è superiore a quella del metodo di Gauss. Infatti al k-esimo passo la costruzione della matrice $J^{(k)}$ richiede $n-1$ divisioni per il calcolo degli elementi m_{rk}, $r = 1, \ldots, n$, $r \neq k$, mentre per moltiplicare $J^{(k)}$ per $A^{(k)}$ occorrono $(n-1)(n-k)$ operazioni moltiplicative. Quindi, a meno di termini di ordine inferiore, al k-esimo passo occorrono $n(n-k)$ operazioni moltiplicative e il costo computazionale del metodo è dato da

$$\sum_{k=1}^{n} n(n-k) = n \sum_{k=1}^{n-1} k \simeq \frac{n^3}{2}.$$

Il metodo di Gauss-Jordan può essere utilizzato anche per il calcolo della matrice inversa A^{-1}: in tal caso il costo computazionale è lo stesso del metodo di Gauss.

Per quanto riguarda la stabilità del metodo di Gauss-Jordan, conviene distinguere due fasi: una prima fase in cui vengono eliminati i termini che si trovano sotto la diagonale principale, che corrisponde a un'applicazione del metodo di Gauss, e in questa fase si può usare una tecnica di massimo pivot, e una seconda fase, in cui vengono eliminati i termini che si trovano al di sopra della diagonale principale, e che corrisponde a un'applicazione del metodo di Gauss senza pivot; in questa seconda fase non vi è alcun controllo sulla crescita degli elementi $m_{rk}, r = 1, \ldots, k-1$, che possono diventare comunque elevati in modulo.

4.23 Esempio. Applicando il metodo di Gauss-Jordan alla matrice

$$A^{(1)} = A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1+\epsilon & 2 \\ 1 & 1 & 2 \end{bmatrix}, \quad \text{per } \epsilon > 0,$$

si ottiene

$$M^{(1)} = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}, \quad A^{(2)} = \begin{bmatrix} 1 & 1 & 1 \\ 0 & \epsilon & 1 \\ 0 & 0 & 1 \end{bmatrix},$$

$$M^{(2)} = \begin{bmatrix} 1 & -1/\epsilon & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad A^{(3)} = \begin{bmatrix} 1 & 0 & 1-1/\epsilon \\ 0 & \epsilon & 1 \\ 0 & 0 & 1 \end{bmatrix},$$

$$M^{(3)} = \begin{bmatrix} 1 & 0 & 1/\epsilon - 1 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{bmatrix}, \quad A^{(4)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \epsilon & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Se ϵ è molto piccolo, nelle matrici $A^{(3)}$, $M^{(2)}$ e $M^{(3)}$ compaiono elementi molto grandi. ∎

Però Peters e Wilkinson hanno dimostrato [22] che se al passo k si applica una tecnica di pivot parziale alle righe di indice maggiore o uguale a k, anche se durante il procedimento si generano al di sopra della diagonale principale elementi di modulo molto elevato, questi non influenzano l'errore della soluzione. Cioè il metodo di Gauss-Jordan con pivot parziale è stabile quanto il metodo di Gauss con pivot parziale.

4.24 Esempio. Al sistema lineare $A\mathbf{x} = \mathbf{b}$ dell'esempio 4.21 si applica il metodo di Gauss-Jordan, senza varianti del pivot, con la variante del pivot

parziale e con la variante del pivot totale. I valori ottenuti per $\epsilon_x = \dfrac{\|\delta\mathbf{x}\|_2}{\|\mathbf{x}\|_2}$ sono:

n	Gauss-Jordan	pivot parziale	pivot totale
4	$1.82\ 10^{-5}$	$1.33\ 10^{-5}$	$1.60\ 10^{-6}$
6	$8.42\ 10^{-5}$	$1.79\ 10^{-5}$	$2.09\ 10^{-5}$
8	$1.12\ 10^{-3}$	$3.51\ 10^{-4}$	$1.93\ 10^{-5}$
10	$6.64\ 10^{-3}$	$2.46\ 10^{-4}$	$7.50\ 10^{-5}$
12	$2.13\ 10^{-2}$	$3.03\ 10^{-3}$	$1.07\ 10^{-4}$
14	$9.15\ 10^{-1}$	$4.12\ 10^{-3}$	$8.75\ 10^{-4}$
16	$4.04\ 10^{-1}$	$2.30\ 10^{-2}$	$9.15\ 10^{-4}$

Confrontando questi risultati con quelli riportati nell'esempio 4.21 relativi al metodo di Gauss, risulta che quando il metodo di Gauss-Jordan è applicato con le varianti del massimo pivot, il suo comportamento è molto simile a quello di Gauss con le stesse varianti e per alcuni valori di n i risultati sono praticamente identici. In questo caso il metodo di Gauss-Jordan con le sua varianti del massimo pivot ha le stesse caratteristiche di stabilità di quello di Gauss con le medesime varianti. ∎

12. Metodo di Householder

Il procedimento di fattorizzazione della matrice A con matrici di Householder è sempre applicabile. Si segue il procedimento di fattorizzazione con le matrici elementari del paragrafo 5: al primo passo, sia $\mathbf{a}_1$ il vettore formato dagli elementi della prima colonna di $A^{(1)} = A$ e sia

$$\theta_1 = \begin{cases} a_{11}^{(1)}/|a_{11}^{(1)}| & \text{se } a_{11}^{(1)} \neq 0, \\ 1 & \text{se } a_{11}^{(1)} = 0; \end{cases}$$

posto

$$\beta_1 = \frac{1}{\|\mathbf{a}_1\|_2(\|\mathbf{a}_1\|_2 + |a_{11}^{(1)}|)}$$

e

$$\mathbf{v}_1 = \begin{bmatrix} \theta_1(\|\mathbf{a}_1\|_2 + |a_{11}^{(1)}|) \\ a_{21}^{(1)} \\ \vdots \\ a_{n1}^{(1)} \end{bmatrix},$$

la prima matrice elementare di Householder è data da

$$E^{(1)} = P^{(1)} = I - \beta_1 \mathbf{v}_1 \mathbf{v}_1^H.$$

Con la notazione del paragrafo 5, al k-esimo passo, sia $\mathbf{a}_k$ il vettore di ordine $n-k+1$ formato dagli elementi della prima colonna di $B^{(k)}$, cioè dagli elementi della k-esima colonna di $A^{(k)}$ con indice di riga maggiore o uguale a k, e sia

$$\theta_k = \begin{cases} a_{kk}^{(k)}/|a_{kk}^{(k)}| & \text{se } a_{kk}^{(k)} \neq 0, \\ 1 & \text{se } a_{kk}^{(k)} = 0; \end{cases}$$

posto

$$\beta_k = \frac{1}{\|\mathbf{a}_k\|_2(\|\mathbf{a}_k\|_2 + |a_{kk}^{(k)}|)}$$

e

$$\mathbf{v}_k = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \theta_k(\|\mathbf{a}_k\|_2 + |a_{kk}^{(k)}|) \\ a_{k+1,k}^{(k)} \\ \vdots \\ a_{nk}^{(k)} \end{bmatrix} \begin{matrix} \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} k-1 \text{ componenti} \\ \left.\begin{matrix} \\ \\ \\ \\ \\ \end{matrix}\right\} n-k+1 \text{ componenti,} \end{matrix}$$

la k-esima matrice elementare di Householder è data da

$$E^{(k)} = P^{(k)} = I - \beta_k \mathbf{v}_k \mathbf{v}_k^H.$$

Se al k-esimo passo si ha $\mathbf{a}_k = \mathbf{0}$, si pone $P^{(k)} = I$, cioè il k-esimo passo non comporta alcuna operazione e la matrice $A^{(n)}$ ottenuta al termine del procedimento avrà nullo l'elemento $a_{kk}^{(n)}$.

Poiché le matrici $P^{(k)}$ sono hermitiane e unitarie, e quindi

$$[P^{(k)}]^{-1} = P^{(k)},$$

la matrice

$$E = [P^{(1)}]^{-1}[P^{(2)}]^{-1} \dots [P^{(n-1)}]^{-1} = P^{(1)}P^{(2)} \dots P^{(n-1)}$$

è unitaria. Se si pone $Q = E$ e $R = A^{(n)}$, dalla (21) si ottiene

$$A = QR,$$

e cioè la matrice A è fattorizzata nel prodotto di una matrice unitaria per una triangolare superiore.

4.25 Esempio. Si calcoli la fattorizzazione QR della matrice

$$A = \begin{bmatrix} 72 & \text{-}144 & \text{-}144 \\ \text{-}144 & \text{-}36 & \text{-}360 \\ \text{-}144 & \text{-}360 & 450 \end{bmatrix}.$$

Al primo passo si ha

$$\beta_1 = \frac{1}{62208}, \quad \mathbf{v}_1 = [288, \text{-}144, \text{-}144]^T,$$

per cui

$$P^{(1)} = I - \beta_1 \mathbf{v}_1 \mathbf{v}_1^T = \frac{1}{6} \begin{bmatrix} \text{-}2 & 4 & 4 \\ 4 & 4 & \text{-}2 \\ 4 & \text{-}2 & 4 \end{bmatrix}$$

e

$$A^{(2)} = \begin{bmatrix} \text{-}216 & \text{-}216 & 108 \\ 0 & 0 & \text{-}486 \\ 0 & \text{-}324 & 324 \end{bmatrix};$$

al secondo passo si ha

$$\beta_2 = \frac{1}{104976}, \quad \mathbf{v}_2 = [0, 324, \text{-}324]^T,$$

per cui

$$P^{(2)} = I - \beta_2 \mathbf{v}_2 \mathbf{v}_2^T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

e

$$R = A^{(3)} = \begin{bmatrix} \text{-}216 & \text{-}216 & 108 \\ 0 & \text{-}324 & 324 \\ 0 & 0 & \text{-}486 \end{bmatrix}.$$

Inoltre

$$Q = P^{(1)} P^{(2)} = \frac{1}{6} \begin{bmatrix} \text{-}2 & 4 & 4 \\ 4 & \text{-}2 & 4 \\ 4 & 4 & \text{-}2 \end{bmatrix}. \qquad \blacksquare$$

13. Implementazione del metodo di Householder

Il metodo di Householder per risolvere il sistema lineare $A\mathbf{x} = \mathbf{b}$ può essere implementato senza calcolare effettivamente le matrici $P^{(k)}$. Si procede nel modo seguente: si considera la matrice

$$T^{(1)} = [\, A^{(1)} \mid \mathbf{b}^{(1)} \,] = [\, A \mid \mathbf{b} \,]$$

e si costruiscono β_1, $\mathbf{v}_1$ e il vettore riga di $n+1$ componenti $\mathbf{y}_1^H = \mathbf{v}_1^H T^{(1)}$. Allora è

$$T^{(2)} = P^{(1)} T^{(1)} = T^{(1)} - \beta_1 \mathbf{v}_1 \mathbf{y}_1^H .$$

Al k-esimo passo si ha:

$$T^{(k+1)} = P^{(k)} T^{(k)} = T^{(k)} - \beta_k \mathbf{v}_k \mathbf{y}_k^H ,$$

dove $\mathbf{y}_k^H = \mathbf{v}_k^H T^{(k)}$ è un vettore riga di $n+1$ componenti, di cui le prime $k-1$ sono nulle. Dopo n passi si ottiene la matrice

$$T^{(n)} = [\, A^{(n)} \mid \mathbf{b}^{(n)} \,] = [R \mid \mathbf{b}^{(n)}],$$

e quindi il sistema $R\mathbf{x} = \mathbf{b}^{(n)}$ con matrice dei coefficienti triangolare superiore è equivalente al sistema $A\mathbf{x} = \mathbf{b}$.

Al k-esimo passo le prime $k-1$ componenti del vettore $\mathbf{v}_k$ sono nulle e quindi $\mathbf{v}_k$ ha al più $n-k+1$ componenti diverse da zero; per la sua determinazione sono richieste al più $n-k+3$ operazioni moltiplicative oltre a una estrazione di radice quadrata; la determinazione del vettore $\mathbf{y}_k$ richiede al più $(n-k+1)^2$ operazioni moltiplicative; il prodotto esterno $\mathbf{v}_k \mathbf{y}_k^H$ richiede $(n-k+1)^2$ operazioni moltiplicative. Quindi il costo computazionale del k-esimo passo è $2(n-k)^2$, e complessivamente il costo computazionale del metodo di Householder per risolvere il sistema $A\mathbf{x} = \mathbf{b}$ è di $2n^3/3$ operazioni, pari al doppio di quello del metodo di Gauss. Sono inoltre richieste n estrazioni di radice quadrata.

Nell'implementazione del metodo, le matrici $A^{(k)}$ sono memorizzate nella stessa area di memoria della matrice A. Poiché gli $n-k$ elementi della k-esima colonna della matrice $A^{(k+1)}$ al di sotto dell'elemento principale sono nulli, nella costruzione di $A^{(k+1)}$ non conviene annullare le corrispondenti posizioni di memoria, che in tal modo alla fine del procedimento contengono le componenti del vettore $\mathbf{v}_k$ di indice maggiore di k. Quindi al k-esimo passo restano da memorizzare β_k e la k-esima componente di $\mathbf{v}_k$. Globalmente per questi elementi sono richieste altre $2n-2$ posizioni di memoria. Al termine del procedimento nello spazio di memoria inizialmente occupato dalla matrice A e nelle $2n-2$ posizioni aggiuntive, risultano memorizzati la matrice R e gli elementi necessari per ricostruire le matrici elementari $P^{(k)}$ e quindi la matrice Q.

Non conviene in generale costruire la matrice Q: in molte applicazioni operare con le matrici $P^{(k)}$ non richiede un costo computazionale maggiore che operare con la matrice Q. Ad esempio il calcolo del prodotto QB, dove B è una matrice di ordine n, richiede lo stesso numero di operazioni moltiplicative, n^3, sia che si moltiplichino le due matrici Q e B, sia che si usino le matrici elementari $P^{(k)}$ con l'algoritmo:

$$\begin{aligned} &B^{(n)} = B, \\ &B^{(k)} = P^{(k)} B^{(k+1)}, \quad \text{per } k = n-1, \ldots, 1 \\ &QB \ = B^{(1)}. \end{aligned}$$

Infatti il calcolo di

$$P^{(k)} B^{(k+1)} = B^{(k+1)} - \beta_k \mathbf{v}_k \mathbf{y}_k^H,$$

dove

$$\mathbf{y}_k^H = \mathbf{v}_k^H B^{(k+1)},$$

richiede $2n(n-k+1)$ operazioni moltiplicative e quindi complessivamente il costo computazionale è n^3.

La matrice Q, se è specificatamente richiesta, può essere così calcolata, partendo dall'ultima matrice $P^{(n-1)}$ fino alla prima $P^{(1)}$:

$$Q = P^{(1)}(P^{(2)} \ldots (P^{(n-2)} P^{(n-1)}) \ldots).$$

Così procedendo si può sfruttare il fatto che il prodotto $P^{(k+1)} \ldots P^{(n-1)}$ coincide con la matrice identica ad eccezione degli elementi delle ultime $n-k$ righe e colonne. In tal modo il costo computazionale del calcolo di Q è pari a $2n^3/3$.

Poiché $A^{-1} = R^{-1} Q^H$, il metodo di Householder può essere usato anche per calcolare l'inversa di una matrice A, con il procedimento

$$B^{(n)} = R^{-1}, \quad B^{(k)} = B^{(k+1)} P^{(k)}, \quad \text{per } k = n-1, \ldots, 1.$$

Il costo computazionale della fattorizzazione QR è $2n^3/3$, del calcolo della matrice triangolare superiore R^{-1} è $n^3/6$, della moltiplicazione delle matrici elementari è $2n^3/3$: quindi complessivamente il costo computazionale del calcolo di A^{-1} con il metodo di Householder è $3n^3/2$.

Il metodo di Householder può essere applicato anche a matrici A non quadrate. Se $A \in \mathbf{C}^{m \times n}$, $m > n$, si ha $A = QR$, dove $Q \in \mathbf{C}^{m \times m}$ è unitaria e

$$R = A^{(n+1)} = \begin{bmatrix} T \\ \\ O \end{bmatrix} \begin{matrix} \} \\ \\ \} \end{matrix} \begin{matrix} n \text{ righe} \\ \\ m-n \text{ righe}, \end{matrix}$$

dove $T \in \mathbf{C}^{n\times n}$ è una matrice triangolare superiore. Il costo computazionale della fattorizzazione QR con il metodo di Householder è in questo caso $n^2(m - n/3)$.

14. Analisi dell'errore del metodo di Householder

Lo studio dell'errore della fattorizzazione QR con il metodo di Householder è assai più complesso che per la fattorizzazione LU. Ci si limita quindi a riportare i risultati fondamentali, limitatamente al caso reale [18]. Nei teoremi che seguono intervengono le seguenti matrici:

$\widetilde{A}^{(k)}$ $k = 2, \ldots, n$, è la k-esima matrice effettivamente calcolata nel procedimento di fattorizzazione QR di A (si noti che $\widetilde{A}^{(1)} = A^{(1)} = A$);

$\widetilde{R}$ è la matrice triangolare superiore $\widetilde{A}^{(n)}$ effettivamente calcolata al termine del procedimento di fattorizzazione,

$\widehat{P}^{(k)}$ è la matrice elementare di Householder che si calcolerebbe partendo da $\widetilde{A}^{(k)}$ se non intervenissero gli errori di arrotondamento,

$\widetilde{P}^{(k)}$ è la matrice effettivamente calcolata partendo da $\widetilde{A}^{(k)}$, cioè tale che

$$\widetilde{A}^{(k+1)} = fl(\widetilde{P}^{(k)}\widetilde{A}^{(k)}),$$

dove $fl(\widetilde{P}^{(k)}\widetilde{A}^{(k)})$ è il risultato del calcolo di $\widetilde{P}^{(k)}\widetilde{A}^{(k)}$ effettivamente eseguito in aritmetica di macchina. Allora la matrice

$$\widehat{Q} = \widehat{P}^{(n-1)} \ldots \widehat{P}^{(1)}$$

è una matrice unitaria, mentre la matrice

$$\widetilde{Q} = fl(\widetilde{P}^{(n-1)}(\ldots fl(\widetilde{P}^{(2)}\widetilde{P}^{(1)})\ldots))$$

in generale non lo è.

Anziché valutare gli errori nei singoli elementi dei vettori, o delle matrici, essi verranno valutati globalmente in norma 2 nel caso di vettori e in norma di Frobenius nel caso di matrici.

4.26 Teorema. *Applicando ad un vettore* $\mathbf{z}$ *successivamente le* $n-1$ *matrici di Householder che intervengono nella fattorizzazione* QR *di* A*, il vettore* $\mathbf{w}$ *effettivamente calcolato*

$$\mathbf{w} = fl(\widetilde{P}^{(n-1)}(\ldots fl(\widetilde{P}^{(1)}\mathbf{z})\ldots)),$$

risulta essere

$$\mathbf{w} = \widehat{Q}(\mathbf{z} + \mathbf{e}),$$

dove

$$\|\mathbf{e}\|_2 \leq \gamma n^2 u \|\mathbf{z}\|_2 + O(u^2),$$

in cui $O(u^2)$ *è una funzione di* u *di ordine maggiore o uguale al secondo e* γ *è una costante positiva.* ∎

Se $A = I$, applicando il teorema precedente al caso degli n vettori $\mathbf{e}_i$, $i = 1, \ldots, n$, si ottiene la relazione

$$\widetilde{Q} = \widehat{Q}(I + E) + O(u^2), \tag{41}$$

con

$$\|E\|_F \leq \gamma n^{5/2} u + O(u^2). \quad ∎$$

Si considerano ora gli errori da cui sono affette le matrici $A^{(k)}$.

4.27 Teorema. *Al* k*-esimo passo si ha:*

$$\widetilde{A}^{(k+1)} = \widehat{P}^{(k)} \ldots \widehat{P}^{(1)}(A + G^{(k)}),$$

dove

$$\|G^{(k)}\|_F \leq \gamma n^2 u \ \|A\|_F + O(u^2). \quad ∎$$

In particolare, per la matrice $\widetilde{R}$ si ha

$$\widetilde{R} = \widehat{Q}(A + G), \tag{42}$$

dove

$$\|G\|_F \leq \gamma n^2 u \|A\|_F + O(u^2).$$

Si osservi che la matrice $\widehat{Q}$ che compare nella (42) non è uguale alla matrice $\widetilde{Q}$ effettivamente calcolata, ma per la (41) differisce "di poco" da questa.

I risultati dei teoremi 4.26 e 4.27 sono ora utilizzati per individuare una limitazione dell'errore nella risoluzione del sistema lineare.

4.28 Teorema. *Sia* $\widetilde{\mathbf{x}}$ *il vettore effettivamente calcolato nella risoluzione del sistema lineare* $A\mathbf{x} = \mathbf{b}$*, effettuata mediante i seguenti passi:*

a) calcolo delle matrici elementari $\widetilde{P}^{(k)}$*,* $k = 1, \ldots, n-1$*,*

b) applicazione successiva di tali matrici alla matrice A *e al vettore* $\mathbf{b}$*, per calcolare la matrice triangolare superiore* $\widetilde{R}$ *e il vettore* $\widetilde{\mathbf{y}}$*,*

c) determinazione del vettore $\widetilde{\mathbf{x}}$ *effettivamente calcolato nella risoluzione del sistema* $\widetilde{R}\mathbf{x} = \widetilde{\mathbf{y}}$*.*

Allora risulta

$$(A + \Delta A)\widetilde{\mathbf{x}} = \mathbf{b} + \Delta\mathbf{b}, \tag{43}$$

con

$$\|\Delta A\|_F \leq u(\gamma n^2 \|A\|_F + n\|\widetilde{R}\|_F) + O(u^2), \quad \|\Delta\mathbf{b}\|_2 \leq \gamma n^2 u \|\mathbf{b}\|_2 + O(u^2).$$

Dim. La matrice $\widetilde{R}$ del sistema al punto c) è triangolare superiore: allora, per il teorema 4.15, la soluzione effettivamente calcolata $\widetilde{\mathbf{x}}$ è tale che

$$(\widetilde{R} + F)\,\widetilde{\mathbf{x}} = \widetilde{\mathbf{y}}, \quad |F| \leq nu\,|\widetilde{R}| + O(u^2), \tag{44}$$

per cui

$$\|F\|_F \leq nu\,\|\widetilde{R}\|_F + O(u^2).$$

Per il teorema 4.26, si ha che per il vettore

$$\widetilde{\mathbf{y}} = fl(\widetilde{P}^{(n-1)} \ldots fl(\widetilde{P}^{(1)}\mathbf{b}) \ldots),$$

vale

$$\widetilde{\mathbf{y}} = \widehat{Q}(\mathbf{b} + \mathbf{e}), \quad \|\mathbf{e}\|_2 \leq \gamma n^2 u\,\|\mathbf{b}\|_2 + O(u^2). \tag{45}$$

Da (44) e (45) segue che

$$(\widetilde{R} + F)\widetilde{\mathbf{x}} = \widehat{Q}(\mathbf{b} + \mathbf{e}),$$

e, per la (42),

$$[\widehat{Q}(A + G) + F]\,\widetilde{\mathbf{x}} = \widehat{Q}(\mathbf{b} + \mathbf{e}).$$

Essendo $\widehat{Q}$ unitaria, si ha:

$$(A + G + \widehat{Q}^H F)\,\widetilde{\mathbf{x}} = \mathbf{b} + \mathbf{e},$$

e quindi vale la (43) con $\Delta\mathbf{b} = \mathbf{e}$, e con

$$\|\Delta A\|_F = \|G + \widehat{Q}^H F\|_F.$$

Poiché $\|\widehat{Q}^H F\|_F = \|F\|_F$, risulta $\|\Delta A\|_F \leq u(\gamma n^2 \|A\|_F + n\|\widetilde{R}\|_F) + O(u^2)$. ∎

Dal teorema 4.28 risulta che la stabilità del metodo di Householder è legata alla norma della matrice A e alla norma della matrice triangolare $\widetilde{R}$, che differisce dalla norma di R per un termine dell'ordine di u. Poiché

tutte le matrici $A^{(k)}$ ottenute con il metodo di Householder hanno la stessa norma di Frobenius, essendo

$$\begin{aligned}\mathrm{tr}([A^{(k+1)}]^H A^{(k+1)}) &= \mathrm{tr}([A^{(k)}]^H [P^{(k)}]^H P^{(k)} A^{(k)}) \\ &= \mathrm{tr}([A^{(k)}]^H A^{(k)}), \quad k = 1, \ldots, n-1\end{aligned}$$

risulta

$$\|A^{(k)}\|_F = \|A\|_F, \quad k = 1, \ldots, n, \text{ e } \quad \|R\|_F = \|A\|_F.$$

Quindi, a differenza del metodo di Gauss, il metodo di Householder non richiede scambi di righe nè richiede applicazioni di strategie, del tipo di quella del massimo pivot, per aumentarne la stabilità.

In alcune applicazioni, come ad esempio per calcolare il rango di una matrice, il metodo di Householder viene utilizzato con una tecnica analoga a quella del massimo pivot totale per il metodo di Gauss, allo scopo di ottenere una matrice R con gli elementi principali ordinati in ordine di modulo non crescente. Per questo al k-esimo passo, costruita la matrice $A^{(k)}$ della forma (18), si determina la colonna di $B^{(k)}$ la cui norma 2 è massima; sia j, $1 \leq j \leq n-k+1$, l'indice di tale colonna: se $j \neq 1$, si scambiano fra loro la $(k+1)$-esima e la $(k+j)$-esima colonna della matrice $A^{(k)}$ e si costruisce la matrice elementare $P^{(k)}$ a partire della matrice così ottenuta. Se il rango di A è $r < n$, il procedimento termina dopo r passi e si ottiene una decomposizione del tipo

$$A\Pi = QR,$$

dove $\Pi \in \mathbf{R}^{n\times n}$ è una matrice di permutazione, $Q \in \mathbf{C}^{n\times n}$ è unitaria e R è della forma

$$R = \begin{bmatrix} R_1 & S \\ O & O \end{bmatrix} \begin{matrix} \} & r & \text{righe} \\ \} & n-r & \text{righe,} \end{matrix}$$

dove $R_1 \in \mathbf{C}^{r\times r}$ è triangolare superiore non singolare, con gli elementi principali ordinati in ordine di modulo non crescente.

La crescita dell'errore algoritmico come funzione di n è in generale molto inferiore a quella indicata dalle maggiorazioni dei teoremi 4.26, 4.27 e 4.28.

4.29 Esempio. Si confrontano i valori di $\epsilon_x = \dfrac{\|\delta\mathbf{x}\|_2}{\|\mathbf{x}\|_2}$, ottenuti applicando al sistema lineare $A\mathbf{x} = \mathbf{b}$ dell'esempio 4.21 il metodo di Householder con i valori ottenuti con il metodo di Gauss con il massimo pivot parziale e totale.

n	Householder	Gauss, pivot parziale	Gauss, pivot totale
4	$1.59\ 10^{-5}$	$1.30\ 10^{-5}$	$1.60\ 10^{-6}$
6	$3.43\ 10^{-5}$	$1.83\ 10^{-5}$	$2.09\ 10^{-5}$
8	$1.54\ 10^{-4}$	$3.54\ 10^{-4}$	$1.93\ 10^{-5}$
10	$6.57\ 10^{-4}$	$3.05\ 10^{-4}$	$7.52\ 10^{-5}$
12	$2.09\ 10^{-3}$	$3.08\ 10^{-3}$	$1.07\ 10^{-4}$
14	$6.56\ 10^{-3}$	$3.01\ 10^{-3}$	$8.75\ 10^{-4}$
16	$3.06\ 10^{-2}$	$2.49\ 10^{-2}$	$9.15\ 10^{-4}$

Risulta quindi che l'errore generato dal metodo di Householder è, in questo caso, confrontabile con quello generato dal metodo di Gauss con pivot parziale. ∎

15. Fattorizzazione QR con le matrici di Givens

L'applicazione di una matrice di Householder alla matrice A consente di annullare simultaneamente tutti gli elementi di una colonna di A al di sotto di quello principale, ed è per questo che il metodo di Householder viene utilizzato per fattorizzare in modo efficiente la matrice A nella forma QR. Vi sono però dei casi, ad esempio per matrici con specifiche proprietà di struttura, in cui è necessario annullare solo particolari elementi. In questi casi è conveniente utilizzare le matrici di Givens.

4.30 Definizione. Siano $1 \leq i, j \leq n$ con $i \neq j$, $\phi \in \mathbf{R}$ e $\psi \in \mathbf{C}$, con $|\psi| = 1$. La matrice

$$
G_{ij} = \begin{bmatrix}
1 & & & & & & & & \\
& \ddots & & & & & & & \\
& & 1 & & & & & & \\
& & & c & & & -\overline{s} & & \\
& & & & 1 & & & & \\
& & & & & \ddots & & & \\
& & & & & & 1 & & \\
& & & s & & & c & & \\
& & & & & & & 1 & \\
& & & & & & & & \ddots \\
& & & & & & & & & 1
\end{bmatrix}
\begin{matrix}
\\ \\ \\ \leftarrow i \\ \\ \\ \\ \leftarrow j \\ \\ \\ \\
\end{matrix}
$$

$$
\begin{matrix}
\\ \\ \\ \uparrow \\ i \\ \\ \\ \uparrow \\ j \\ \\ \\
\end{matrix}
$$

in cui $c = \cos\phi$ e $s = \psi \sin\phi$, è detta matrice di *rotazione di Givens*.

Le matrici di Givens si ottengono dalla matrice identica per mezzo di correzioni di rango 2, sono unitarie e, nel caso reale, esprimono una rotazione di un angolo ϕ nel piano individuato dai vettori $\mathbf{e}_i$ ed $\mathbf{e}_j$. La matrice $G_{12} \in \mathbf{R}^{2\times 2}$ è quella dell'esempio 2.2. L'interpretazione geometrica dell'applicazione di G_{12} ad un vettore $\mathbf{x} \in \mathbf{R}^2$ è riportata nella figura 4.4.

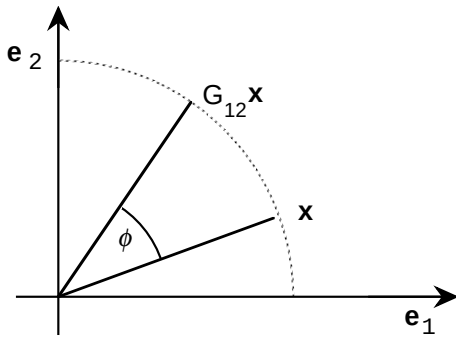


Fig. 4.4 - Rotazione di un vettore $\mathbf{x} \in \mathbf{R}^2$.

4.31 Esempio. Una matrice ortogonale $U \in \mathbf{R}^{2\times 2}$, o è una matrice di Householder, o è una matrice di Givens. Infatti, sia

$$U = \begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix}.$$

Dalla ortogonalità di U, essendo

$$\alpha^2 + \gamma^2 = 1,$$

si può porre

$$\alpha = \cos\phi \quad \text{e} \quad \gamma = \sin\phi, \quad \phi \in \mathbf{R},$$

ed essendo

$$\alpha\beta + \gamma\delta = 0,$$

deve risultare

$$\beta = -\sin\phi \quad \text{e} \quad \delta = \cos\phi,$$

oppure

$$\beta = \sin\phi \quad \text{e} \quad \delta = -\cos\phi.$$

Nel primo caso la matrice U è la seguente matrice di Givens

$$U = \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix},$$

nel secondo caso si ha

$$U = \begin{bmatrix} \cos\phi & \sin\phi \\ \sin\phi & \text{-}\cos\phi \end{bmatrix} = I - 2\mathbf{v}\mathbf{v}^T,$$

dove

$$\mathbf{v} = [\text{-}\sin\frac{\phi}{2}\ ,\ \cos\frac{\phi}{2}]^T,$$

e quindi U è una matrice elementare di Householder. ∎

Sia $A \in \mathbf{C}^{n\times n}$, fissati gli indici i e j, la matrice $B = G_{ij}A$ differisce dalla matrice A per i soli elementi delle righe di indici i e j. Se $a_{ji} \neq 0$, è possibile scegliere ϕ e ψ in modo tale che l'elemento b_{ji} sia uguale a 0. Se $a_{ii} \neq 0$, posto

$$\begin{aligned} \psi &= -\ \frac{a_{ji}}{a_{ii}}\ \left|\frac{a_{ii}}{a_{ji}}\right|, \\ \cos\phi &= \frac{|a_{ii}|}{\sqrt{|a_{ii}|^2 + |a_{ji}|^2}}, \\ \sin\phi &= \frac{|a_{ji}|}{\sqrt{|a_{ii}|^2 + |a_{ji}|^2}}, \end{aligned} \tag{46}$$

risulta

$$b_{ji} = sa_{ii} + ca_{ji} = \psi\sin\phi\ a_{ii} + \cos\phi\ a_{ji} = \frac{\psi|a_{ji}|a_{ii} + |a_{ii}|a_{ji}}{\sqrt{|a_{ii}|^2 + |a_{ji}|^2}}$$

$$= \frac{|a_{ji}|a_{ii}}{\sqrt{|a_{ii}|^2 + |a_{ji}|^2}} \left[\psi + \frac{a_{ji}}{a_{ii}}\ \left|\frac{a_{ii}}{a_{ji}}\right|\ \right] = 0.$$

Se $a_{ii} = 0$, ψ può assumere un qualsiasi valore.

Se $A \in \mathbf{R}^{n\times n}$, la matrice di Givens G_{ij} è reale e

$$\begin{aligned} c &= \frac{|a_{ii}|}{\sqrt{|a_{ii}|^2 + |a_{ji}|^2}}, \\ s &= -\ \mathrm{sgn}\left(\frac{a_{ji}}{a_{ii}}\right) \frac{|a_{ji}|}{\sqrt{|a_{ii}|^2 + |a_{ji}|^2}}. \end{aligned}$$

Al posto delle (46) conviene usare le seguenti formule numericamente più stabili:

$$\text{se } |a_{ji}| \geq |a_{ii}|,\ \text{si pone } t = \left|\frac{a_{ii}}{a_{ji}}\right|,\ \sin\phi = \frac{1}{\sqrt{1+t^2}},\quad \cos\phi = t\sin\phi,$$

$$\text{se } |a_{ji}| < |a_{ii}|,\ \text{si pone } t = \left|\frac{a_{ji}}{a_{ii}}\right|,\quad \cos\phi = \frac{1}{\sqrt{1+t^2}},\ \sin\phi = t\cos\phi.$$

La fattorizzazione QR con le matrici di Givens viene realizzata costruendo a partire dalla matrice A una successione di matrici, in cui ogni matrice è ottenuta moltiplicando a sinistra la matrice precedentemente calcolata per la matrice $G_{ij},\ j = i+1,\dots,n,\ i = 1,\dots,n-1$, che è tale da annullare l'elemento di posto (j,i). Ad esempio, se $n = 4$, la matrice R è ottenuta nel modo seguente

$$R = G_{34}G_{24}G_{23}G_{14}G_{13}G_{12}A,$$

e quindi in questo caso la matrice Q^H è data dal prodotto

$$Q^H = G_{34}G_{24}G_{23}G_{14}G_{13}G_{12}.$$

Questo algoritmo richiede per il calcolo di R $4n^3/3$ operazioni moltiplicative. Infatti per ogni i e j, $j = i+1,\dots,n,\ i = 1,\dots,n-1$, la costruzione di G_{ij} richiede 4 operazioni moltiplicative e 1 estrazione di radice quadrata, e la moltiplicazione per G_{ij} richiede $4(n-i+1)$ operazioni moltiplicative. Complessivamente sono richieste

$$\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} 4(n-i+1)+4 \simeq 4\sum_{i=1}^{n}\sum_{j=i}^{n}(n-i) = 4\sum_{i=1}^{n}(n-i)^2 \simeq \frac{4n^3}{3}$$

operazioni moltiplicative.

4.32 Esempio. Sia

$$A = \begin{bmatrix} -2 & \sqrt{2} & 0 & 1 \\ 2 & 0 & 2 & -2 \\ 0 & -1 & \sqrt{2} & \frac{3}{2}\sqrt{2} \\ -2\sqrt{2} & 1 & -3\sqrt{2} & -\frac{1}{2}\sqrt{2} \end{bmatrix}.$$

Si applica ad A il metodo di Givens

$$G_{12} = \begin{bmatrix} \frac{1}{2}\sqrt{2} & -\frac{1}{2}\sqrt{2} & 0 & 0 \\ \frac{1}{2}\sqrt{2} & \frac{1}{2}\sqrt{2} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$A^{(2)} = G_{12}A = \begin{bmatrix} -2\sqrt{2} & 1 & -\sqrt{2} & \frac{3}{2}\sqrt{2} \\ 0 & 1 & \sqrt{2} & -\frac{1}{2}\sqrt{2} \\ 0 & -1 & \sqrt{2} & \frac{3}{2}\sqrt{2} \\ -2\sqrt{2} & 1 & -3\sqrt{2} & -\frac{1}{2}\sqrt{2} \end{bmatrix},$$

$$G_{13} = I,$$

$$G_{14} = \begin{bmatrix} \frac{1}{2}\sqrt{2} & 0 & 0 & \frac{1}{2}\sqrt{2} \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -\frac{1}{2}\sqrt{2} & 0 & 0 & \frac{1}{2}\sqrt{2} \end{bmatrix},$$

$$A^{(3)} = G_{14}A^{(2)} = \begin{bmatrix} -4 & \sqrt{2} & -4 & 1 \\ 0 & 1 & \sqrt{2} & -\frac{1}{2}\sqrt{2} \\ 0 & -1 & \sqrt{2} & \frac{3}{2}\sqrt{2} \\ 0 & 0 & -2 & -2 \end{bmatrix},$$

$$G_{23} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \frac{1}{2}\sqrt{2} & -\frac{1}{2}\sqrt{2} & 0 \\ 0 & \frac{1}{2}\sqrt{2} & \frac{1}{2}\sqrt{2} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$A^{(4)} = G_{23}A^{(3)} = \begin{bmatrix} -4 & \sqrt{2} & -4 & 1 \\ 0 & \sqrt{2} & 0 & -2 \\ 0 & 0 & 2 & 1 \\ 0 & 0 & -2 & -2 \end{bmatrix},$$

$$G_{24} = I,$$

$$G_{34} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{1}{2}\sqrt{2} & -\frac{1}{2}\sqrt{2} \\ 0 & 0 & \frac{1}{2}\sqrt{2} & \frac{1}{2}\sqrt{2} \end{bmatrix},$$

$$R = A^{(5)} = G_{34}A^{(4)} = \begin{bmatrix} -4 & \sqrt{2} & -4 & 1 \\ 0 & \sqrt{2} & 0 & -2 \\ 0 & 0 & 2\sqrt{2} & \frac{3}{2}\sqrt{2} \\ 0 & 0 & 0 & -\frac{1}{2}\sqrt{2} \end{bmatrix}.$$

Si ha perciò $A = QR$, dove

$$Q^H = G_{34}G_{23}G_{14}G_{12} = \frac{1}{2}\begin{bmatrix} 1 & -1 & 0 & \sqrt{2} \\ 1 & 1 & -\sqrt{2} & 0 \\ \sqrt{2} & 0 & 1 & -1 \\ 0 & \sqrt{2} & 1 & 1 \end{bmatrix}.$$

∎

Il metodo di Givens per la fattorizzazione QR richiede quindi circa il doppio delle operazioni richieste dal metodo di Householder. Perciò è conveniente utilizzare questo metodo solo quando la matrice A ha particolari proprietà di struttura: in tal caso, se la matrice A è sparsa, il metodo di Givens può essere competitivo con quello di Householder. Più in generale si utilizza il metodo di Givens quando è richiesto l'annullamento selettivo di particolari elementi della matrice A, come nel caso di alcuni metodi per

calcolare autovalori di matrici e risolvere problemi di minimi quadrati (si vedano i capitoli 6 e 7). Dal punto di vista della stabilità numerica, il metodo di Givens ha un comportamento analogo a quello di Householder.

4.33 Esempio. Se $A \in \mathbf{C}^{n\times n}$ è tridiagonale, il numero di rotazioni di Givens richieste è dato da $n-1$. Cioè

$$R = G_{n-1,n}G_{n-2,n-1}\dots G_{12}A.$$

Il costo computazionale del metodo di Givens è in questo caso di $12n$ operazioni moltiplicative ed n estrazioni di radici quadrate, ed è uguale a quello richiesto dal metodo di Householder. ∎

16. Tecniche compatte

Le matrici L e U della fattorizzazione possono essere calcolate oltre che con il metodo di Gauss che utilizza matrici elementari, anche con le tecniche di seguito riportate. Dalla relazione $A = LU$, poiché L e U sono matrici rispettivamente triangolare inferiore e triangolare superiore, si ha:

$$a_{ij} = \sum_{k=1}^{\min(i,j)} l_{ik}u_{kj}, \quad i,j = 1,\dots,n. \tag{47}$$

Seguendo un ordine particolare la (47) consente di calcolare gli elementi di L e di U. Procedendo per righe si ha:

$$a_{ij} = \sum_{k=1}^{i-1} l_{ik}u_{kj} + u_{ij}, \quad j = i,\dots,n, \quad i = 1,\dots,n, \tag{48}$$

$$a_{ij} = \sum_{k=1}^{j-1} l_{ik}u_{kj} + l_{ij}u_{jj}, \quad j = 1,\dots,i-1, \quad i = 2,\dots,n, \tag{49}$$

dove il risultato delle sommatorie di (48) e (49) è da intendersi uguale a zero se il secondo estremo risulta nullo. Dalle (48) e (49) si ha:

$$u_{ij} = a_{ij} - \sum_{k=1}^{i-1} l_{ik}u_{kj}, \quad j = i,\dots,n, \quad i = 1,\dots,n, \tag{50}$$

$$l_{ij} = \frac{1}{u_{jj}}\left[a_{ij} - \sum_{k=1}^{j-1} l_{ik}u_{kj}\right], \quad j = 1,\dots,i-1, \quad i = 2,\dots,n, \tag{51}$$

dove il risultato delle sommatorie di (50) e (51) è da intendersi uguale a zero se il secondo estremo risulta nullo. Il calcolo procede allora nel modo seguente:

per $i = 1,$ dalla (50) si ha $u_{1j} = a_{1j},\ j = 1,\dots,n$;

per $i = 2$, dalla (51) si ha $l_{21} = a_{21}/u_{11}$ e dalla (50) $u_{2j} = a_{2j} - l_{21}u_{1j},\ j = 2, \ldots, n$; si calcolano l_{21} e $u_{2j},\ j = 2, \ldots, n$ utilizzando gli u_{1j} calcolati precedentemente;

$\vdots$

per $i = n$, dalla (51) si possono calcolare gli $l_{nj},\ j = 1, \ldots, n-1$, utilizzando gli $l_{nk},\ k < j$, della stessa riga già calcolati e gli u_{kj} calcolati precedentemente, e dalla (50) si può calcolare u_{nn}.

Per $n = 4$ l'ordinamento con cui vengono ricavati gli elementi di L e di U è schematizzato nella figura 4.5.

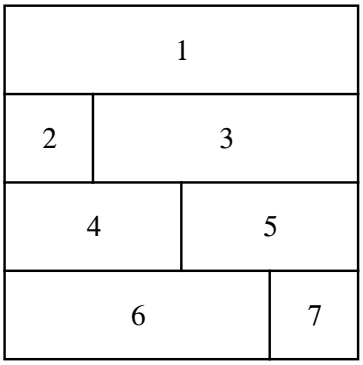


Fig.4.5 - Ordinamento del calcolo nella tecnica compatta.

La determinazione di L e U può essere ottenuta anche seguendo l'ordinamento di *Crout*, come è indicato per il caso $n = 4$ nella figura 4.6.

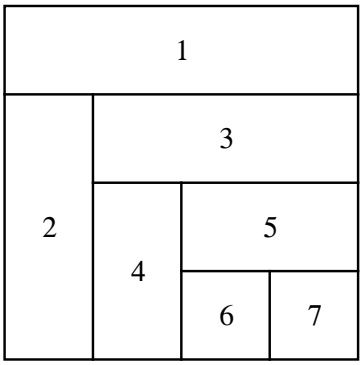


Fig. 4.6 - Ordinamento nel metodo di Crout.

Il costo computazionale di questi metodi è lo stesso, qualunque sia l'ordinamento che si segue, ed è uguale a quello richiesto dal metodo di Gauss. Questi metodi non sono applicabili a matrici che non ammettono

la fattorizzazione LU, non essendo possibile effettuare scambi di righe o di colonne. Anche le tecniche di massimo pivot non sono applicabili, per cui se si presentano problemi di instabilità numerica, un modo per contenere gli errori di arrotondamento è quello di calcolare le quantità

$$\sum_{k=1}^{\min(i,j)} l_{ik}u_{kj}$$

delle (48) e (49) usando una precisione superiore a quella usata negli altri calcoli.

17. Metodo di Cholesky

Nel paragrafo 3 è stato dimostrato che una matrice A definita positiva è fattorizzabile nel prodotto

$$A = LL^H,$$

cioè in componenti

$$a_{ij} = \sum_{k=1}^{\min(i,j)} l_{ik}\bar{l}_{jk}.$$

Poiché la matrice A è hermitiana, è possibile eseguire il calcolo in modo da utilizzare solo gli elementi $a_{ij},\ i \geq j$; si ha

$$\begin{aligned} a_{jj} &= \sum_{k=1}^{j} |l_{jk}|^2, \quad \text{per } j = 1, \ldots, n, \\ a_{ij} &= \sum_{k=1}^{j} l_{ik}\bar{l}_{jk}, \quad \text{per } i = j+1, \ldots, n, \quad j = 1, \ldots, n-1, \end{aligned} \tag{52}$$

da cui si ricavano le seguenti relazioni che definiscono il *metodo di Cholesky*

$$\left.\begin{aligned} l_{jj} &= \sqrt{a_{jj} - \sum_{k=1}^{j-1} |l_{jk}|^2}, \\ l_{ij} &= \frac{1}{l_{jj}}\left[a_{ij} - \sum_{k=1}^{j-1} l_{ik}\bar{l}_{jk}\right], \quad \begin{array}{l}\text{per } i = j+1, \ldots, n \\ \text{e } \ j \neq n,\end{array} \end{aligned}\right\} \quad \text{per } j = 1, \ldots, n,$$

dove il risultato delle sommatorie è da intendersi uguale a zero se il secondo estremo risulta nullo.

Quindi il calcolo degli elementi di L procede per colonne, come è schematizzato per il caso $n = 4$ nella figura 4.7.

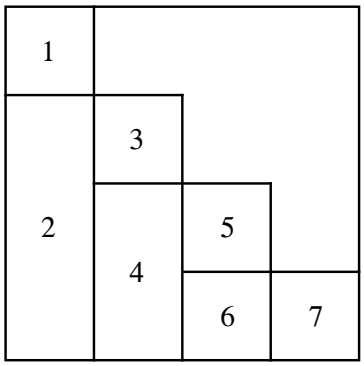


Fig. 4.7 - Ordinamento nel metodo di Cholesky.

4.34 Esempio. La matrice

$$A = \begin{bmatrix} 9 & 1+2\mathbf{i} & \text{-}1+2\mathbf{i} \\ 1-2\mathbf{i} & 6 & \text{-}1-2\mathbf{i} \\ \text{-}1-2\mathbf{i} & \text{-}1+2\mathbf{i} & 9 \end{bmatrix}$$

per il teorema 2.41 è definita positiva. Applicando il metodo di Cholesky si ha:

$$\begin{aligned} l_{11} &= \sqrt{a_{11}} = 3, \\ l_{21} &= \frac{a_{21}}{l_{11}} = \frac{1-2\mathbf{i}}{3}, \quad l_{22} = \sqrt{a_{22} - |l_{21}|^2} = \frac{7}{3}, \\ l_{31} &= \frac{a_{31}}{l_{11}} = -\frac{1+2\mathbf{i}}{3}, \quad l_{32} = \frac{a_{32} - l_{31}\bar{l}_{21}}{l_{22}} = \frac{\text{-}12+22\mathbf{i}}{21}, \\ & \qquad\qquad l_{33} = \sqrt{a_{33} - |l_{31}|^2 - |l_{32}|^2} = \frac{2\sqrt{86}}{7}. \end{aligned}$$

Quindi la matrice L, tale che $A = LL^H$, è

$$L = \begin{bmatrix} 3 & 0 & 0 \\ \dfrac{1-2\mathbf{i}}{3} & \dfrac{7}{3} & 0 \\ -\dfrac{1+2\mathbf{i}}{3} & \dfrac{\text{-}12+22\mathbf{i}}{21} & \dfrac{2\sqrt{86}}{7} \end{bmatrix}$$

∎

Il metodo di Cholesky, come le altre tecniche compatte esposte nel paragrafo 16, non consente di usare varianti di massimo pivot, ma risulta comunque stabile: è possibile infatti dimostrare un risultato analogo a quello

del teorema 4.17, in cui la stabilità del metodo è legata al modulo degli elementi della matrice L. Dalla (52) si ha che

$$|l_{ik}| \leq \sqrt{a_{ii}}, \quad k = 1, \ldots, i \quad i = 1, \ldots, n,$$

e quindi tutti gli elementi di L sono limitati da

$$l_M \leq \sqrt{a_M},$$

in cui l_M e a_M sono rispettivamente il massimo modulo degli elementi di L e di A.

Per il calcolo di l_{ij} sono richieste j operazioni moltiplicative e quindi per il calcolo degli elementi della i-esima riga sono richieste

$$\sum_{j=1}^{i-1} j \simeq \frac{i^2}{2}$$

operazioni moltiplicative, oltre al numero delle operazioni per il calcolo di l_{ii} che è di ordine inferiore (compresa una estrazione di radice quadrata). Per il calcolo delle n righe di L il numero delle operazioni moltiplicative richieste è dato da

$$\sum_{i=1}^{n} \frac{i^2}{2} \simeq \frac{n^3}{6}.$$

Anche il metodo di Gauss, che nel caso generale ha un costo computazionale di $n^3/3$, nel caso di matrici definite positive può essere implementato in modo che il costo computazionale si riduca a $n^3/6$ (si veda l'esercizio 4.22)

18. Considerazioni sul costo computazionale.

La tabella di figura 4.8, ripresa da [17], riporta, a meno dei termini di ordine inferiore, il numero di operazioni richieste per risolvere il sistema lineare $A\mathbf{x} = \mathbf{b}$ di ordine n con i vari metodi diretti presentati.

Nel 1965 è stato dimostrato in [16] che per risolvere il sistema lineare $A\mathbf{x} = \mathbf{b}$, con matrice A qualsiasi, il metodo di Gauss è ottimo, dal punto di vista della complessità computazionale, fra i metodi che utilizzano solo combinazioni lineari di righe e colonne. Recentemente però sono stati proposti metodi basati su tecniche diverse, che permettono di risolvere un sistema di equazioni lineari con un costo computazionale di ordine inferiore. Tali metodi si basano sull'equivalenza, dal punto di vista del costo computazionale, del problema dell'inversione di una matrice di ordine n e del problema della moltiplicazione di due matrici di ordine n.

metodo	oper. moltipl.	oper. addit.	rad. quadr.
Gauss	$n^3/3$	$n^3/3$	-
Gauss-Jordan	$n^3/2$	$n^3/2$	-
Householder	$2n^3/3$	$2n^3/3$	$2n$
Givens	$4n^3/3$	$2n^3/3$	$n^2/2$
Cholesky$^{(*)}$	$n^3/6$	$n^3/6$	n

Fig. 4.8 - Costo computazionale dei metodi diretti.
$^{(*)}$ il metodo di Cholesky si può applicare solo nel caso di matrici definite positive.

4.35 Teorema. *Se il numero delle operazioni aritmetiche sufficienti a calcolare il prodotto di due matrici di ordine n è al più kn^θ, k, θ costanti positive, $2 \leq \theta \leq 3$, allora hn^θ operazioni aritmetiche, h costante positiva, sono sufficienti a invertire una matrice non singolare di ordine n.*

Dim. Sia $A \in \mathbf{C}^{n\times n}$ una matrice non singolare e si supponga per semplicità che $n = 2^p$, p intero positivo (per n qualsiasi è sufficiente considerare la matrice

$$A' = \begin{bmatrix} A & O \\ O & I_m \end{bmatrix},$$

dove $m = 2^p - n$, $p = \lceil \log n \rceil$). Vale la relazione

$$A^{-1} = (A^H A)^{-1} A^H$$

e per $n \geq 2$ si calcola $B = (A^H A)^{-1}$ nel seguente modo: si partiziona la matrice B

$$B = \begin{bmatrix} B_{11} & B_{12} \\ B_{12}^H & B_{22} \end{bmatrix},$$

dove $B_{ij} \in \mathbf{C}^{n/2\times n/2}$ per $i, j = 1, 2$. Poiché B è definita positiva, B_{11} è definita positiva, e quindi è non singolare, e anche il *complemento di Schur* di B_{11} $S = B_{22} - B_{12}^H B_{11}^{-1} B_{12}$ (si veda l'esercizio 1.43) risulta non singolare e vale

$$B^{-1} = \begin{bmatrix} B_{11}^{-1} + B_{11}^{-1} B_{12} S^{-1} \left[B_{11}^{-1} B_{12}\right]^H & -B_{11}^{-1} B_{12} S^{-1} \\ -\left[B_{11}^{-1} B_{12} S^{-1}\right]^H & S^{-1} \end{bmatrix}.$$

Poiché B^{-1} è definita positiva, anche S è definita positiva, e quindi è possibile ripetere lo stesso procedimento per calcolare le inverse di B_{11} e S. Indicando con $I(n)$ il numero di operazioni sufficienti a invertire B, vale la relazione

$$I(n) = 2I\left(\frac{n}{2}\right) + 4M\left(\frac{n}{2}\right) + 2A\left(\frac{n}{2}\right),$$

dove si è indicato con $M\left(\frac{n}{2}\right)$ e $A\left(\frac{n}{2}\right)$ il numero delle operazioni sufficienti a calcolare il prodotto e la somma di matrici di ordine $\frac{n}{2}$. Quindi, poiché

$$A\left(\frac{n}{2}\right) = \left(\frac{n}{2}\right)^2 \quad \text{e} \quad M\left(\frac{n}{2}\right) \leq k\left(\frac{n}{2}\right)^\theta$$

e inoltre $I(1) = 1$, si ha

$$\begin{aligned} I(n) &\leq 2I\left(\frac{n}{2}\right) + (4k+2)\left(\frac{n}{2}\right)^\theta \\ I(1) &= 1, \end{aligned}$$

da cui

$$I(n) \leq \left(\frac{n}{2}\right)^\theta (4k+2) \sum_{i=0}^{p-1} 2^{(1-\theta)i} < \sigma n^\theta, \quad \text{con } \sigma = \frac{2k+1}{2^{\theta-1}-1}.$$

Quindi il numero di operazioni aritmetiche sufficienti a invertire la matrice A è minore di hn^θ, con $h = \sigma + k$. ∎

Nel 1969 Strassen [24] ha individuato il seguente metodo per calcolare il prodotto $C = AB$ di due matrici A e B di ordine 2 con 7 moltiplicazioni e 18 addizioni

$$\begin{array}{ll} s_1 = (a_{11} + a_{22})(b_{11} + b_{22}) & s_2 = (a_{21} + a_{22})b_{11} \\ s_3 = a_{11}(b_{12} - b_{22}) & s_4 = a_{22}(b_{21} - b_{11}) \\ s_5 = (a_{11} + a_{12})b_{22} & s_6 = (a_{21} - a_{11})(b_{11} + b_{12}) \\ s_7 = (a_{12} - a_{22})(b_{21} + b_{22}) & \\ c_{11} = s_1 + s_4 - s_5 + s_7 & c_{12} = s_3 + s_5 \\ c_{21} = s_2 + s_4 & c_{22} = s_1 - s_2 + s_3 + s_6. \end{array}$$

Poiché in queste relazioni non viene utilizzata la proprietà commutativa della moltiplicazione, è possibile applicare tali formule anche nel caso in cui gli elementi a_{ij}, b_{ij}, c_{ij} sono sostituiti con matrici A_{ij}, B_{ij}, C_{ij}.

Se A e B sono due matrici di ordine $n = 2^p$, p intero maggiore di 1, si partizionano le matrici A, B e C in sottomatrici di ordine $n/2$

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \quad B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}, \quad C = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}$$

e si applicano in modo ricorsivo le relazioni precedenti, eseguendo cioè ciascuna delle 7 moltiplicazioni di matrici di ordine $n/2$ con lo stesso metodo. Se $M(n)$ denota il numero di operazioni aritmetiche sufficienti a moltiplicare matrici di ordine n con questo algoritmo, si ha allora

$$\begin{aligned} M(n) &= 7M\left(\frac{n}{2}\right) + O(n^2) \\ M(1) &= 1, \end{aligned}$$

da cui si ottiene $M(n) = 7^p + O(n^2) = n^\theta + O(n^2)$, dove $\theta = \log_2 7 = 2.807\ldots$.

Successivamente l'ordine della complessità computazionale della moltiplicazione di matrici è stato ridotto a $n^\phi, \phi = 2.38\ldots$ [7]. Il problema della determinazione di un algoritmo asintoticamente ottimo è ancora aperto: risulta comunque che kn^2 operazioni sono necessarie per moltiplicare matrici di ordine n, dove k è una costante. È opportuno rilevare che alcuni di questi metodi, che sono asintoticamente più veloci di quelli esposti, hanno solo interesse teorico, in quanto diventano convenienti per valori molto elevati di n.

Esercizi proposti

4.1 Si calcoli il numero di condizionamento in norma 2 e in norma ∞ delle seguenti matrici

$$A = \begin{bmatrix} 1 & 2 \\ 1.001 & 2 \end{bmatrix}, \quad B = \begin{bmatrix} 39 & 16 \\ 71 & 29 \end{bmatrix}, \quad C = \begin{bmatrix} 100 & 99 \\ 99 & 98 \end{bmatrix},$$

$$D = \begin{bmatrix} 1 & -1 & 1 \\ -1 & \epsilon & \epsilon \\ 1 & \epsilon & \epsilon \end{bmatrix}, \ 0 < \epsilon < \frac{1}{2}\,.$$

(Risposta: $\mu_2(A) = 5001$, $\mu_\infty(A) = 6002$, $\mu_2(B) = 1532$, $\mu_\infty(B) = 2200$, $\mu_2(C) = 39206$, $\mu_\infty(C) = 39601$, $\mu_2(D) = \dfrac{1}{\epsilon}$, $\mu_\infty(D) = \dfrac{3}{2}(1 + \dfrac{1}{\epsilon})$.)

4.2 Sia $A \in \mathbf{C}^{n\times n}$, triangolare e non singolare. Si dimostri che

$$\mu_2(A) \geq \frac{\max\limits_{i=1,\ldots,n} |a_{ii}|}{\min\limits_{i=1,\ldots,n} |a_{ii}|}\,.$$

(Traccia: è $\|A\|_2^2 \geq \max\limits_{i=1,\ldots,n} \|A\mathbf{e}_i\|_2^2 \geq \max\limits_{i=1,\ldots,n} |a_{ii}|^2$. Per A^{-1} si proceda in modo analogo.)

4.3 Sia $A \in \mathbf{C}^{n\times n}$ non singolare. Si dimostri che $\mu_2(A) = 1$ se e solo se αA è una matrice unitaria per qualche $\alpha \in \mathbf{C},\ \alpha \neq 0$.
(Traccia: sia $\mu_2(A) = 1$; poiché (si veda l'esercizio 2.13)

$$\|A^{-1}\|_2^2 = \rho[(AA^H)^{-1}] = \rho[(A^HA)^{-1}] = \frac{1}{\min\limits_{i=1,\ldots,n} \lambda_i},$$

dove λ_i sono gli autovalori di A^HA, ne segue che $\max\limits_{i=1,\ldots,n} \lambda_i / \min\limits_{i=1,\ldots,n} \lambda_i = 1$.)

4.4 Si studi il numero di condizionamento in norma ∞ della matrice

$$A = \begin{bmatrix} \text{cotg } \alpha & \text{cosec } \alpha \\ -\text{cosec } \alpha & -\text{cotg } \alpha \end{bmatrix}$$

come funzione di α.
(Risposta: $\mu_\infty(A) = \dfrac{(1 + |\cos\alpha|)^2}{\sin^2\alpha}$; la matrice A è mal condizionata per $\alpha \approx k\pi$, k intero.)

4.5 Data la matrice

$$A = \begin{bmatrix} \alpha & 3\alpha \\ 1 & 1 \end{bmatrix}, \quad \alpha > 0$$

si determini il valore del parametro α per cui il numero di condizionamento $\mu_\infty(A)$ è minimo.
(Risposta: $\alpha = 1/2$.)

4.6 Si determini una maggiorazione del numero di condizionamento in norma 2 della matrice

$$A = \begin{bmatrix} 6.4 & 0.2 & 0.4 \\ 0.2 & 4.2 & 0.5 \\ 0.4 & 0.5 & 7.1 \end{bmatrix}.$$

(Traccia: si tenga conto che la matrice è simmetrica e si applichi la (5) e il primo teorema di Gerschgorin.)

4.7 Sia $A \in \mathbf{C}^{n\times n}$ non singolare.

a) Si dimostri che per il numero di condizionamento di A vale la limitazione inferiore

$$\mu(A) \geq \frac{\|A\|}{\|\Delta A\|},$$

per ogni $\Delta A \in \mathbf{C}^{n\times n}$ tale che $A + \Delta A$ sia singolare, e che il segno di uguaglianza vale se la norma usata è la norma 2 e

$$\Delta A = -\frac{\mathbf{y}\mathbf{z}^H}{\mathbf{z}^H\mathbf{z}},$$

in cui $\mathbf{y}$ è un vettore tale che $\|\mathbf{y}\|_2 = 1$ e $\|A^{-1}\mathbf{y}\| = \|A^{-1}\|$ e $\mathbf{z} = A^{-1}\mathbf{y}$;

b) si diano delle limitazioni inferiori del numero di condizionamento delle matrici dell'esercizio 4.1.

(Traccia: a) esiste $\mathbf{x} \neq \mathbf{0}$ tale che $A(I + A^{-1}\Delta A)\mathbf{x} = \mathbf{0}$, per cui

$$\frac{\|A^{-1}\Delta A\mathbf{x}\|}{\|\mathbf{x}\|} = 1;$$

si verifichi che $A - \dfrac{\mathbf{y}\mathbf{z}^H}{\mathbf{z}^H\mathbf{z}}$ è singolare e che $\|\Delta A\|_2 = \dfrac{1}{\|A^{-1}\|_2}$, in quanto $\|\mathbf{y}\mathbf{z}^H\|_2 = \|\mathbf{y}\|_2\,\|\mathbf{z}\|_2$ (si veda l'esercizio 3.15); b) per opportune scelte della matrice ΔA si ottengono le limitazioni $\mu_\infty(A) \geq 3001$, $\mu_\infty(B) \geq 1420$, $\mu_\infty(C) \geq 19900$, $\mu_\infty(D) \geq \dfrac{3}{2\epsilon}$.)

4.8 Sia $A \in \mathbf{C}^{n\times n}$. Si dimostri che se A è non singolare, allora

$$\mu_2(A) \geq \frac{\|A\|_2}{\|\mathbf{a}_i - \mathbf{a}_j\|_2},$$

per ogni coppia $\mathbf{a}_i$ e $\mathbf{a}_j$, $i \neq j$, di colonne di A.

(Traccia: se A è non singolare, sia $\Delta A = (\mathbf{a}_i - \mathbf{a}_j)\mathbf{e}_j^T$ e quindi $A + \Delta A$ è singolare. Riferendosi all'esercizio 4.7, si determini $\|\Delta A\|_2$.)

4.9 Sia $A \in \mathbf{C}^{n\times n}$ non singolare.

a) Si dica di quanto possono differire al più i numeri di condizionamento di A rispetto alla norma 2 e rispetto alla norma ∞, cioè si determinino due costanti α e $\beta \in \mathbf{R}$, $0 < \alpha \leq \beta$, tali che

$$\alpha\mu_\infty(A) \leq \mu_2(A) \leq \beta\mu_\infty(A);$$

b) se A è hermitiana, si dimostri che

$$\mu_2(A) \leq \mu(A),$$

in cui $\mu(A)$ è il numero di condizionamento di A rispetto a una qualunque norma indotta;

c) si consideri in particolare il caso delle matrici $A \in \mathbf{R}^{2\times 2}$ della forma

$$A = \begin{bmatrix} a & b \\ b & a \end{bmatrix}, \quad a,\ b \in \mathbf{R}.$$

(Traccia: a) si sfruttino le relazioni del paragrafo 4, capitolo 3; b) vale $\|A\|_2 = \rho(A) \leq \|A\|$; c) $\mu_2(A) = \mu_\infty(A) = \dfrac{|a| + |b|}{\big||a| - |b|\big|}$.)

4.10 Sia $L \in \mathbf{C}^{n\times n}$ una matrice triangolare inferiore con elementi principali uguali a 1 ed elementi non principali di modulo minore o uguale a 1.

a) Si determini il massimo valore possibile per $\mu_\infty(L)$;

b) si indichi una matrice per la quale tale massimo viene raggiunto.

(Risposta: a) $n2^{n-1}$; b)

$$l_{ij} = \begin{cases} 1 & \text{se } i = j, \\ -1 & \text{se } i > j, \\ 0 & \text{se } i < j. \end{cases}$$

4.11 Sia $A \in \mathbf{C}^{n\times n}$ una matrice non singolare e siano ΔA, $\delta\mathbf{a}$ e α una matrice, un vettore e uno scalare di perturbazione. Si dimostri che

a) la matrice $\widetilde{A} = A + \Delta A$ è non singolare se $A^{-1}\Delta A$ non ha l'autovalore -1;

b) la matrice $\widetilde{A}$, ottenuta da A sostituendo alla sua i-esima colonna $\mathbf{a}_i$ il vettore $\mathbf{a}_i + \delta\mathbf{a}$, è non singolare se $\mathbf{e}_i^T A^{-1}\delta\mathbf{a} \neq -1$;

c) la matrice $\widetilde{A}$, ottenuta da A sostituendo al suo elemento a_{ij} l'elemento $a_{ij} + \alpha$, è non singolare se $\alpha b_{ji} \neq -1$, dove $B = A^{-1}$.

(Risposta: b) si veda l'esercizio 1.45.)

4.12 Si consideri il sistema lineare $A\mathbf{x} = \mathbf{b}$, $\mathbf{b} \neq \mathbf{0}$. Si supponga di dare al vettore $\mathbf{b}$ la perturbazione $\delta\mathbf{b}$, e sia $\mathbf{x} + \delta\mathbf{x}$ la soluzione del sistema così ottenuto:

$$A(\mathbf{x} + \delta\mathbf{x}) = \mathbf{b} + \delta\mathbf{b}.$$

a) Si dimostri che

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \mu(A)\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|};$$

b) si costruisca una matrice A e dei vettori $\mathbf{b}$ e $\delta\mathbf{b}$ tali che la relazione precedente valga con il segno di uguaglianza per la norma 2;

c) sono dati i due sistemi lineari $A\mathbf{x} = \mathbf{b}$ e $B\mathbf{y} = \mathbf{b}$, dove

$$A = \begin{bmatrix} 151 & 30.1 \\ 30.1 & 6 \end{bmatrix}, \quad B = \begin{bmatrix} 151 & 30.1 \\ 30.1 & -6 \end{bmatrix},$$

si dica per quale dei due sistemi la soluzione è più sensibile agli errori del termine noto. Se in entrambi i casi si vuole ottenere la soluzione con la stessa precisione rispetto alla norma 2, si dica quanto più preciso deve essere $\mathbf{b}$ nel caso peggiore.

(Traccia: b) si consideri una matrice A normale con autovalore di modulo massimo $\lambda_{\max}$ e minimo $\lambda_{\min}$ e i vettori $\mathbf{b}$ e $\delta\mathbf{b}$, autovettori di A corrispondenti a $\lambda_{\max}$ e $\lambda_{\min}$; c) si calcoli $\mu_2(A)/\mu_2(B)$.)

4.13 Sia $\mathbf{x}^*$ la soluzione del sistema lineare $A\mathbf{x} = \mathbf{b}$ e $\widetilde{\mathbf{x}}$ la soluzione calcolata in aritmetica finita.

a) Si dimostri che

$$\frac{\|\mathbf{x}^* - \widetilde{\mathbf{x}}\|}{\|\mathbf{x}^*\|} \leq \mu(A)\ \frac{\|\mathbf{r}\|}{\|\mathbf{b}\|},$$

dove $\mathbf{r} = A\widetilde{\mathbf{x}} - \mathbf{b}$ è il *residuo* di $\widetilde{\mathbf{x}}$; ne segue che la stima "a posteriori" dell'errore di una soluzione calcolata basata solo sulla grandezza del residuo può essere priva di significato;

b) siano

$$A = \begin{bmatrix} 1 & 1 \\ \text{-}1 & \text{-}1 + \epsilon \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 1 \\ \text{-}1 \end{bmatrix},$$

e si supponga che la soluzione calcolata sia

$$\widetilde{\mathbf{x}} = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}.$$

Si confrontino l'errore relativo di $\widetilde{\mathbf{x}}$ e il suo residuo;

c) siano

$$A = \begin{bmatrix} 0.81 & \text{-}0.56 \\ 0.16 & \text{-}0.11 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 0.25 \\ 0.05 \end{bmatrix};$$

la soluzione esatta è $\mathbf{x}^* = [1, 1]^T$ e sono date due soluzioni approssimate $\widetilde{\mathbf{x}}_1 = [1.1, 0.9]^T$ e $\widetilde{\mathbf{x}}_2 = [0.5, 0.31]^T$. Si calcolino gli errori relativi ϵ_1 e ϵ_2 delle due soluzioni e i corrispondenti residui;

d) siano $A\mathbf{x} = \mathbf{b}$ e $A\mathbf{y} = \mathbf{c}$ due sistemi lineari, in cui

$$A = \begin{bmatrix} 1 & 1.001 \\ 1 & 1 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 2.001 \\ 2 \end{bmatrix}, \quad \mathbf{c} = \begin{bmatrix} 1 \\ 0 \end{bmatrix},$$

e si supponga che un algoritmo fornisca le seguenti soluzioni approssimate

$$\widetilde{\mathbf{x}} = \begin{bmatrix} 2 \\ 0 \end{bmatrix}, \qquad \widetilde{\mathbf{y}} = \begin{bmatrix} \text{-}1001 \\ 1000 \end{bmatrix};$$

si calcolino gli errori relativi ϵ_x e ϵ_y delle due soluzioni e i corrispondenti residui.

Si osservi, come risulta dai punti c) e d), che ad una soluzione più precisa può corrispondere un residuo maggiore.

(Risposta: c) in norma ∞ risulta $\epsilon_1 = 0.1$, $\epsilon_2 = 0.69$, $\|\mathbf{r}_1\| = 0.137$, $\|\mathbf{r}_2\| = 0.0186$; d) è $\mathbf{x}^* = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, $\mathbf{y}^* = \begin{bmatrix} \text{-}1000 \\ 1000 \end{bmatrix}$, per cui $\epsilon_x = 1$, $\epsilon_y = 0.001$, $\|\mathbf{r}_x\| = 0.001$, $\|\mathbf{r}_y\| = 1$. Sia in c) che in d) la matrice A è malcondizionata: in c) è $\mu(A) \approx 2.7\ 10^3$, in d) è $\mu(A) \approx 4\ 10^3$.)

4.14 Sia $T \in \mathbf{R}^{n\times n}$ triangolare superiore.

a) Sia $\mathbf{d}$ un vettore ad n componenti tutte uguali a 1 o -1. Si dia un algoritmo per determinare i segni delle componenti di $\mathbf{d}$ in modo che il vettore $\mathbf{y} = T^{-1}\mathbf{d}$ abbia componenti di modulo grande;

b) il vettore $\mathbf{y}$ così ottenuto è usato per stimare la $\|T^{-1}\|_\infty$, infatti

$$\|T^{-1}\|_\infty \geq \frac{\|\mathbf{y}\|_\infty}{\|\mathbf{d}\|_\infty} = \|\mathbf{y}\|_\infty.$$

Si applichi tale limitazione al caso della matrice T di elementi

$$t_{ij} = \begin{cases} 1 & \text{se } i = j, \\ -1 & \text{se } i < j, \\ 0 & \text{se } i > j. \end{cases}$$

c) Siano L e U le matrici della fattorizzazione LU di una matrice $A \in \mathbf{C}^{n\times n}$. Si dimostri che

$$\mu_\infty(A) \geq \frac{\mu_\infty(U)}{\mu_\infty(L)};$$

d) si sfrutti la tecnica descritta in a) per determinare una stima di $\mu_\infty(A)$ quando sia stata calcolata la fattorizzazione LU di A, nell'ipotesi che la matrice L sia ben condizionata, e si dica qual è il costo computazionale di questa stima.

(Traccia: a) si pone $d_n = 1$ e si assegna a d_i il segno opposto a quello di

$$\sum_{j=i+1}^{n} t_{ij} y_j, \quad \text{per } j = n-1, \ldots, 1;$$

b) posto $\mathbf{d} = [1, 1 \ldots, 1]^T$, risulta $\mathbf{y} = [2^{n-1}, 2^{n-2} \ldots, 2, 1]^T$ e quindi

$$\|\mathbf{y}\|_\infty = 2^{n-1} \leq \|T^{-1}\|_\infty;$$

c) poiché $U = L^{-1}A$, è $\|U\| \leq \|L^{-1}\| \ \|A\|$ e $\|U^{-1}\| \leq \|L\| \ \|A^{-1}\|$; d) si può applicare alla matrice U la tecnica descritta al punto a) e se $\mu_\infty(L)$ è piccolo, si può stimare $\mu_\infty(A) \approx \|A\|_\infty \|\mathbf{y}\|_\infty$. Il costo computazionale è $n^2/2$.)

4.15 Si determini una matrice di permutazione Π tale che la matrice

$$A = \begin{bmatrix} 1 & 1 & 0 & 3 \\ 2 & 1 & -1 & 1 \\ -1 & 2 & 3 & -1 \\ 3 & -1 & -1 & 2 \end{bmatrix}$$

ammetta fattorizzazione $\Pi A = LU$.

4.16 Si verifichi con il metodo di Gauss che il sistema

$$\begin{bmatrix} 3 & 1 & 0 \\ 2 & -1 & -1 \\ 4 & 3 & 1 \end{bmatrix} \quad \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \quad = \quad \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

è consistente e se ne calcoli una soluzione.

4.17 Si risolvano i seguenti sistemi lineari con il metodo di Gauss, utilizzando un'aritmetica in base 10 e 4 cifre significative, con arrotondamento dei risultati intermedi e si confrontino i risultati ottenuti con le soluzioni esatte $\mathbf{x}^*$ indicate. Si ripeta il calcolo con la variante del massimo pivot parziale e totale.

(1) $$\begin{bmatrix} 0.02 & 0.8 \\ 0.3 & 12.5 \end{bmatrix} \quad \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad = \quad \begin{bmatrix} 0.806 \\ 12.59 \end{bmatrix}, \quad \mathbf{x}^* = \begin{bmatrix} 0.3 \\ 1.0 \end{bmatrix}$$

(2) $$\begin{bmatrix} 0.02 & 0.63 \\ 0.15 & 2.35 \end{bmatrix} \quad \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad = \quad \begin{bmatrix} 0.797 \\ 2.89 \end{bmatrix}, \quad \mathbf{x}^* = \begin{bmatrix} -1.1 \\ 1.3 \end{bmatrix}$$

(3) $$\begin{bmatrix} 1 & 10^5 \\ 1 & 1 \end{bmatrix} \quad \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad = \quad \begin{bmatrix} 10^5 \\ 2 \end{bmatrix}, \quad \mathbf{x}^* = \begin{bmatrix} 1.\overline{00001} \\ 0.\overline{99998} \end{bmatrix}$$

$$(4)\qquad \begin{bmatrix} 0.5 & -1.04 & 1.475 \\ -2 & 4.1 & -4 \\ 1 & 2.7 & -5.2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 3.085 \\ -9.65 \\ -7.63 \end{bmatrix}, \quad \mathbf{x}^* = \begin{bmatrix} 1 \\ -0.5 \\ 1.4 \end{bmatrix}$$

$$(5)\qquad \begin{bmatrix} 0.1 & 1 & 3 \\ 0.5 & 1.5 & 1 \\ -4 & 1 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -2.05 \\ 0.25 \\ -4.2 \end{bmatrix}, \quad \mathbf{x}^* = \begin{bmatrix} 1.5 \\ 0.2 \\ -0.8 \end{bmatrix}$$

4.18 Sia $A \in \mathbf{C}^{n\times n}$. Se esiste la fattorizzazione LU di A, si dica quale struttura hanno le matrici L e U quando

a) A è tridiagonale;

b) A è una matrice a banda di ampiezza $k < n$ (per la definizione di matrice a banda si veda l'esercizio 1.25);

c) gli elementi di A verificano la relazione $a_{ij} = 0$ per $i > j+1$ (la matrice A è detta *in forma di Hessenberg superiore*).

(Risposta: a) L e U bidiagonali, b) L e U a banda inferiore e superiore di ampiezza k, c) L bidiagonale, U triangolare superiore.)

4.19 Si determini la fattorizzazione LU della matrice $A \in \mathbf{R}^{n\times n}$ (in forma di Hessenberg superiore, si veda l'esercizio 4.18)

$$A = \begin{bmatrix} n & n-1 & n-2 & \dots & 1 \\ n-1 & n-1 & n-2 & \dots & 1 \\ & n-2 & n-2 & \dots & 1 \\ & & \ddots & \ddots & \vdots \\ & & & 1 & 1 \end{bmatrix}$$

(Risposta:

$$L = \begin{bmatrix} 1 & & & & \\ \frac{n-1}{n} & 1 & & & \\ & \frac{(n-2)n}{n-1} & 1 & & \\ & & \ddots & \ddots & \\ & & & \frac{3}{2} & 1 \end{bmatrix}$$

$$U = \begin{bmatrix} n & n-1 & n-2 & \dots & 1 \\ & \dfrac{n-1}{n} & \dfrac{n-2}{n} & \dots & \dfrac{1}{n} \\ & & \dfrac{n-2}{n-1} & \dots & \dfrac{1}{n-1} \\ & & & \ddots & \vdots \\ & & & & \dfrac{1}{2} \end{bmatrix} .)$$

4.20 Sia $A \in \mathbf{C}^{n\times n}$ una matrice hermitiana e siano $A^{(k)}$ le matrici ottenute applicando ad A il metodo di Gauss e $B^{(k)}$ le sottomatrici di $A^{(k)}$ ottenute cancellando le prime $k-1$ righe e colonne.

a) Nell'ipotesi che il metodo possa essere applicato senza scambi di righe, si dimostri che le $B^{(k)}$ sono hermitiane;

b) se A è anche definita positiva, si dimostri che le $B^{(k)}$ sono definite positive, e quindi il metodo di Gauss è sempre applicabile senza scambi di righe;

c) se A è anche a predominanza diagonale in senso stretto, si dimostri che le $B^{(k)}$ sono a predominanza diagonale in senso stretto e che il metodo di Gauss, applicato con la variante del massimo pivot parziale, non richiede scambi di righe.

(Traccia: Si dimostri per induzione; si ha

$$B^{(k)} = \begin{bmatrix} \alpha & \mathbf{b}^H \\ \mathbf{b} & C^{(k)} \end{bmatrix} \quad \text{e} \quad B^{(k+1)} = C^{(k)} - \frac{1}{\alpha}\,\mathbf{b}\mathbf{b}^H,$$

a) essendo $C^{(k)}$ hermitiana per ipotesi induttiva, $B^{(k+1)}$ risulta hermitiana; b) si veda l'esercizio 1.43 f), notando che $B^{(k+1)}$ è il complemento di Schur di α in $B^{(k)}$, oppure, direttamente, $B^{(k)}$, essendo definita positiva per ipotesi induttiva, ammette la fattorizzazione LL^H dove

$$L = \begin{bmatrix} \beta & \mathbf{0}^H \\ \mathbf{c} & N \end{bmatrix},$$

con $\beta > 0$ e N triangolare inferiore non singolare, da cui

$$B^{(k)} = \begin{bmatrix} \beta^2 & \beta\mathbf{c}^H \\ \beta\mathbf{c} & \mathbf{c}\mathbf{c}^H + NN^H \end{bmatrix}.$$

Ne segue che $\alpha = \beta^2$, $\mathbf{b} = \beta\mathbf{c}$, $C^{(k)} = \mathbf{c}\mathbf{c}^H + NN^H = \dfrac{1}{\alpha}\,\mathbf{b}\mathbf{b}^H + NN^H$ e $B^{(k+1)} = NN^H$, per cui $B^{(k+1)}$ è definita positiva; c) indicati con b_j e c_{ij} gli elementi di $\mathbf{b}$ e $C^{(k)}$, per ipotesi induttiva è

$$|\alpha| > \sum_{j=1}^{n-k} |b_j|$$

$$|c_{ii}| > \sum_{\substack{j=1\\ j\neq i}}^{n-k} |c_{ij}| + |b_i| > \sum_{\substack{j=1\\ j\neq i}}^{n-k} |c_{ij}| + \frac{|b_i|}{|\alpha|} \sum_{j=1}^{n-k} |b_j| \geq \frac{|b_i|^2}{|\alpha|} + \sum_{\substack{j=1\\ j\neq i}}^{n-k} |c_{ij} - \frac{1}{\alpha}\, b_i \bar{b}_j|$$

da cui

$$|c_{ii} - \frac{1}{\alpha}\,|b_i|^2| \geq |c_{ii}| - \frac{1}{|\alpha|}|b_i|^2 > \sum_{\substack{j=1\\ j\neq i}}^{n-k} |c_{ij} - \frac{1}{\alpha}\, b_i \bar{b}_j|. \ )$$

4.21 Sia $A \in \mathbf{C}^{n\times n}$ e si indichino con a_M e $a_M^{(k)}$ rispettivamente il massimo modulo degli elementi di A e di $A^{(k)}$; si dimostri che

a) se A è hermitiana e a predominanza diagonale in senso stretto, allora $a_M^{(k)} < 2a_M$ per $k = 2, \ldots, n$;

b) se A è definita positiva, allora $a_M^{(k)} \geq a_M^{(k+1)}$ per $k = 1, \ldots, n-1$ e quindi $a_M^{(n)} \leq a_M$;

mentre usando la tecnica del massimo pivot parziale

c) se A è tridiagonale, allora $a_M^{(k)} \leq 2a_M$ per $k = 2, \ldots, n$;

d) se A è in forma di Hessenberg superiore (per la definizione si veda l'esercizio 4.18), allora $a_M^{(k)} \leq k a_M$ per $k = 2, \ldots, n$.

(Traccia: poiché anche le $A^{(k)}$ hanno predominanza diagonale in senso stretto (si veda l'esercizio 4.20), per ogni $s \geq 1$ e i tale che $s+1 \leq i \leq n$ è $|m_{is}| \sum_{j=s+1}^{i} \left|a_{sj}^{(s)}\right| < \left|a_{is}^{(s)}\right|$, e dalla (22) si ha

$$\sum_{j=s+1}^{i} \left|a_{ij}^{(s+1)}\right| \leq \sum_{j=s+1}^{i} \left|a_{ij}^{(s)}\right| + |m_{is}| \sum_{j=s+1}^{i} \left|a_{sj}^{(s)}\right| < \sum_{j=s}^{i} \left|a_{ij}^{(s)}\right|.$$

Si dimostri per induzione su r che

$$|a_{kk}^{(k)}| < \sum_{j=k-r}^{k} \left|a_{kj}^{(k-r)}\right|, \quad \text{per } r = 1, \ldots, k-1.$$

Per $r = 1$ la relazione discende dalla (22) e dal fatto che $a^{(k)}_{k,k+1} = a^{(k)}_{k+1,k}$ perché A è hermitiana, per $r > 1$ si ha $\sum_{j=k-r+1}^{k} \left|a^{(k-r+1)}_{kj}\right| < \sum_{j=k-r}^{k} \left|a^{(k-r)}_{kj}\right|$ e si sfrutti l'ipotesi induttiva per $r - 1$. Ne segue che

$$|a^{(k)}_{kk}| < \sum_{i=1}^{k} |a_{ki}| < 2|a_{kk}|.$$

b) $\quad 0 < a^{(k+1)}_{k+1,k+1} = a^{(k)}_{k+1,k+1} - \dfrac{[a^{(k+1)}_{k,k+1}]^2}{a^{(k)}_{kk}} \leq a^{(k)}_{k+1,k+1}.$)

4.22 Sia $A \in \mathbf{C}^{n\times n}$. Si dica qual è il numero di operazioni moltiplicative richieste, quando non si facciano scambi di righe, dal metodo di Gauss per la fattorizzazione LU di

a) matrici tridiagonali;

b) matrici a banda di ampiezza $k \ll n$ (per la definizione di matrice a banda si veda l'esercizio 1.25);

c) matrici in forma di Hessenberg superiore (per la definizione si veda l'esercizio 4.18);

d) matrici hermitiane.

(Risposta: a) $2n$, b) $n(k + k^2)$, c) $n^2/2$, d) $n^3/6$, si veda l'esercizio 4.20.)

4.23 Si dica quante operazioni moltiplicative sono richieste per calcolare l'inversa di una matrice $A \in \mathbf{C}^{n\times n}$ tridiagonale con il metodo di Gauss, purché non si facciano scambi di righe. E se A è anche hermitiana?

(Risposta: $5n^2/2$, se A è hermitiana $3n^2/2$.)

4.24 Sia $A \in \mathbf{C}^{n\times n}$, n pari, una matrice i cui elementi a_{ij} sono non nulli se e solo se $|i - j| = 1$, e $\mathbf{b} \in \mathbf{C}^n$.

a) Si determini un algoritmo per risolvere il sistema $A\mathbf{x} = \mathbf{b}$ con $2n$ operazioni moltiplicative;

b) indicata con $\widetilde{\mathbf{x}}$ la soluzione effettivamente calcolata con questo algoritmo, si verifichi che

$$(A + \Delta A)\widetilde{\mathbf{x}} = \mathbf{b}, \quad \text{dove } |\Delta A| \leq 2u|A| + O(u^2),$$

in cui u è la precisione di macchina.

(Traccia: a) si calcoli x_2 dalla prima equazione, si sostituisca nella terza, e così via fino a x_n; per le componenti di indice dispari si proceda all'inverso; b) con questo algoritmo si ha

$$\widetilde{x}_2 = fl\left(\frac{b_1}{a_{12}}\right) = \frac{b_1}{a_{12}+\delta_{12}}, \quad |\delta_{12}| \le u|a_{12}|,$$

$$\widetilde{x}_4 = fl\left(\frac{b_3 - a_{32}\widetilde{x}_2}{a_{34}}\right) = \frac{b_3 - (a_{32}+\delta_{32})\widetilde{x}_2}{a_{34}+\delta_{34}},$$

$$|\delta_{32}| \le u|a_{32}|, \ |\delta_{34}| \le 2u|a_{34}|,$$

e così via.)

4.25 Sia $A \in \mathbf{R}^{n\times n}$ la matrice ad albero simmetrica (per la definizione di matrice ad albero si veda l'esercizio 1.59)

$$A = \begin{bmatrix} \alpha_1 & \alpha_2 & \alpha_3 & \dots & \alpha_n \\ \alpha_2 & \beta_2 & & & \\ \alpha_3 & & \beta_3 & & \\ \vdots & & & \ddots & \\ \alpha_n & & & & \beta_n \end{bmatrix};$$

a) si dica a quali condizioni devono soddisfare gli α_j e i β_j affinché esista la fattorizzazione LU di A e si dica quante operazioni moltiplicative sono richieste dal calcolo della fattorizzazione con il metodo di Gauss;

b) si determini una matrice Π di permutazione tale che la matrice $B = \Pi A \Pi^T$ possa essere fattorizzata con un costo computazionale dell'ordine di n e si calcoli L e U tali che

$$\Pi A \Pi^T = LU;$$

c) si esamini il caso che $\alpha_1 = \alpha_2 = \ldots = \alpha_n = 1$ e $\beta_2 = \beta_3 = \ldots = \beta_n = -1$.

(Risposta: a) si usi l'esercizio 1.59 e il teorema 4.4, sono richieste $\dfrac{n^3}{6}$ operazioni moltiplicative; b) Π è ottenuta da I scambiando la prima e l'ultima colonna, costo computazionale $2n$.)

4.26 Siano A e $B \in \mathbf{C}^{n\times n}$ due matrici non singolari che differiscono solo per una colonna e si supponga di conoscere A^{-1}. Si descriva un metodo per calcolare B^{-1} che sfrutti A^{-1} con un costo computazionale di $2n^2$.

(Traccia: siano $\mathbf{a}_i$ e $\mathbf{b}_i$ le colonne di A e di B diverse. Posto $\mathbf{u} = \mathbf{b}_i - \mathbf{a}_i$, si ha per la formula di Sherman-Morrison (si veda l'esercizio 1.45)

$$B^{-1} = (A + \mathbf{u}\mathbf{e}_i^T)^{-1} = A^{-1} - \frac{A^{-1}\mathbf{u}\mathbf{e}_i^T A^{-1}}{1 + \mathbf{e}_i^T A^{-1}\mathbf{u}}\,. \ )$$

4.27 a) Posto $\alpha_r = \dfrac{1}{r}$, per $r = 1, 2, \ldots$, si dimostri la relazione

$$\alpha_r = \sum_{j=r+1}^{n-1} \frac{1}{j(j-1)} \; + \; \frac{1}{n-1} \; , \quad \text{per } r = 1, \ldots, n-2, \; n = 3, 4, \ldots$$

b) Si applichi ad una matrice $A \in \mathbf{C}^{n\times n}$ il metodo di Gauss con la variante del massimo pivot totale. Si dimostri che, detto $a^{(k)}_{\max}$ il massimo modulo degli elementi della matrice $A^{(k)}$ ottenuta al $(k-1)$-esimo passo del metodo, si ha

$$a^{(k)}_{\max} \le g(n) a^{(1)}_{\max}, \quad \text{dove } g(n) = \sqrt{n \prod_{j=2}^{n} j^{1/(j-1)}}.$$

(Traccia: a) si proceda per induzione su r e su n; b) sia $B^{(k)}$ la sottomatrice ottenuta cancellando in $A^{(k)}$ le prime $k-1$ righe e colonne. Quindi $B^{(k)}$ ha ordine $r = n-k+1$ e risulta

$$|\det B^{(k)}| = p_k p_{k+1} \ldots p_n,$$

dove $p_k = \max\limits_{i,j=k,\ldots,n} |a^{(k)}_{ij}|$ è il modulo del pivot al k-esimo passo. Per la disuguaglianza di Hadamard (si veda l'esercizio 2.28) è

$$|\det B^{(k)}| \le \left(\sqrt{r} p_k\right)^r, \quad r = n-k+1, \; k = 1, \ldots, n,$$

e quindi $\quad p_k p_{k+1} \ldots p_n \le r^{r/2} p_k^r, \quad$ da cui

$$p_{k+1} \ldots p_n \le r^{r/2} p_k^{r-1}, \quad r = n-k+1, \; k = 1, \ldots, n-1.$$

Per $k = 1$ si ottiene

$$q_1 = (p_2 \ldots p_n)^{1/(n-1)} \le p_1 \sqrt{n} \; \sqrt{n^{1/(n-1)}},$$

e per $k = 2, \ldots, n-1$, si ottiene

$$q_k = (p_{k+1} \ldots p_n)^{1/[r(r-1)]} \le p_k^{1/r} \sqrt{r^{1/(r-1)}}.$$

Moltiplicando fra loro queste relazioni si ha

$$q_1 \prod_{k=2}^{n-1} q_k = \prod_{r=1}^{n-1} p_k^{\alpha_r} \le p_1 \prod_{r=2}^{n-1} p_k^{1/r} g(n),$$

dove α_r è la quantità definita al punto a), e quindi

$$p_n \leq p_1 g(n). \ )$$

4.28 Siano $A \in \mathbf{C}^{n\times n}$ e $\mathbf{b} \in \mathbf{C}^n$. Il sistema lineare $A\mathbf{x} = \mathbf{b}$ può essere risolto utilizzando due diversi algoritmi:

a) si opera in aritmetica complessa, memorizzando A e $\mathbf{b}$ in $2n^2 + 2n$ locazioni di memoria (ogni numero complesso richiede 2 locazioni),

b) posto $A = A_1 + \mathbf{i}A_2,\ \mathbf{b} = \mathbf{b}_1 + \mathbf{i}\mathbf{b}_2$, con $A_1, A_2 \in \mathbf{R}^{n\times n},\ \mathbf{b}_1, \mathbf{b}_2 \in \mathbf{R}^n$, si risolve il sistema

$$\begin{bmatrix} A_1 & -A_2 \\ A_2 & A_1 \end{bmatrix} \quad \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} \quad = \quad \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix},$$

ottenendo $\mathbf{x} = \mathbf{x}_1 + \mathbf{i}\mathbf{x}_2$ (questo sistema richiede $4n^2 + 2n$ locazioni di memoria).

Si confrontino i due metodi sulla base della complessità computazionale e del condizionamento in norma 2.

(Risposta: l'algoritmo a) richiede un numero di operazioni moltiplicative fra numeri complessi dell'ordine di $n^3/3$, e poiché il prodotto di due numeri complessi può essere calcolato con 4 moltiplicazioni fra numeri reali, la complessità è dell'ordine di $4n^3/3$ (esiste anche un algoritmo per calcolare il prodotto di due numeri complessi con 3 moltiplicazioni e 5 addizioni); l'algoritmo b) richiede un numero di operazioni moltiplicative reali dell'ordine di $(2n)^3/3 = 8n^3/3$. Il secondo algoritmo quindi, oltre a richiedere il doppio di memoria, ha anche una complessità che è il doppio di quella del primo. Posto $B = \begin{bmatrix} A_1 & -A_2 \\ A_2 & A_1 \end{bmatrix}$, si verifichi che $B^H B = \begin{bmatrix} S & -T \\ T & S \end{bmatrix}$, dove $S + \mathbf{i}T = A^H A$, e si applichi l'esercizio 2.26 per dimostrare che $\mu_2(A) = \mu_2(B)$.)

4.29 Sia $A \in \mathbf{C}^{n\times n}$.

a) Sotto le ipotesi del teorema 4.4 la matrice A è fattorizzabile nel prodotto

$$A = LDR,$$

in cui L ed R sono matrici triangolari rispettivamente inferiore e superiore con gli elementi principali uguali a 1 e D è una matrice diagonale;

b) si descriva un algoritmo per il calcolo della fattorizzazione LDR;

c) si dimostri che se A è non singolare anche D è non singolare;

d) si dimostri che se A non verifica le ipotesi del teorema 4.4, l'algoritmo è applicabile con la variante del massimo pivot parziale, cioè esiste una matrice di permutazione Π tale che

$$\Pi A = LDR,$$

e se A è di rango $m < n$ e si applica l'algoritmo con la variante del massimo pivot totale, allora

$$\Pi A \Pi' = \begin{bmatrix} L_{11} & O \\ L_{21} & I_{n-m} \end{bmatrix} \begin{bmatrix} D_1 & O \\ O & O \end{bmatrix} \begin{bmatrix} R_{11} & R_{12} \\ O & I_{n-m} \end{bmatrix},$$

in cui $L_{11}, D_1, R_{11} \in \mathbf{C}^{m\times m}$, L_{11} e R_{11} sono triangolari, rispettivamente inferiore e superiore, con gli elementi principali uguali a 1 e D_1 è una matrice diagonale e non singolare;

e) si considerino in particolare le matrici $A = \mathbf{x}\mathbf{y}^H$ e $B = \mathbf{x}\mathbf{y}^H + \mathbf{u}\mathbf{v}^H$, dove $\mathbf{x}, \mathbf{y}, \mathbf{u}, \mathbf{v} \in \mathbf{C}^n$;

f) si applichi l'algoritmo descritto al punto b) alla matrice

$$A = \begin{bmatrix} 6 & -4 & 8 & 2 & -10 \\ -9 & 6 & -12 & -3 & 15 \\ 15 & -10 & 20 & 5 & -25 \\ -12 & 8 & -16 & -4 & 20 \\ 3 & -2 & 4 & 1 & -5 \end{bmatrix}.$$

La fattorizzazione LDR può essere ottenuta senza effettuare scambi di righe. È unica?

(Traccia: a) posto $A = LU$, si verifichi che esiste D diagonale tale che $U = DR$; b) si apportino le necessarie modifiche al metodo di Gauss; d) si segua un ragionamento analogo a quello della dimostrazione del teorema 4.5; f) la fattorizzazione è $A =$

$$\begin{bmatrix} 1 & & & & \\ -3/2 & 1 & & & \\ 5/2 & & 1 & & \\ -2 & & & 1 & \\ 1/2 & & & & 1 \end{bmatrix} \begin{bmatrix} 6 & & & & \\ & 0 & & & \\ & & 0 & & \\ & & & 0 & \\ & & & & 0 \end{bmatrix} \begin{bmatrix} 1 & -2/3 & 4/3 & 1/3 & -5/3 \\ & 1 & & & \\ & & 1 & & \\ & & & 1 & \\ & & & & 1 \end{bmatrix}$$

e non è unica.)

4.30 Si dica per quali valori dei parametri le seguenti matrici sono definite positive e se ne dia la fattorizzazione LL^H:

(1) $$A = \begin{bmatrix} 2 & 1+\mathbf{i}\alpha & 0 \\ 1-\mathbf{i}\alpha & 2 & \alpha \\ 0 & \alpha & 2 \end{bmatrix}, \quad \alpha \in \mathbf{R},$$

(2) $$A = \begin{bmatrix} \alpha & \beta & 0 \\ -\beta & \alpha & \beta \\ 0 & -\beta & \alpha \end{bmatrix}, \quad \alpha, \beta \in \mathbf{C}.$$

(Risposta: (1) $|\alpha| < \sqrt{\dfrac{3}{2}}$, $L = \sqrt{2} \begin{bmatrix} 1 & 0 & 0 \\ \dfrac{1-\mathbf{i}\alpha}{2} & \dfrac{\sqrt{3-\alpha^2}}{2} & 0 \\ 0 & \dfrac{\alpha}{\sqrt{3-\alpha^2}} & \dfrac{\sqrt{3-2\alpha^2}}{\sqrt{3-\alpha^2}} \end{bmatrix}$;

(2) $\alpha \in \mathbf{R}$, $\beta = \mathbf{i}b$, $b \in \mathbf{R}$, $\alpha > \sqrt{2}\,|b|$,

$$L = \sqrt{\alpha} \begin{bmatrix} 1 & 0 & 0 \\ -\dfrac{\beta}{\alpha} & \dfrac{\sqrt{\alpha^2-b^2}}{\alpha} & 0 \\ 0 & -\dfrac{\beta}{\sqrt{\alpha^2-b^2}} & \dfrac{\sqrt{\alpha^2-2b^2}}{\sqrt{\alpha^2-b^2}} \end{bmatrix}. \ )$$

4.31 Sia $A \in \mathbf{C}^{n\times n}$ una matrice definita positiva a banda di ampiezza k (per la definizione di matrice a banda si veda l'esercizio 1.25). Si dimostri che la matrice L della fattorizzazione LL^Hè a banda inferiore di ampiezza k, e si dia il costo computazionale del calcolo di L con il metodo di Cholesky per $k \ll n$.

(Risposta: $n(k^2+3k)/2$ operazioni moltiplicative e n estrazioni di radice quadrata.)

4.32 Sia $A \in \mathbf{C}^{n\times n}$ una matrice definita positiva. Si dimostri che esiste ed è unica la fattorizzazione LDL^H di A, in cui L è una matrice triangolare inferiore con gli elementi principali uguali a 1 e D è una matrice diagonale, con gli elementi principali positivi. Si descriva un algoritmo per il calcolo di questa fattorizzazione e se ne valuti il costo computazionale.

4.33 Sia $A \in \mathbf{C}^{(n+m)\times(n+m)}$ una matrice definita positiva, con

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{12}^H & A_{22} \end{bmatrix},$$

dove $A_{11} \in \mathbf{C}^{n\times n},\ A_{12} \in \mathbf{C}^{n\times m},\ A_{22} \in \mathbf{C}^{m\times m}$, e sia $L \in \mathbf{C}^{(n+m)\times(n+m)}$ la matrice della fattorizzazione LL^H di A. Posto

$$L = \begin{bmatrix} L_{11} & O \\ \\ L_{12} & L_{22} \end{bmatrix},$$

con $L_{11} \in \mathbf{C}^{n\times n},\ L_{12} \in \mathbf{C}^{m\times n},\ L_{22} \in \mathbf{C}^{m\times m}$. Si dimostri che

$$L_{22}L_{22}^H = S, \quad \text{dove } S = A_{22} - A_{12}^H A_{11}^{-1} A_{12}$$

è il complemento di Schur di A_{11} in A (si veda l'esercizio 1.43). (Traccia: si imponga la condizione $LL^H = A$.)

4.34 Per un generico vettore $v \in \mathbf{R}^n$ si definiscono i due vettori

$$\mathbf{v}_+ = \frac{1}{2}(\mathbf{v} + J\mathbf{v}), \quad \mathbf{v}_- = \frac{1}{2}(\mathbf{v} - J\mathbf{v}).$$

Si dimostri che se A è centrosimmetrica e se $A\mathbf{x} = \mathbf{b}$, allora

$$A\mathbf{x}_+ = \mathbf{b}_+, \quad \text{e} \quad A\mathbf{x}_- = \mathbf{b}_-$$

(per la definizione di matrice centrosimmetrica e della matrice J si veda l'esercizio 2.35).

4.35 Si dica a quale ipotesi deve verificare il vettore $\mathbf{v} \in \mathbf{R}^n$ affinché la matrice di Householder

$$I - \beta\mathbf{v}\mathbf{v}^T, \quad \beta \in \mathbf{R},$$

sia persimmetrica (per la definizione di matrice persimmetrica si veda l'esercizio 2.36).

4.36 Si determini la fattorizzazione QR delle seguenti matrici con il metodo di Householder e con il metodo di Givens

$$(1) \qquad A = \begin{bmatrix} 1 & 1 & 1 \\ 2 & -1 & -1 \\ 2 & -4 & 5 \end{bmatrix}, \qquad (2) \qquad B = \frac{1}{9}\begin{bmatrix} 4 & 1 & -8 \\ 7 & 4 & 4 \\ 4 & -8 & 1 \end{bmatrix}.$$

4.37 Sia $A \in \mathbf{C}^{n\times n}$.

a) Si dica in che cosa differiscono le diverse fattorizzazioni QR di una matrice A non singolare;

b) si dimostri che è unica la fattorizzazione QR di una matrice A non singolare se si impone l'ulteriore condizione che gli elementi diagonali di R siano reali e positivi;

c) si dimostri che se A è singolare, allora due diverse fattorizzazioni QR di A non differiscono necessariamente per una trasformazione di fase;

d) si determinino tutte le possibile fattorizzazioni QR delle matrici

$$(1) \qquad A = \begin{bmatrix} 8 & 0 & 3 \\ 4 & -2 & 1 \\ 1 & -1 & -1 \end{bmatrix}, \qquad (2) \qquad B = \begin{bmatrix} 8 & -16 & 24 \\ 4 & -8 & 12 \\ 1 & -2 & 3 \end{bmatrix}.$$

(Traccia: a) si tenga conto del fatto che una matrice unitaria e triangolare superiore è diagonale, e quindi è una matrice di fase; c) si consideri ad esempio il caso in cui

$$R = \begin{bmatrix} S & T \\ O & O \end{bmatrix},$$

dove $S \in \mathbf{C}^{k \times k}$, $k \leq n-2$, è triangolare superiore; d) posto

$$Z = \frac{1}{9} \begin{bmatrix} 8 & 4 & 1 \\ 4 & -7 & -4 \\ 1 & -4 & 8 \end{bmatrix},$$

si ha $A = QR$, dove

$$Q = ZU, \qquad R = U^H \begin{bmatrix} 9 & -1 & 3 \\ & 2 & 1 \\ & & -1 \end{bmatrix}, \qquad U = \begin{bmatrix} \mathrm{e}^{i\theta_1} & & \\ & \mathrm{e}^{i\theta_2} & \\ & & \mathrm{e}^{i\theta_3} \end{bmatrix},$$

in cui $\theta_1, \theta_2, \theta_3 \in \mathbf{R}$ e $B = Q'R'$, dove

$$Q' = ZV, \qquad R' = V^H \begin{bmatrix} 9 & -18 & 27 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad V = \begin{bmatrix} \mathrm{e}^{i\theta_1} & 0 \\ 0 & V' \end{bmatrix},$$

in cui $V' \in \mathbf{C}^{2 \times 2}$ è una qualunque matrice unitaria.)

4.38 Sia $A \in \mathbf{C}^{n \times n}$ e sia $A = QR$ la sua fattorizzazione QR. Si dica quale struttura ha la matrice R se la A è una matrice a banda di ampiezza $k < n$ (per la definizione di matrice a banda si veda l'esercizio 1.25) e quale è il costo computazionale del metodo di Householder e del metodo di Givens.

(Risposta: se $2k < n$, R è a banda superiore di ampiezza $2k$; se $k \ll n$, il costo computazionale è $4nk^2$ con il metodo di Householder, $8nk^2$ con il metodo di Givens.)

4.39 Sia $A\mathbf{x} = \mathbf{b}$ un sistema di ordine n e $D \in \mathbf{C}^{n\times n}$ una matrice diagonale non singolare. Se $\mu(DA)$ è minore di $\mu(A)$, il sistema $DA\mathbf{x} = D\mathbf{b}$ è meglio condizionato di quello di partenza. In tal caso è possibile che la soluzione di questo secondo sistema sia affetta da un errore minore. Questa operazione viene detta *scalatura* per righe del sistema.

a) Si calcoli $\mu_\infty(A)$ e $\mu_\infty(DA)$ nel caso

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & \alpha & \alpha^2 \\ 1 & \alpha^2 & \alpha^4 \end{bmatrix},\ \alpha \gg 1,\ \text{e } D = \begin{bmatrix} \frac{1}{3} & & \\ & \frac{1}{\alpha^2+\alpha+1} & \\ & & \frac{1}{\alpha^4+\alpha^2+1} \end{bmatrix}$$

b) Non sempre però ad una diminuzione del numero di condizionamento segue un'effettiva diminuzione dell'errore della soluzione calcolata. Si consideri il caso del sistema $A\mathbf{x} = \mathbf{b}$, con

$$\mathbf{b} = \left[\frac{1}{3}, \frac{1}{\alpha^2+\alpha+1}, \frac{1}{\alpha^4+\alpha^2+1}\right]^T,\ \text{per } \alpha = 10,$$

operando in base 10 con quattro cifre significative e arrotondamento dei risultati intermedi.

(Risposta: a)

$$\mu_\infty(A) = \frac{(\alpha^4+\alpha^2+1)(\alpha^2+1)}{(\alpha-1)^2} \quad \text{e} \quad \mu_\infty(DA) = \frac{2\alpha^3+3\alpha^2+1}{(\alpha-1)^2}.\ )$$

4.40 Sia $A \in \mathbf{C}^{(n\times m)\times(n\times m)}$ la matrice tridiagonale a blocchi

$$A = \begin{bmatrix} A_1 & C_1 & & \\ B_1 & A_2 & \ddots & \\ & \ddots & \ddots & C_{m-1} \\ & & B_{m-1} & A_m \end{bmatrix},$$

in cui i blocchi sono matrici di ordine n.

a) Si dica sotto quale ipotesi esiste la fattorizzazione LU a blocchi di A, cioè

$$A = LU = \begin{bmatrix} I & & & \\ L_1 & I & & \\ & \ddots & \ddots & \\ & & L_{m-1} & I \end{bmatrix} \begin{bmatrix} U_1 & C_1 & & \\ & U_2 & \ddots & \\ & & \ddots & C_{m-1} \\ & & & U_m \end{bmatrix},$$

in cui i blocchi sono matrici di ordine n;

b) si descriva un algoritmo per calcolare tale fattorizzazione e se ne dia il costo computazionale;

c) si esamini in particolare il caso in cui $B_i = C_i = I$ per $i = 1, \ldots, m-1$, e $A_i = T$ per $i = 1, \ldots, m$;

d) si dimostri che se la matrice A ha *predominanza diagonale a blocchi in senso stretto per colonne*, cioè se per $i = 1, \ldots, m$, A_i è non singolare e

$$|A_i^{-1}\|_1(\|B_i\|_1 + \|C_{i-1}\|_1) < 1, \quad \|C_0\|_1 = \|B_m\|_1 = 0,$$

allora la fattorizzazione LU a blocchi esiste.

(Traccia: a) da $A = LU$ si ottengono le relazioni

$$U_1 = A_1, \qquad L_iU_i = B_i, \qquad L_iC_i + U_{i+1} = A_{i+1}, \quad \text{per } i = 1, \ldots, m-1,$$

per cui la fattorizzazione esiste se U_i è non singolare per $i = 1, \ldots, m-1$, e quindi se le sottomatrici principali di testa di A di ordine $n,\ 2n, \ldots, (m-1)n$, sono non singolari; b)

$$U_1 = A_1, \qquad L_i = B_iU_i^{-1}, \qquad U_{i+1} = A_{i+1} - L_iC_i, \quad \text{per } i = 1, \ldots, m-1,$$

con un costo computazionale di $(m-1)(h_1 + h_2)$ operazioni, in cui h_1 e h_2 sono le operazioni moltiplicative richieste dalla risoluzione dei sistemi $L_iU_i = B_i$ e dalle moltiplicazioni di L_i e C_i; c) $L_i = U_i^{-1}$, $U_{i+1} = T - L_i$, con un costo computazionale di $(m-1)h_1$ operazioni; d) si dimostri per induzione che $\|U_i^{-1}\|_1\|B_i\|_1 < 1$, infatti

$$U_i = A_i(I - A_i^{-1}B_{i-1}U_{i-1}^{-1}C_{i-1})$$

e

$$\begin{aligned} \|A_i^{-1}B_{i-1}U_{i-1}^{-1}C_{i-1}\|_1 &\le (\|A_i^{-1}\|_1\ \|C_{i-1}\|_1)\ (\|U_{i-1}^{-1}\|_1\ \|B_{i-1}\|_1) \\ &< 1 - \|A_i^{-1}\|_1\ \|B_i\|_1 \le 1, \end{aligned}$$

e quindi per il teorema 3.13 esiste U_i^{-1} e

$$\|U_i^{-1}\|_1 \le \frac{\|A_i^{-1}\|_1}{1 - \|A_i^{-1}B_{i-1}U_{i-1}^{-1}C_{i-1}\|_1} < \frac{1}{\|B_i\|_1}\ .$$

Ne segue che $\|L_i\|_1 < 1$.)

4.41 Sia $A\mathbf{x} = \mathbf{b}$ un sistema lineare di ordine n^2 in cui A è la matrice a blocchi

$$A = \begin{bmatrix} T & I & & \\ I & T & \ddots & \\ & \ddots & \ddots & I \\ & & I & T \end{bmatrix}, \quad \text{con } T = \begin{bmatrix} 4 & -1 & & \\ -1 & 4 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 4 \end{bmatrix} \in \mathbf{R}^{n\times n}.$$

Si dimostri che la matrice A è definita positiva e si determini il costo computazionale del metodo di Gauss a elementi e a blocchi (si veda l'esercizio 4.40) e del metodo di Cholesky.

(Traccia: si applichi il teorema 2.41; la matrice A è simmetrica e a banda di ampiezza n, per cui, tenendo conto della simmetria, il metodo di Gauss ha costo computazionale $n^4/2$ (si veda l'esercizio 4.22) per la fattorizzazione LU; applicando il metodo di Gauss a blocchi, le matrici U_i sono simmetriche (piene) e quindi il costo computazionale è $n^4/2$; il metodo di Cholesky ha costo computazionale $n^4/2$ (si veda l'esercizio 4.31) per la fattorizzazione LL^T; in ogni caso il costo computazionale della risoluzione dei sistemi finali a matrice triangolare o triangolare a blocchi è dell'ordine di n^3.)

4.42 Per risolvere un sistema lineare $A\mathbf{y} = \mathbf{b}$, dove $A \in \mathbf{C}^{n\times n}$ è una matrice di Vandermonde (si veda l'esercizio 1.54) conviene procedere nel modo seguente

(1) si costruisce il vettore $\mathbf{c} \in \mathbf{C}^n$ con l'algoritmo:

per $i = 1, \ldots, n$
  $c_i = b_i$,
  se $i \neq 1$ per $j = 2, \ldots, i$, $\quad c_i = \dfrac{c_i - c_{j-1}}{x_i - x_{j-1}}$,

(2) si costruisce il vettore $\mathbf{y}$ con l'algoritmo:

$y_1 = c_n$
per $i = n-1, \ldots, 1$
  $y_{n-i+1} = y_{n-i}$,
  se $i \leq n-2$ per $j = n-i, \ldots, 2$, $\quad y_j = y_{j-1} - x_i y_j$,
  $y_1 = c_i - x_i y_1$.

Si dimostri che il vettore $\mathbf{y}$ così ottenuto è la soluzione del sistema dato e si valuti il costo computazionale di questo procedimento.

(Traccia: l'algoritmo descritto calcola i coefficienti del polinomio $p(x) = y_1 x^{n-1} + y_2 x^{n-2} + \ldots + y_n$ di interpolazione per i punti (x_i, b_i), per $i = 1, \ldots, n$, nella forma di Newton. Il vettore $\mathbf{c}$ costruito contiene gli elementi

diagonali della tabella delle differenze divise, per cui

$$p(x) = c_1 + c_2(x - x_1) + c_3(x - x_1)(x - x_2) \\ + \ldots + c_n(x - x_1)(x - x_2)\ldots(x - x_{n-1}).$$

I coefficienti di $p(x)$ vengono calcolati ricorsivamente tenendo conto che

$$p_n(x) = c_n$$
$$p_i(x) = (x - x_i)p_{i+1}(x) + c_i, \text{ per } i = n-1, \ldots, 1$$
$$p(x) = p_1(x).$$

Questo procedimento richiede n^2 operazioni moltiplicative.)

4.43 Sia p un numero primo; si consideri l'insieme $\mathbf{Z}_p = \{\ 0,\ 1,\ \ldots,\ p-1\ \}$ e su di esso si definisca *l'aritmetica intera modulo p*, nel modo seguente: siano $a,\ b,\ c \in \mathbf{Z}_p$ e sia *op* è un'operazione aritmetica, allora

(1) se $op \in \{+,\ -,\ *\}$, è $c = (a\ op\ b) \bmod p$, se esiste un intero k tale che $(a\ op\ b) - c = kp$,

(2) se *op* è l'operazione aritmetica $\div$, è $c = (a\ \div\ b) \bmod p$, se esiste un intero k tale che $a - cb = kp$.

L'insieme $\mathbf{Z}_p$ con questa aritmetica è un campo.

a) Si dimostri che se A è una matrice di ordine n ad elementi interi, tale che

$$\det A_k \neq 0 \bmod p, \quad k = 1, \ldots, n-1,$$

dove A_k sono le sottomatrici principali di testa di ordine k di A, allora esistono due matrici con elementi in $\mathbf{Z}_p$ L triangolare inferiore con elementi principali uguali a 1 e U triangolare superiore, tali che $A = LU \mod p$ e il metodo di Gauss è applicabile con l'aritmetica modulo p.

b) Se si applica il metodo di Gauss alla matrice A con l'aritmetica intera modulo p e con la tecnica del massimo pivot totale nel caso in cui risulti un pivot nullo, allora si ha che il rango di A è maggiore o uguale al numero σ_p degli elementi principali non nulli di U. Se

$$p > q = 2\ \left(\prod_{j=1}^{n}\sum_{i=1}^{n}|a_{ij}|^2\right)^{\frac{1}{2}},$$

allora il rango di A è uguale a σ_p.

c) Si calcoli una limitazione inferiore del rango della matrice

$$A = \begin{bmatrix} 0.58 & -1.1 & -0.52 \\ -0.56 & 1.12 & 0.56 \\ 0.02 & 0.02 & 0.04 \end{bmatrix}$$

dell'esempio 4.22.

(Traccia: a) si applichino i teoremi 4.4 e 4.12, che valgono su un campo qualsiasi e quindi anche su $\mathbf{Z}_p$; b) il metodo di Gauss, applicato su $\mathbf{Z}_p$, permette di calcolare il rango di A su $\mathbf{Z}_p$, cioè la massima dimensione delle sottomatrici B di A per cui $\det B \neq 0 \bmod p$. Inoltre poiché $\det B = 0 \bmod p$ se e solo se $\det B$ è multiplo intero di p, si ha che il rango di A è maggiore o uguale al rango di A su $\mathbf{Z}_p$, e l'uguaglianza vale se $p > 2 \det B$ per ogni sottomatrice B di A. Infatti se $p > 2 \det B$, allora $\det B = 0 \bmod p$ se e solo se $\det B = 0$. Per la disuguaglianza di Hadamard, posto $p > q$ vale $p > 2 \det B$ per ogni sottomatrice B di A; c) si moltiplichino per 100 gli elementi di A e si operi con $p = 5$, ottenendo

$$(100 \det A) \bmod 5 = \begin{bmatrix} 3 & 0 & 3 \\ 4 & 2 & 1 \\ 2 & 2 & 4 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ 4 & 1 & 1 \end{bmatrix} \begin{bmatrix} 3 & 0 & 3 \\ 0 & 2 & 2 \\ 0 & 0 & 0 \end{bmatrix} \bmod 5,$$

da cui segue che rango di $A \geq 2$.)

Commento bibliografico

"Quando si risolve un problema matematico con un calcolatore è inevitabile che si generino deviazioni della *soluzione* effettivamente calcolata dalla soluzione esatta. Quindi la soluzione ottenuta non serve a niente se non è accompagnata da una relativa stima dell'errore commesso". Questo in sostanza afferma Von Neumann nel primo dei due articoli scritti con Goldstine nel 1947 e nel 1951 [27, 28] sul calcolo dell'inversa di matrici, in un'epoca in cui si avvia l'utilizzazione dei calcolatori per risolvere problemi matematici, anche di dimensioni inimmaginabili solo pochi anni prima. Dai due articoli, che riportano la prima analisi sistematica degli errori generati dall'uso di un'aritmetica finita con numeri rappresentati in virgola fissa, risulta che il numero delle cifre esatte che si può sperare di ottenere per gli elementi della matrice inversa è assai minore del numero delle cifre utilizzate nei calcoli e decresce fortemente all'aumentare dell'ordine della matrice. Questi risultati così pessimistici hanno scoraggiato per un certo periodo l'uso dei metodi diretti per la risoluzione dei sistemi lineari, a favore dei metodi iterativi che apparivano più affidabili.

Però Wilkinson nel 1961 [29], utilizzando la tecnica dell'analisi all'indietro per l'errore generato da un'aritmetica finita in virgola mobile, riesce a valutare gli errori in modo più sintetico, mettendoli in relazione con le perturbazioni della matrice originale del problema. Da questa analisi risulta che anche i metodi diretti possono essere sufficientemente accurati.

Per lungo tempo si è cercato di individuare un parametro che misurasse la difficoltà di risoluzione di un sistema lineare, cioè il condizionamento. Inizialmente è stata suggerita come misura del condizionamento la grandezza del determinante, e quindi le matrici con determinante piccolo erano considerate mal condizionate. Il numero di condizionamento $\mu(A) = \|A\|\|A^{-1}\|$ è stato introdotto da Turing nel 1948 [26] in norma di Frobenius. Von Neumann in [27] aveva proposto per le matrici definite positive il numero

$$\frac{\max\limits_i \lambda_i}{\min\limits_i \lambda_i}.$$

Molti dei metodi diretti usati per risolvere sistemi lineari sono stati descritti indipendentemente da vari autori. Il metodo di eliminazione di Gauss, sviluppato all'inizio del 19^o secolo, e quindi con più di un secolo di sperimentazione numerica, è il metodo più usato per risolvere sistemi lineari con matrici dense. Il metodo, oggi conosciuto come metodo di Gauss-Jordan, è stato descritto per la prima volta da Clasen nel 1888. Mentre il nome di Gauss è appropriato in quanto si tratta di una modifica del metodo di Gauss, il nome di Jordan si presta ad un equivoco: si riferisce infatti al matematico tedesco Wilhelm Jordan, e non al più famoso matematico francese Camille Jordan, padre della forma normale di Jordan di una matrice. Le tecniche compatte di calcolo sono più recenti: il metodo di Crout è del 1941 e il metodo di Cholesky è stato descritto nel 1924 [6]. Una variante della decomposizione di Cholesky che evita il calcolo delle radici quadrate è proposta in [19].

Il metodo di Gauss è stato uno dei primi metodi ad essere programmato su un calcolatore: già nel 1947 come ricorda Wilkinson in una conferenza sulla storia dei calcolatori [33], Alway e Wilkinson realizzarono i primi programmi di algebra lineare per la versione VII del calcolatore ACE, alla cui progettazione presso il National Physical Laboratory a Teddington lavorava in quel tempo Turing. L'obiettivo era quello di risolvere un sistema lineare di 8 equazioni in 8 incognite.

Le matrici di Givens, da lui presentate nel 1951 per il calcolo degli autovalori di matrici simmetriche, sono state poi descritte nel 1954 in un lavoro [11] in cui compare anche un'analisi dettagliata dell'errore, e viene sviluppata, in modo indipendente da Wilkinson, la tecnica dell'analisi dell'errore all'indietro.

Nel 1958, alla conferenza dell'ACM a Urbana, Householder propose di usare matrici elementari hermitiane unitarie, per ridurre il costo computazionale del metodo di Givens nella riduzione di matrici simmetriche a forma tridiagonale e nella riduzione unitaria di matrici a forma triangolare. Queste matrici, oggi note come matrici di Householder, erano già state studiate da Turnbull e Aitken nel 1930. Un'analisi dettagliata degli errori che si generano con le trasformazioni di Householder, si trova nel libro di Lawson e Hanson [18].

La maggiorazione della funzione di crescita degli elementi nel metodo di Gauss con il massimo pivot totale è stata ottenuta da Wilkinson [29]. Da un'ampia sperimentazione numerica Wilkinson ha anche suggerito la congettura che con il massimo pivot totale risulta

$$\max_{i,j} |a_{ij}^{(k)}| \leq n \max_{i,j} |a_{ij}|, \quad k = 1, \ldots, n-1.$$

Questa congettura è stata dimostrata per matrici ad elementi reali di ordine $n \leq 4$ da Cryer [8], che ha anche trovato che vale il segno di uguaglianza nel caso di certe matrici i cui elementi sono $+1$ e -1, introdotte da Hadamard. Per matrici ad elementi complessi la congettura non è vera, essendo stata costruita una matrice che non la verifica già per $n = 3$. Tecniche come quelle della scalatura e dell'equilibratura della matrice A possono essere utilizzate per diminuire il numero di condizionamento e quindi gli effetti indotti dagli errori di arrotondamento; per questo si veda [10]. L'analisi dell'errore del metodo di Gauss-Jordan è riportata in [22], mentre quella del metodo di Cholesky si trova in [32].

Attualmente la ricerca sui metodi diretti riguarda principalmente i metodi per risolvere sistemi con matrici di grandi dimensioni e sparse [4] e [5]: la tecnologia moderna, che consente un'agevole acquisizione di grandi masse di dati, ha permesso a ricercatori di vari settori di formulare modelli più raffinati e quindi più complessi, con un numero sempre crescente di variabili per descrivere i fenomeni. Ciò ha imposto lo studio di specifiche tecniche per affrontare questo tipo di problemi. Di particolare interessse sono le tecniche di partizionamento di matrici che consentono di sfruttarne la struttura, e le varianti di metodi classici, che mantengono la sparsità delle matrici: si veda ad esempio il lavoro di Björck e Duff in [4], in cui si suggerisce una tecnica di pivot per il metodo di Cholesky che mantiene la sparsità della matrice.

Molti sono i libri nei quali sono esposti i metodi diretti per risolvere problemi lineari: si veda ad esempio [1], [15], [23]; fra quelli che trattano questo problema in modo specifico sono da citare [10], [12], [14], [25]. L'analisi dell'errore è trattata in [9], [30], [31]. Nel libro di Wilkinson e Reinsch [34] sono raccolte le liste di programmi in Algol per la risoluzione dei più significativi problemi di algebra lineare.

Gli studi nel campo della complessità computazionale sono stati avviati nel 1954 con un lavoro di Ostrowski sul problema della complessità del calcolo dei polinomi. Nel campo dell'algebra lineare nel 1965 Klyuev e Kokovkin-Shcherbak [16] hanno dimostrato che il metodo di Gauss è ottimo fra i metodi che utilizzano solo combinazioni di righe e colonne, e nel 1967 Winograd propone un algoritmo per moltiplicare matrici che riduce il numero di operazioni richieste dal metodo classico e che consente quindi di calcolare la decomposizione di Cholesky di una matrice definita positiva con un costo computazionale di $n^3/12$ operazioni moltiplicative. Fondamentale è il risultato ottenuto da Strassen nel 1969 [24] che propone un algoritmo per moltiplicare matrici di ordine n con $O(n^{2.808})$ operazioni, e dimostra la riducibilità del problema della moltiplicazione di due matrici a quello dell'inversione. Per circa dieci anni sono stati fatti tentativi, senza successo, per trovare degli algoritmi migliori di quello di Strassen, poi dal 1978 al 1980 le ricerche condotte da Pan [20] e da Bini, Capovani, Lotti e Romani [2] hanno consentito di ridurre progressivamente l'esponente. Nel 1986 Coppersmith e Winograd [7] hanno ridotto l'esponente a 2.38. Una rassegna sistematica delle ricerche condotte sul problema della moltiplicazione di matrici, con un elenco cronologico dei risultati conseguiti, è riportata nel libro di Pan [21]. Per una presentazione dei più importanti risultati di complessità computazionale numerica si vedano i libri di Kronsjö [17] e di Bini, Capovani, Lotti, Romani [3].

Bibliografia

[1] K. E. Atkinson, *An Introduction to Numerical Analysis*, John Wiley and Sons, New York, 1978.

[2] D. Bini, M. Capovani, G. Lotti, F. Romani, "$O(n^{2.7799})$ Complexity for Matrix Multiplication", *Information Processing Letters*, 8, 5, 1979, pp. 234-235.

[3] D. Bini, M. Capovani, G. Lotti, F. Romani, *Complessità numerica*, Boringhieri, Torino, 1981.

[4] Å. Björk, R. J. Plemmons, H. Schneider, *Large Scale Matrix Problems*, North Holland, New York, 1981.

[5] J. R. Bunch, D. J. Rose, *Sparse Matrix Computations*, Academic Press, New York, 1976.

[6] Commandant Bénoit, "Note sur une méthode de résolution des équations normales, etc. (Procédé du Comm. Cholesky)", *Bull. Géod.*, Toulouse, 1924, 2, pp. 5-77.

[7] D. Coppersmith, S. Winograd, "Matrix Multiplication via Arithmetic Progressions", *Proc. 19th Ann. ACM Symp. on Theory of Computing*, 1987, pp. 1-6.

[8] C. W. Cryer "Pivot Size in Gaussian Elimination", *Numer. Math.* 12, 1968, pp. 335-345.

[9] G. E. Forsythe, M. A. Malcom, C. B. Moler, *Computer Methods for Mathematical Computations*, Prentice Hall, Englewood Cliffs, N. J., 1977.

[10] G. E. Forsythe e C. B. Moler, *Computer Solution of Linear Algebraic Systems*, Prentice Hall, Englewood Cliffs, N. J., 1967.

[11] J. W. Givens, *Numerical Computation of Characteristic Values of a Real Symmetric Matrix*, Oak Ridge National Laboratory, ORNL-1574, 1954.

[12] G. H. Golub, C. F. Van Loan, *Matrix Computations*, 2nd Edition, The Johns Hopkins University Press, Baltimore, Maryland, 1989.

[13] R. T. Gregory, D. L. Karney, *A Collection of Matrices for Testing Computational Algorithms*, J. Wiley and Sons, New York, 1969.

[14] A. S. Householder, *The Theory of Matrices in Numerical Analysis*, Blaisdell, Boston, 1964.

[15] E. Isaacson e H. B. Keller, *Analysis of Numerical Methods*, John Wiley and Sons, New York, 1966.

[16] V. V. Klyuev, H. I. Kokovkin-Shcherbak, "Minimization of the number of arithmetic operations in the solution of linear algebraic systems of equations" *U.S.S.R. Computational Mathematics and Mathematical Physics*, 5, 1965, pp. 25-43.

[17] L. I. Kronsjö, *Algorithms, their Complexity and Efficiency*, J. Wiley and Sons, New York, 1979.

[18] C. L. Lawson, R. J. Hanson, *Solving Least Squares Problems*, Prentice Hall, Englewood Cliffs, N. J., 1974.

[19] R. S. Martin, G. Peters, J. H. Wilkinson, "Symmetric Decomposition of a Positive Definite Matrix", *Numer. Math.*, 7, 1965, pp. 362-383.

[20] V. Y. Pan, "Strassen Algorithm is not Optimal. Trilinear Technique of Aggregating, Uniting and Cancelling for Constructing Fast Algorithms for Matrix Multiplication", *Proc. Nineteenth Ann. Symp. on Foundations of Computer Science*, 1978, pp. 166-176.

[21] V. Y. Pan, "How to Multiply Matrices Faster", Lecture notes in *Computer Science* 179, Springer-Verlag, Berlin, 1984.

[22] G. Peters, J. H. Wilkinson, "On the Stability of Gauss-Jordan Elimination with Pivoting", *Comm. A.C.M.*, 18, 1975, pp. 20-24.

[23] J. Stoer, R. Bulirsch, *Introduction to Numerical Analysis*, Springer-Verlag, New York, 1980.

[24] V. Strassen, "Gaussian Elimination is not Optimal", *Numer. Math.*, 13, 1969, pp. 354-356.

[25] G. W. Stewart, *Introduction to Matrix Computations*, Academic Press, New York, 1973.

[26] A. M. Turing, "Rounding-off Errors in Matrix Processes" *Quart. Jour. Mech. Applied Math.* 1, 1948, pp. 287-308.

[27] J. Von Neumann, H. H. Goldstine "Numerical Inverting of Matrices of High Order", *Amer. Math. Soc. Bull.* 53, 1947, pp. 1021-1099.

[28] J. Von Neumann, H. H. Goldstine "Numerical Inverting of Matrices of High Order II", *Amer. Math. Soc. Proc.* 2, 1951, pp. 188-202.

[29] J. H. Wilkinson, "Error Analysis of Direct Methods of Matrix Inversion", *J. Assoc. Comp. Mach.* 8, 1961, pp. 281-330.

[30] J. H. Wilkinson, *Rounding Errors in Algebraic Processes*, Prentice Hall, Englewood Cliffs, N. J., 1963.

[31] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.

[32] J. H. Wilkinson, "A Priori Error Analysis of Algebraic Processes", *Proc. International Congress Math.*, Izdat. Mir, Moskow, 1968, pp. 629-639.

[33] J. H. Wilkinson, "Turing's Work at the National Physical Laboratory and the Construction of the Pilot ACE, DEUCE, and ACE", in *A History of Computing in the Twentieth Century*, ed. by N. Metropolis, J. Howlett, G-C. Rota, Academic Press, New York, 1980.

[34] J. H. Wilkinson, C. Reinsch, *Handbook for Automatic Computation, vol. 2, Linear Algebra*, Springer-Verlag, New York, 1971.

Capitolo 5

METODI ITERATIVI PER LA RISOLUZIONE DI SISTEMI DI EQUAZIONI LINEARI

1. Successioni di vettori e di matrici

Per risolvere un sistema lineare $A\mathbf{x} = \mathbf{b}$, oltre ai metodi diretti, si possono utilizzare anche i metodi iterativi, che risultano particolarmente convenienti se la matrice A è sparsa, cioè se il numero degli elementi non nulli di A è dell'ordine della dimensione della matrice. Infatti quando si utilizza un metodo di risoluzione diretto, ad esempio il metodo di Gauss, può accadere che nelle matrici intermedie vengano generati molti elementi diversi da zero in corrispondenza ad elementi nulli della matrice iniziale (questo fenomeno si chiama *fill-in*). Poiché i metodi diretti non sfruttano adeguatamente la sparsità della matrice, per questo tipo di problemi, soprattutto se A è di grandi dimensioni, può essere più conveniente utilizzare un metodo iterativo. Esistono però dei casi nei quali la matrice A è sparsa, ma è comunque conveniente applicare dei metodi diretti che sfruttano specifiche proprietà di struttura della matrice.

5.1 Definizione. Una successione $\{\mathbf{x}^{(k)}\}$ di vettori di $\mathbf{C}^n$ si dice *convergente* al vettore $\mathbf{x}^*$ di $\mathbf{C}^n$ se esiste una norma per cui risulta

$$\lim_{k\to\infty} \|\mathbf{x}^{(k)} - \mathbf{x}^*\| = 0; \tag{1}$$

in tal caso si pone

$$\lim_{k\to\infty} \mathbf{x}^{(k)} = \mathbf{x}^*. \qquad \blacksquare$$

Per il teorema 3.4 di equivalenza delle norme su $\mathbf{C}^n$, la definizione 5.1 non dipende dalla particolare norma considerata. La condizione di convergenza data dalla (1) si traduce in una condizione di convergenza delle successioni formate dalle singole componenti. Infatti, considerando la norma ∞, poiché è

$$|x_i^{(k)} - x_i^*| \le \|\mathbf{x}^{(k)} - \mathbf{x}^*\|_\infty, \quad i = 1, \ldots, n,$$

dalla (1) si ha

$$\lim_{k\to\infty} |x_i^{(k)} - x_i^*| = 0,$$

e quindi

$$\lim_{k\to\infty} x_i^{(k)} = x_i^*, \quad i = 1, ..., n; \tag{2}$$

viceversa, se vale la (2), è ovviamente verificata la condizione (1), per la norma ∞.

Per le successioni di matrici $\{A^{(k)}\}$ si può dare una definizione di convergenza analoga alla 5.1.

Il seguente teorema è di fondamentale importanza nello studio della convergenza dei metodi iterativi per la risoluzione dei sistemi lineari.

5.2 Teorema. *Sia $A \in \mathbf{C}^{n\times n}$, allora*

$$\lim_{k\to\infty} A^k = O \quad \textit{se e solo se} \quad \rho(A) < 1.$$

Dim. Per il teorema 2.18 esiste una matrice non singolare $T \in \mathbf{C}^{n\times n}$, tale che $A = TJT^{-1}$, dove J è la forma normale di Jordan di A; allora risulta

$$A^k = TJ^kT^{-1}. \tag{3}$$

Usando la notazione del teorema 2.18, risulta

$$J^k = \begin{bmatrix} J_1^k & & & \\ & J_2^k & & \\ & & \ddots & \\ & & & J_p^k \end{bmatrix},$$

dove

$$J_i^k = \begin{bmatrix} [C_i^{(1)}]^k & & & \\ & [C_i^{(2)}]^k & & \\ & & \ddots & \\ & & & [C_i^{(\tau(\lambda_i))}]^k \end{bmatrix},$$

per $i = 1, \ldots, p$, e i blocchi $C_i^{(j)} \in \mathbf{C}^{\nu_i(j)\times\nu_i(j)}$ per $j = 1, \ldots, \tau(\lambda_i)$ sono della forma

$$C_i^{(j)} = \lambda_i I + U,$$

in cui

$$U = \begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \\ & & & 0 \end{bmatrix}.$$

Per ogni i e j, $1 \le i \le p$, $1 \le j \le \tau(\lambda_i)$, risulta

$$[C_i^{(j)}]^k = (\lambda_i I + U)^k = \sum_{r=0}^{k} \binom{k}{r} \lambda_i^{k-r} U^r,$$

assumendo $U^0 = I$. Posto $s = \nu_i^{(j)}$, per $r \ge s$ risulta $U^r = O$, e quindi per $k \ge s$ è

$$[C_i^{(j)}]^k = \sum_{r=0}^{s-1} \binom{k}{r} \lambda_i^{k-r} U^r = \begin{bmatrix} \lambda_i^k & \binom{k}{1}\lambda_i^{k-1} & \cdots & \binom{k}{s-1}\lambda_i^{k-s+1} \\ & \lambda_i^k & \ddots & \\ & & \ddots & \binom{k}{1}\lambda_i^{k-1} \\ & & & \lambda_i^k \end{bmatrix}. \quad (4)$$

Ne segue che condizione necessaria e sufficiente affinché λ_i^k e $\binom{k}{r}\lambda_i^{k-r}$ tendano a zero per $k \to \infty$ è che sia $|\lambda_i| < 1$ per $i = 1, \ldots, n$, cioè $\rho(A) < 1$.

∎

5.3 Esempio. La matrice

$$E = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

ha autovalori $\lambda_1 = \lambda_2 = 0$, $\lambda_3 = 3$. Quindi risulta

$$\rho(E) = 3 \text{ e } \lim_{k\to\infty} E^k \ne O.$$

Si osservi infatti che per $k \ge 1$ è

$$E^k = 3^{k-1} E.$$

La matrice $F = \frac{1}{4} E$ ha autovalori $\lambda_1 = \lambda_2 = 0$, $\lambda_3 = \frac{3}{4}$. Quindi risulta

$$\rho(F) = \frac{3}{4} < 1 \text{ e } \lim_{k\to\infty} F^k = O.$$

Si osservi infatti che per $k \ge 1$ è

$$F^k = \left(\frac{3}{4}\right)^{k-1} F.$$

∎

5.4 Teorema. *Sia $A \in \mathbf{C}^{n\times n}$. Allora*

$$\det(I-A)\neq 0 \ e \ \lim_{k\to\infty}\sum_{i=0}^{k} A^i = (I-A)^{-1} \quad \textit{se e solo se} \quad \rho(A)<1.$$

Dim. Sia $\rho(A)<1$, allora gli autovalori di A hanno tutti modulo minore di 1, quindi la matrice $I-A$ non ha autovalori nulli e risulta non singolare. Inoltre, poiché

$$(I-A)\sum_{i=0}^{k} A^i = I - A^{k+1},$$

si ha

$$\sum_{i=0}^{k} A^i = (I-A)^{-1}(I-A^{k+1}),$$

e quindi

$$\begin{aligned}\lim_{k\to\infty}\sum_{i=0}^{k} A^i &= \lim_{k\to\infty}(I-A)^{-1}(I-A^{k+1})\\ &= (I-A)^{-1}\lim_{k\to\infty}(I-A^{k+1}) = (I-A)^{-1},\end{aligned}$$

in quanto per il teorema 5.2 si ha $\lim_{k\to\infty} A^{k+1} = O$. Viceversa, sia $I-A$ non singolare e

$$\lim_{k\to\infty}\sum_{i=0}^{k} A^i = (I-A)^{-1}.$$

Indicato con λ un autovalore di A tale che $|\lambda| = \rho(A)$ e con $\mathbf{x}$ un autovettore corrispondente a λ, è $\lambda \neq 1$ perché $I-A$ è non singolare, ed inoltre vale

$$\lim_{k\to\infty}\sum_{i=0}^{k} A^i\mathbf{x} = (I-A)^{-1}\mathbf{x}$$

e quindi

$$\lim_{k\to\infty}\Big(\sum_{i=0}^{k}\lambda^i\Big)\mathbf{x} = \frac{1}{1-\lambda}\,\mathbf{x}.$$

Ne segue la convergenza della serie numerica

$$\sum_{i=0}^{\infty}\lambda^i = \frac{1}{1-\lambda},$$

per cui $|\lambda| < 1$. ∎

Come per le serie numeriche, si usa scrivere

$$\sum_{i=0}^{\infty} A^i = (I - A)^{-1}.$$

2. Generalità sui metodi iterativi

Sia $A \in \mathbf{C}^{n\times n}$ una matrice non singolare e si consideri la decomposizione di A nella forma

$$A = M - N, \tag{5}$$

dove M è una matrice non singolare. Dalla (5), sostituendo nel sistema lineare

$$A\mathbf{x} = \mathbf{b}, \tag{6}$$

risulta

$$M\mathbf{x} - N\mathbf{x} = \mathbf{b},$$

cioè

$$\mathbf{x} = M^{-1}N\mathbf{x} + M^{-1}\mathbf{b}.$$

Posto

$$P = M^{-1}N \quad \text{e} \quad \mathbf{q} = M^{-1}\mathbf{b}, \tag{7}$$

si ottiene il seguente sistema

$$\mathbf{x} = P\mathbf{x} + \mathbf{q}, \tag{8}$$

equivalente al sistema (6).

Dato un vettore iniziale $\mathbf{x}^{(0)}$, si considera la successione $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots$, così definita

$$\mathbf{x}^{(k)} = P\mathbf{x}^{(k-1)} + \mathbf{q}, \quad k = 1, 2, \ldots\ . \tag{9}$$

Se la successione $\mathbf{x}^{(k)}$ è convergente e si indica con

$$\mathbf{x}^* = \lim_{k\to\infty} \mathbf{x}^{(k)},$$

allora passando al limite nella (9) risulta

$$\mathbf{x}^* = P\mathbf{x}^* + \mathbf{q}, \tag{10}$$

cioè $\mathbf{x}^*$ è la soluzione del sistema (8) e quindi del sistema (6).

La relazione (9) individua un *metodo iterativo* in cui, partendo da un vettore iniziale $\mathbf{x}^{(0)}$, la soluzione viene approssimata utilizzando una successione $\{\mathbf{x}^{(k)}\}$ di vettori. La matrice P si dice *matrice di iterazione del metodo*.

Al variare del vettore iniziale $\mathbf{x}^{(0)}$ si ottengono dalla (9) diverse successioni $\{\mathbf{x}^{(k)}\}$, alcune delle quali possono essere convergenti ed altre no. Un metodo iterativo è detto *convergente* se, qualunque sia il vettore iniziale $\mathbf{x}^{(0)}$, la successione $\{\mathbf{x}^{(k)}\}$ è convergente.

5.5 Esempio. Si consideri il sistema (8) in cui

$$P = \begin{bmatrix} \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & 2 \end{bmatrix}, \quad \mathbf{q} = \mathbf{0}, \quad \text{e quindi} \quad \mathbf{x}^* = \mathbf{0}.$$

Allora

$$P^k = \begin{bmatrix} \left(\frac{1}{2}\right)^k & 0 & 0 \\ 0 & \left(\frac{1}{2}\right)^k & 0 \\ 0 & 0 & 2^k \end{bmatrix}.$$

Se $\mathbf{x}^{(0)} = [1, 0, 0]^T$, si ottiene la successione

$$\mathbf{x}^{(k)} = [\left(\tfrac{1}{2}\right)^k, 0, 0]^T, \ k = 1, 2, \ldots,$$

che converge alla soluzione del sistema. Se invece $\mathbf{x}^{(0)} = [0, 1, 1]^T$, si ottiene la successione

$$\mathbf{x}^{(k)} = [0, \left(\tfrac{1}{2}\right)^k, 2^k]^T, \ k = 1, 2, \ldots,$$

che non converge. Questo è un esempio di metodo non convergente. ∎

5.6 Teorema. *Il metodo iterativo (9) è convergente se e solo se* $\rho(P) < 1$.

Dim. Sia $\mathbf{x}^*$ la soluzione del sistema (6), che soddisfa quindi la (10). Sottraendo membro a membro la (9) dalla (10) risulta

$$\mathbf{x}^* - \mathbf{x}^{(k)} = P(\mathbf{x}^* - \mathbf{x}^{(k-1)}), \quad k = 1, 2, \ldots \ . \tag{11}$$

Indicato con

$$\mathbf{e}^{(k)} = \mathbf{x}^* - \mathbf{x}^{(k)},$$

il vettore *errore* alla k-esima iterazione, si ha dalla (11)

$$\mathbf{e}^{(k)} = P\mathbf{e}^{(k-1)}, \quad k = 1, 2, \ldots \tag{12}$$

e quindi

$$\mathbf{e}^{(k)} = P\mathbf{e}^{(k-1)} = P^2\mathbf{e}^{(k-2)} = \ldots = P^k\mathbf{e}^{(0)}. \tag{13}$$

Se $\rho(P) < 1$, per il teorema 5.2 risulta

$$\lim_{k\to\infty} P^k = O,$$

e dalla (13), per ogni vettore $\mathbf{e}^{(0)}$, segue che

$$\lim_{k\to\infty} \mathbf{e}^{(k)} = \mathbf{0}. \tag{14}$$

Viceversa, se il metodo è convergente, la (14) vale per ogni $\mathbf{x}^{(0)}$, e in particolare deve valere se $\mathbf{x}^{(0)}$ è tale che il vettore $\mathbf{e}^{(0)} = \mathbf{x}^* - \mathbf{x}^{(0)}$ è un autovettore di P corrispondente ad un autovalore λ di modulo massimo, cioè $|\lambda| = \rho(P)$. In questo caso risulta

$$P\mathbf{e}^{(0)} = \lambda\mathbf{e}^{(0)}$$

e quindi

$$\mathbf{e}^{(k)} = P^k\mathbf{e}^{(0)} = \lambda^k\mathbf{e}^{(0)}.$$

Ne segue che

$$\lim_{k\to\infty} [\rho(P)]^k = 0$$

e quindi $\rho(P) < 1$. ∎

La condizione $\rho(P) < 1$, necessaria e sufficiente per la convergenza del metodo (9), non è in generale di agevole verifica. Conviene allora utilizzare, quando è possibile, delle condizioni sufficienti di convergenza di più facile verifica. Una tale condizione è data nel seguente teorema.

5.7 Teorema. *Se esiste una norma matriciale indotta* $\| \cdot \|$ *per cui* $\|P\| < 1$, *il metodo iterativo (9) è convergente.*

Dim. La tesi segue dal teorema 5.6 e dalla proprietà

$$\rho(P) \leq \|P\|,$$

dimostrata nel teorema 3.10. ∎

Poiché il determinante di una matrice è uguale al prodotto degli autovalori, se $|\det P| \geq 1$, almeno uno degli autovalori di P è in modulo maggiore o uguale a 1 e quindi il metodo (9) non è convergente. Poiché la traccia di una matrice è uguale alla somma degli autovalori, se $|\text{tr } P| \geq n$, almeno uno degli autovalori di P è in modulo maggiore o uguale a 1 e quindi il metodo (9) non è convergente. Quindi le condizioni $|\det P| < 1$ e $|\text{tr } P| < n$ sono necessarie affinché il metodo iterativo (9) sia convergente.

3. Controllo della convergenza

Fissata una norma vettoriale $\| \cdot \|$ e la corrispondente norma matriciale indotta, dalla (13) si ottiene la seguente maggiorazione della norma dell'errore da cui è affetto $\mathbf{x}^{(k)}$ rispetto alla soluzione del sistema $\mathbf{x}^*$:

$$\|\mathbf{e}^{(k)}\| \leq \|P^k\| \, \|\mathbf{e}^{(0)}\|, \tag{15}$$

dove il segno di uguaglianza vale per particolari vettori $\mathbf{e}^{(0)}$, perché la norma matriciale considerata è indotta. Quindi $\|P^k\|$ esprime la riduzione, rispetto all'errore iniziale, dell'errore al k-esimo passo. Questa misura risulta però inadatta per una valutazione della velocità di convergenza di un metodo, che sia indipendente dal numero delle iterazioni. Infatti, se P e Q sono due matrici di iterazione associate a due diversi metodi, può accadere che per una particolare norma $\| \cdot \|$ esistano due interi j e k, con $k \neq j$, tali che

$$\|P^k\| < \|Q^k\| \quad \text{e} \quad \|P^j\| > \|Q^j\|.$$

5.8 Esempio. Siano

$$P = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.6 \end{bmatrix}, \quad Q = \begin{bmatrix} 0.5 & 0.25 \\ 0 & 0.5 \end{bmatrix}.$$

Si ha

$$P^k = \begin{bmatrix} 0.5^k & 0 \\ 0 & 0.6^k \end{bmatrix}, \quad Q^k = \begin{bmatrix} 0.5^k & k\ 0.5^{k+1} \\ 0 & 0.5^k \end{bmatrix}.$$

Utilizzando la norma ∞ risulta

$$\|P^k\|_\infty = 0.6^k \quad \text{e} \quad \|Q^k\|_\infty = (2+k)0.5^{k+1}.$$

Per $k = 1, \ldots, 15$ si ottengono i valori

k	$\|P^k\|_\infty$	$\|Q^k\|_\infty$
1	0.6000000	0.7500000
2	0.3600000	0.5000000
3	0.2160000	0.3125000
.	.	.
.	.	.
8	$0.1679614\ 10^{-1}$	$0.1953125\ 10^{-1}$
9	$0.1007769\ 10^{-1}$	$0.1074219\ 10^{-1}$
10	$0.6046608\ 10^{-2}$	$0.5859375\ 10^{-2}$
11	$0.3627964\ 10^{-2}$	$0.3173828\ 10^{-2}$
.	.	.
.	.	.
15	$0.4701840\ 10^{-3}$	$0.2593994\ 10^{-3}$

Si noti che $\|P^k\|_\infty < \|Q^k\|_\infty$ per $k \leq 9$, e $\|P^k\|_\infty > \|Q^k\|_\infty$ per $k \geq 10$. Utilizzando la norma 2, per $k = 1, \ldots, 15$, si ottengono i valori

k	$\|P^k\|_2$	$\|Q^k\|_2$
1	0.6000000	0.6403882
2	0.3600000	0.4045085
3	0.2160000	0.2500000
.	.	.
.	.	.
6	$0.4665595\ 10^{-1}$	$0.5160587\ 10^{-1}$
7	$0.2799357\ 10^{-1}$	$0.2941847\ 10^{-1}$
8	$0.1679614\ 10^{-1}$	$0.1654713\ 10^{-1}$
9	$0.1007769\ 10^{-1}$	$0.9203542\ 10^{-2}$
.	.	.
.	.	.
15	$0.4701840\ 10^{-3}$	$0.2328810\ 10^{-3}$

e quindi $\|P^k\|_2 < \|Q^k\|_2$ per $k \leq 7$, e $\|P^k\|_2 > \|Q^k\|_2$ per $k \geq 8$. ∎

Se $\mathbf{e}^{(k-1)} \neq \mathbf{0}$, la quantità $\|\mathbf{e}^{(k)}\|/\|\mathbf{e}^{(k-1)}\|$ esprime la riduzione dell'errore al k-esimo passo e la media geometrica delle riduzioni dell'errore sui primi k passi:

$$\sigma_k = \sqrt[k]{\frac{\|\mathbf{e}^{(1)}\|}{\|\mathbf{e}^{(0)}\|}\ \frac{\|\mathbf{e}^{(2)}\|}{\|\mathbf{e}^{(1)}\|}\ \cdots\ \frac{\|\mathbf{e}^{(k)}\|}{\|\mathbf{e}^{(k-1)}\|}} = \sqrt[k]{\frac{\|\mathbf{e}^{(k)}\|}{\|\mathbf{e}^{(0)}\|}}$$

esprime la *riduzione media per passo* dell'errore relativo ai primi k passi. Dalla (15) risulta

$$\sigma_k \leq \sqrt[k]{\|P^k\|},$$

dove il segno di uguaglianza vale per particolari vettori $\mathbf{e}^{(0)}$. La quantità che si ottiene facendo tendere k all'infinito esprime la *riduzione asintotica media per passo* e, come risulta dal seguente teorema, è indipendente dalla particolare norma utilizzata.

5.9 Teorema. *Sia $A \in \mathbf{C}^{n\times n}$ e sia $\|\ .\ \|$ una qualunque norma indotta. Allora*

$$\lim_{k\to\infty} \sqrt[k]{\|A^k\|} = \rho(A).$$

Dim. Si dimostra prima che il limite, se esiste, non dipende dalla particolare norma usata. Per l'equivalenza delle norme, se $\|\ .\ \|'$ e $\|\ .\ \|''$ sono due norme matriciali indotte, esistono due costanti α e β positive, tali che

$$\alpha\ \|A^k\|'' \leq \|A^k\|' \leq \beta\ \|A^k\|'',$$

per cui

$$\sqrt[k]{\alpha}\,\sqrt[k]{\|A^k\|''} \leq \sqrt[k]{\|A^k\|'} \leq \sqrt[k]{\beta}\,\sqrt[k]{\|A^k\|''}.$$

Poiché

$$\lim_{k\to\infty} \sqrt[k]{\alpha} = \lim_{k\to\infty} \sqrt[k]{\beta} = 1,$$

dalla relazione precedente segue che se esiste

$$\lim_{k\to\infty} \sqrt[k]{\|A^k\|''},$$

allora esiste anche

$$\lim_{k\to\infty} \sqrt[k]{\|A^k\|'},$$

e tali limiti coincidono. Si dimostra adesso che il limite esiste per un'opportuna norma indotta. Dalla (3) si ha $A^k = TJ^kT^{-1}$, dove J^k è una matrice diagonale formata dai blocchi $[C_i^{(j)}]^k$, $i = 1, \ldots, p$, $j = 1, \ldots, \tau(\lambda_i)$, in cui λ_i, $i = 1, \ldots, p$, sono gli autovalori distinti di A e i blocchi $[C_i^{(j)}]^k$ sono quelli riportati nella (4). Per il teorema 3.11 l'applicazione

$$A \to \|T^{-1}AT\|_\infty$$

è una norma indotta di A e, indicando con $\|\ .\ \|$ tale norma, risulta $\|A^k\| = \|J^k\|_\infty$. Se λ_1 è l'autovalore di A per cui $|\lambda_1| = \rho(A)$, e, fra tutti i blocchi relativi a $\lambda_1, C_1^{(1)}$ è quello di ordine s massimo, allora esiste un intero k_0 tale che per ogni $k \geq k_0$ si ha

$$\|A^k\| = \|[C_1^{(1)}]^k\|_\infty = \sum_{r=0}^{s-1} \binom{k}{r} |\lambda_1|^{k-r} = [\rho(A)]^k \sum_{r=0}^{s-1} \binom{k}{r} [\rho(A)]^{-r}.$$

La quantità

$$p(k) = \sum_{r=0}^{s-1} \binom{k}{r} [\rho(A)]^{-r}$$

è un polinomio in k di grado $s-1$, e quindi

$$\lim_{k\to\infty} \sqrt[k]{p(k)} = 1.$$

Ne segue che il

$$\lim_{k\to\infty} \sqrt[k]{\|A^k\|}$$

esiste e vale $\rho(A)$. ∎

La quantità $\rho(P)$, indipendente dalla norma utilizzata e dall'indice di iterazione k, viene quindi assunta come misura della velocità di convergenza del metodo (9). Il numero k di iterazioni richieste per ridurre l'errore di $1/10$ (cioè, approssimativamente, per ottenere una cifra decimale in più) è tale che

$$[\rho(P)]^k \approx \frac{1}{10}, \quad \text{da cui} \quad k \approx -1/\log_{10}\rho(P).$$

5.10 Definizione. Si definisce *tasso asintotico di convergenza* del metodo iterativo (9) la costante $R = -\log_{10}\rho(P)$. ∎

Poiché con un metodo iterativo non è ovviamente possibile calcolare in generale la soluzione con un numero finito di iterazioni, occorre individuare dei criteri per l'arresto del procedimento. I criteri più comunemente usati, fissata una tolleranza ϵ, che tiene conto anche della precisione utilizzata nei calcoli, sono i seguenti:

$$\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\| \leq \epsilon, \tag{16}$$

oppure, se $\mathbf{x}^{(k)} \neq \mathbf{0}$,

$$\frac{\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|}{\|\mathbf{x}^{(k)}\|} \leq \epsilon. \tag{17}$$

Si noti però che le condizioni (16) e (17) non garantiscono che la soluzione sia stata approssimata con la precisione ϵ. Infatti per la (12) è:

$$\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)} = [\mathbf{x}^* - \mathbf{x}^{(k-1)}] - [\mathbf{x}^* - \mathbf{x}^{(k)}] = \mathbf{e}^{(k-1)} - \mathbf{e}^{(k)} = (I - P)\mathbf{e}^{(k-1)}$$

e, passando alle norme, se $\|P\| < 1$, per il teorema 3.13 si ha:

$$\|\mathbf{e}^{(k-1)}\| \leq \|(I-P)^{-1}\|\ \|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\| \leq \frac{\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|}{1 - \|P\|},$$

per cui può accadere che $\|\mathbf{e}^{(k-1)}\|$ sia elevata anche se la condizione (16) è verificata.

In un programma che implementa un metodo iterativo deve essere comunque previsto un controllo per interrompere l'esecuzione quando il numero delle iterazioni diventa troppo elevato. Può anche accadere che un metodo iterativo la cui matrice di iterazione P è tale che $\rho(P) < 1$, per gli effetti indotti dagli errori di arrotondamento non converga in pratica, e questo accade, in particolare, quando la matrice A è fortemente mal condizionata e $\rho(P)$ è molto vicino ad 1.

È opportuno rilevare che un metodo iterativo rispetto ad un metodo diretto è in generale meno sensibile alla propagazione degli errori. Infatti

il vettore $\mathbf{x}^{(k)}$ può essere considerato come il vettore generato con una sola iterazione a partire dal vettore iniziale $\mathbf{x}^{(k-1)}$, e quindi risulta affetto dagli errori di arrotondamento generati dalla sola ultima iterazione.

In un metodo iterativo ad ogni iterazione il costo computazionale è principalmente determinato dalla operazione di moltiplicazione della matrice P per un vettore, che richiede n^2 operazioni moltiplicative se la matrice A non ha specifiche proprietà. Se invece A è sparsa, cioè ha un numero di elementi non nulli dell'ordine di n, la moltiplicazione di P per un vettore richiede un numero di operazioni moltiplicative dell'ordine di n. In questo caso i metodi iterativi possono risultare competitivi con quelli diretti. Particolarmente interessante è il caso in cui la matrice, oltre a essere sparsa, ha specifiche proprietà di struttura, che possono essere convenientemente sfruttate anche per ridurre l'ingombro di memoria richiesto.

4. Metodi iterativi di Jacobi e Gauss-Seidel

Fra i metodi iterativi individuati da una particolare scelta della decomposizione (5) sono particolarmente importanti il metodo di Jacobi e il metodo di Gauss-Seidel, per i quali è possibile dare delle condizioni sufficienti di convergenza verificate da molte delle matrici che si ottengono risolvendo problemi differenziali.

Si consideri la decomposizione della matrice A

$$A = D - B - C$$

dove

$$d_{ij} = \begin{cases} a_{ij} & \text{se } i = j \\ 0 & \text{se } i \neq j, \end{cases} \qquad b_{ij} = \begin{cases} \text{-}a_{ij} & \text{se } i > j \\ 0 & \text{se } i \leq j, \end{cases} \qquad c_{ij} = \begin{cases} 0 & \text{se } i \geq j \\ \text{-}a_{ij} & \text{se } i < j. \end{cases}$$

Scegliendo $M = D, \quad N = B + C$, si ottiene il *metodo di Jacobi*.
Scegliendo $M = D - B, \quad N = C$, si ottiene il *metodo di Gauss-Seidel*.

Per queste decomposizioni risulta $\det M \neq 0$ se e solo se tutti gli elementi principali di A sono non nulli.

Indicando con J la matrice di iterazione del metodo di Jacobi, dalla (7) si ha

$$J = D^{-1}(B + C),$$

per cui la (9) diviene:

$$\mathbf{x}^{(k)} = J\mathbf{x}^{(k-1)} + D^{-1}\mathbf{b}$$

e, in termini di componenti :

$$x_i^{(k)} = \frac{1}{a_{ii}} \left[b_i - \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} x_j^{(k-1)} \right], \quad i = 1, 2, \ldots, n. \tag{18}$$

Il metodo di Jacobi è detto anche *metodo degli spostamenti simultanei*, in quanto le componenti del vettore $\mathbf{x}^{(k)}$ sostituiscono simultaneamente al termine dell'iterazione le componenti di $\mathbf{x}^{(k-1)}$.

Indicando con G la matrice di iterazione del metodo di Gauss-Seidel, dalla (7) si ha

$$G = (D - B)^{-1} C,$$

per cui la (9) diviene:

$$\mathbf{x}^{(k)} = G\mathbf{x}^{(k-1)} + (D - B)^{-1}\mathbf{b}. \tag{19}$$

Per descrivere la (19) in termini di componenti, conviene prima trasformarla nel modo seguente:

$$(D - B)\mathbf{x}^{(k)} = C\mathbf{x}^{(k-1)} + \mathbf{b}$$

$$D\mathbf{x}^{(k)} = B\mathbf{x}^{(k)} + C\mathbf{x}^{(k-1)} + \mathbf{b}$$

$$\mathbf{x}^{(k)} = D^{-1}B\mathbf{x}^{(k)} + D^{-1}C\mathbf{x}^{(k-1)} + D^{-1}\mathbf{b}, \tag{20}$$

ottenendo quindi:

$$x_i^{(k)} = \frac{1}{a_{ii}} \left[b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k-1)} \right], \quad i = 1, 2, \ldots, n. \tag{21}$$

Confrontando la (21) con la (18), risulta che nel metodo di Gauss-Seidel per calcolare le componenti del vettore $\mathbf{x}^{(k)}$ (contrariamente a quanto accade nel metodo di Jacobi) sono utilizzate componenti già calcolate dello stesso vettore. Per questo motivo il metodo prende anche il nome di *metodo degli spostamenti successivi*. Quindi nella implementazione del metodo di Jacobi è necessario disporre, contemporaneamente, di entrambi i vettori $\mathbf{x}^{(k)}$ e $\mathbf{x}^{(k-1)}$, mentre per il metodo di Gauss-Seidel è sufficiente disporre di un solo vettore.

In molte applicazioni il metodo di Gauss-Seidel, che utilizza immediatamente i valori calcolati nella iterazione corrente, risulta più veloce del metodo di Jacobi. Però esistono casi in cui risulta non solo che il metodo di Jacobi sia più veloce del metodo di Gauss-Seidel, ma anche che il metodo di Jacobi sia convergente e quello di Gauss-Seidel no.

5.11 Esempi. Si esamina la convergenza dei metodi di Jacobi e di Gauss-Seidel applicati al sistema $A\mathbf{x} = \mathbf{b}$, per diverse matrici $A \in \mathbf{R}^{3\times 3}$. Il vettore $\mathbf{b}$ è sempre scelto in modo che la soluzione sia $\mathbf{x}^* = [1, 1, 1]^T$. Il criterio di arresto utilizzato è quello espresso dalla (16), in norma ∞, con $\epsilon = 10^{-5}$. Si noti che, poiché $\|\mathbf{x}^*\|_\infty = 1$, in questo caso i criteri di arresto espressi dalla (16) e dalla (17) da un certo valore di k in poi sono equivalenti.

Nelle figure sono riportati i grafici delle norme degli errori assoluti $\|\mathbf{e}^{(k)}\|_\infty$ delle successioni ottenute a partire dal vettore iniziale $\mathbf{x}^{(0)} = \mathbf{0}$. Con i quadratini vuoti sono indicati gli errori generati dal metodo di Jacobi, con i quadratini pieni gli errori generati dal metodo di Gauss-Seidel.

a) Nel caso

$$A = \begin{bmatrix} 3 & 0 & 4 \\ 7 & 4 & 2 \\ -1 & -1 & -2 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 7 \\ 13 \\ -4 \end{bmatrix} \tag{22}$$

risulta

$$J = \begin{bmatrix} 0 & 0 & -\frac{4}{3} \\ -\frac{7}{4} & 0 & -\frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{2} & 0 \end{bmatrix} \quad G = \begin{bmatrix} 0 & 0 & -\frac{4}{3} \\ 0 & 0 & \frac{11}{6} \\ 0 & 0 & -\frac{1}{4} \end{bmatrix}$$

e $\rho(J) = 1.337510$, $\rho(G) = 0.25$. Quindi il metodo di Gauss-Seidel è convergente mentre il metodo di Jacobi non lo è. La successione ottenuta con il metodo di Gauss-Seidel si arresta alla 11-esima iterazione. I grafici degli errori sono riportati nella figura 5.1.

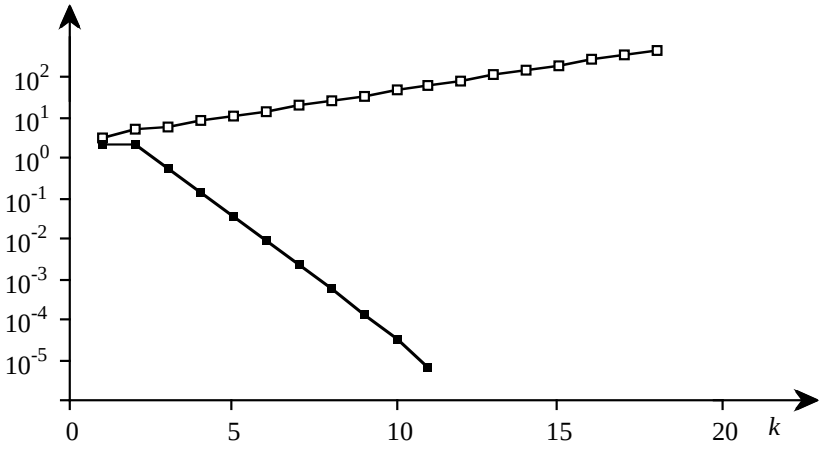


Fig. 5.1 - Grafici degli errori dei metodi di Jacobi e di Gauss-Seidel per il problema (22).

b) Nel caso

$$A = \begin{bmatrix} -3 & 3 & -6 \\ -4 & 7 & -8 \\ 5 & 7 & -9 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} -6 \\ -5 \\ 3 \end{bmatrix} \tag{23}$$

risulta

$$J = \begin{bmatrix} 0 & 1 & -2 \\ \frac{4}{7} & 0 & \frac{8}{7} \\ \frac{5}{9} & \frac{7}{9} & 0 \end{bmatrix} \quad G = \begin{bmatrix} 0 & 1 & -2 \\ 0 & \frac{4}{7} & 0 \\ 0 & 1 & -\frac{10}{9} \end{bmatrix}$$

e $\rho(J) = 0.8133091$, $\rho(G) = 1.111111$. Quindi il metodo di Jacobi è convergente e il metodo di Gauss-Seidel non lo è. La successione ottenuta con il metodo di Jacobi si arresta alla 49-esima iterazione. I grafici degli errori sono riportati nella figura 5.2.

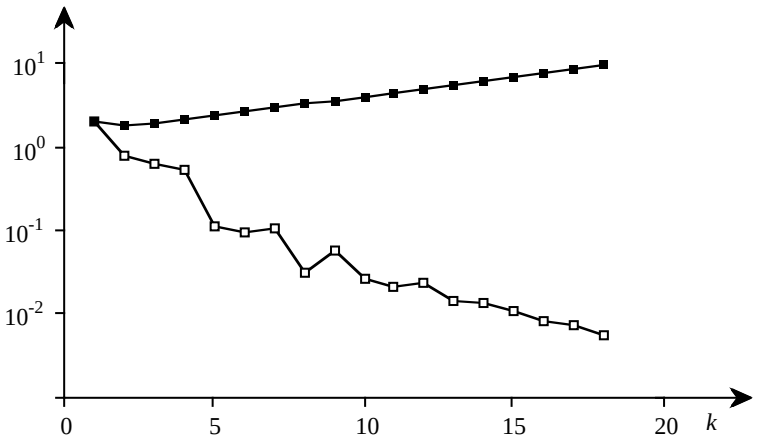


Fig. 5.2 - Grafici degli errori dei metodi di Jacobi e di Gauss-Seidel per il problema (23).

c) Nel caso

$$A = \begin{bmatrix} 4 & 1 & 1 \\ 2 & -9 & 0 \\ 0 & -8 & -6 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 6 \\ -7 \\ -14 \end{bmatrix} \tag{24}$$

risulta

$$J = \begin{bmatrix} 0 & -\frac{1}{4} & -\frac{1}{4} \\ \frac{2}{9} & 0 & 0 \\ 0 & -\frac{4}{3} & 0 \end{bmatrix} \quad G = \begin{bmatrix} 0 & -\frac{1}{4} & -\frac{1}{4} \\ 0 & -\frac{1}{18} & -\frac{1}{18} \\ 0 & \frac{2}{27} & \frac{2}{27} \end{bmatrix}$$

e $\rho(J) = 0.4438188, \quad \rho(G) = 0.01851852$. Quindi entrambi i metodi sono convergenti e la successione generata dal metodo di Gauss-Seidel, che si arresta alla quinta iterazione, converge più rapidamente di quella generata dal metodo di Jacobi, che si arresta alla 16-esima iterazione. I grafici degli errori sono riportati nella figura 5.3.

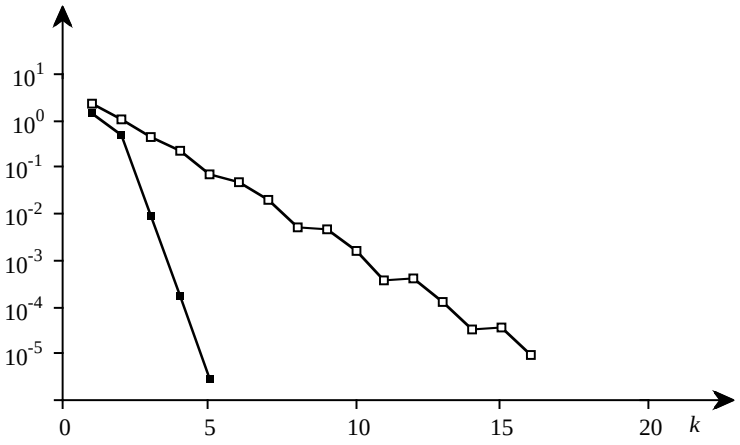


Fig. 5.3 - Grafici degli errori dei metodi di Jacobi e di Gauss-Seidel per il problema (24).

d) Nel caso

$$A = \begin{bmatrix} 7 & 6 & 9 \\ 4 & 5 & -4 \\ -7 & -3 & 8 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 22 \\ 5 \\ -2 \end{bmatrix} \tag{25}$$

risulta

$$J = \begin{bmatrix} 0 & -\frac{6}{7} & -\frac{9}{7} \\ -\frac{4}{5} & 0 & \frac{4}{5} \\ \frac{7}{8} & \frac{3}{8} & 0 \end{bmatrix} \quad G = \begin{bmatrix} 0 & -\frac{6}{7} & -\frac{9}{7} \\ 0 & \frac{24}{35} & \frac{64}{35} \\ 0 & -\frac{69}{140} & -\frac{123}{280} \end{bmatrix}$$

e $\rho(J) = 0.6411328, \quad \rho(G) = 0.7745967$. Quindi entrambi i metodi sono convergenti e la successione generata dal metodo di Jacobi, che si arresta alla 30-esima iterazione, converge più rapidamente di quella generata dal metodo di Gauss-Seidel, che si arresta alla 48-esima iterazione. I grafici degli errori sono riportati nella figura 5.4. ∎

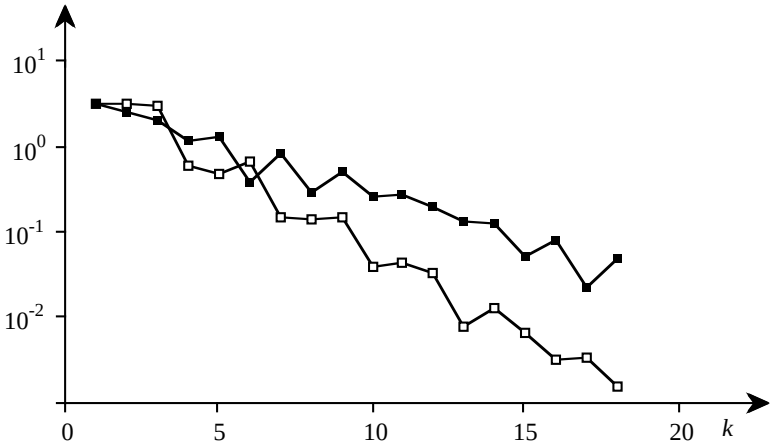


Fig. 5.4 - Grafici degli errori dei metodi di Jacobi e di Gauss-Seidel per il problema (25).

Dai teoremi 5.6 e 5.7 si possono ricavare delle condizioni di convergenza per i metodi di Jacobi e di Gauss-Seidel, applicati alla risoluzione del sistema lineare $A\mathbf{x} = \mathbf{b}$, $A \in \mathbf{C}^{n\times n}$. Particolarmente importanti e di facile verifica sono le condizioni basate sulla proprietà di predominanza della matrice A (si vedano le definizioni 2.40).

5.12 Teorema. *Sia $A = M - N$ la decomposizione della matrice A corrispondente al metodo di Jacobi (cioè $M = D$ e $N = B + C$) o al metodo di Gauss-Seidel (cioè $M = D - B$ e $N = C$). Se vale una delle seguenti ipotesi:*

a) la matrice A è a predominanza diagonale in senso stretto,

b) la matrice A è a predominanza diagonale ed è irriducibile,

c) la matrice A è a predominanza diagonale in senso stretto per colonne,

d) la matrice A è a predominanza diagonale per colonne ed è irriducibile,

allora $\rho(M^{-1}N) < 1$ e quindi il metodo di Jacobi e il metodo di Gauss-Seidel sono convergenti.

Dim. Nelle ipotesi fatte, gli elementi principali di A sono non nulli e quindi la matrice M è non singolare. Un numero complesso λ è autovalore di $M^{-1}N$ se e solo se

$$\det(M^{-1}N - \lambda I) = 0, \tag{26}$$

ed essendo $M^{-1}N - \lambda I = -M^{-1}(\lambda M - N)$, per la regola di Binet dalla (26) segue che λ è autovalore di $M^{-1}N$ se e solo se

$$\det(\lambda M - N) = 0. \tag{27}$$

La matrice $H = \lambda M - N$ ha gli elementi

$$h_{ij} = \begin{cases} \lambda a_{ij} & \text{se } i = j \\ a_{ij} & \text{se } i \neq j \end{cases} \quad \text{per il metodo di Jacobi,}$$

$$h_{ij} = \begin{cases} \lambda a_{ij} & \text{se } i \geq j \\ a_{ij} & \text{se } i < j \end{cases} \quad \text{per il metodo di Gauss-Seidel.}$$

Se $|\lambda| \geq 1$ si ha

$$|h_{ii}| = |\lambda|\ |a_{ii}| \quad \text{e} \quad |h_{ij}| \leq |\lambda|\ |a_{ij}| \quad \text{per} \quad i \neq j,$$

e quindi la matrice H ha le proprietà *a), b) c)* o *d)* della matrice A. In tal caso, per il teorema 2.41 la matrice H è non singolare e quindi un numero λ, tale che $|\lambda| \geq 1$, non può verificare la (27), cioè non può essere autovalore di $M^{-1}N$. Ne segue che gli autovalori di $M^{-1}N$ hanno modulo minore di 1, e per il teorema 5.6 i metodi di Jacobi e di Gauss-Seidel sono convergenti. ∎

5.13 Esempi. a) La matrice

$$A = \begin{bmatrix} \text{-}4 & \text{-}1 & 1 & 1 \\ 0 & \text{-}4 & \text{-}1 & 1 \\ \text{-}1 & \text{-}1 & 4 & 1 \\ 1 & \text{-}1 & 0 & 4 \end{bmatrix}$$

ha predominanza diagonale in senso stretto, sia per righe che per colonne. Per il teorema 5.12 le matrici di iterazione di Jacobi e di Gauss-Seidel hanno entrambe raggio spettrale minore di 1. È infatti

$$J = \frac{1}{4} \begin{bmatrix} 0 & \text{-}1 & 1 & 1 \\ 0 & 0 & \text{-}1 & 1 \\ 1 & 1 & 0 & \text{-}1 \\ \text{-}1 & 1 & 0 & 0 \end{bmatrix}, \qquad G = \frac{1}{16} \begin{bmatrix} 0 & \text{-}4 & 4 & 4 \\ 0 & 0 & \text{-}4 & 4 \\ 0 & \text{-}1 & 0 & \text{-}2 \\ 0 & 1 & \text{-}2 & 0 \end{bmatrix}.$$

Lo spettro degli autovalori di J è dato dall'unione degli spettri delle matrici (si veda l'esercizio 2.26)

$$\frac{1}{4} \left(\begin{bmatrix} 0 & \text{-}1 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 1 & 1 \\ \text{-}1 & 1 \end{bmatrix} \right) \quad \text{e} \quad \frac{1}{4} \left(\begin{bmatrix} 0 & \text{-}1 \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} 1 & 1 \\ \text{-}1 & 1 \end{bmatrix} \right).$$

Gli autovalori di J risultano

$$\lambda_1 = \lambda_2 = \frac{1}{4}, \quad \lambda_3 = \frac{-1 + \mathbf{i}\sqrt{2}}{4}, \quad \lambda_4 = \frac{-1 - \mathbf{i}\sqrt{2}}{4},$$

quindi $\rho(J) = \dfrac{\sqrt{3}}{4}$. Il polinomio caratteristico della G è

$$p(\lambda) = \lambda^4 - \frac{3}{64}\lambda^2 - \frac{\lambda}{256} = \lambda\Big(\lambda - \frac{1}{4}\Big)\Big(\lambda + \frac{1}{8}\Big)^2,$$

per cui gli autovalori di G sono

$$\lambda_1 = 0, \quad \lambda_2 = \frac{1}{4}, \quad \lambda_3 = \lambda_4 = -\frac{1}{8}$$

e il raggio spettrale è $\rho(G) = \dfrac{1}{4}$.

b) La matrice

$$A = \begin{bmatrix} \text{-}4 & \text{-}1 & 1 & 1 \\ 0 & \text{-}4 & \text{-}1 & \text{-}3 \\ \text{-}1 & \text{-}1 & 4 & 1 \\ 1 & 3 & 0 & 4 \end{bmatrix}$$

ha predominanza diagonale ed è irriducibile. Per il teorema 5.12 le matrici di iterazione di Jacobi e di Gauss-Seidel hanno entrambe raggio spettrale minore di 1. È infatti

$$J = \frac{1}{4}\begin{bmatrix} 0 & \text{-}1 & 1 & 1 \\ 0 & 0 & \text{-}1 & \text{-}3 \\ 1 & 1 & 0 & \text{-}1 \\ \text{-}1 & \text{-}3 & 0 & 0 \end{bmatrix}, \qquad G = \frac{1}{16}\begin{bmatrix} 0 & \text{-}4 & 4 & 4 \\ 0 & 0 & \text{-}4 & \text{-}12 \\ 0 & \text{-}1 & 0 & \text{-}6 \\ 0 & 1 & 2 & 8 \end{bmatrix}.$$

Procedendo come nel caso a), si trova che gli autovalori di J risultano

$$\lambda_1 = \frac{1}{4}, \quad \lambda_2 = -\frac{3}{4}, \quad \lambda_3 = \frac{1+\sqrt{2}}{4}, \quad \lambda_4 = \frac{1-\sqrt{2}}{4},$$

da cui $\rho(J) = \dfrac{3}{4}$, e che gli autovalori di G sono

$$\lambda_1 = 0, \quad \lambda_2 = \frac{1}{4}, \quad \lambda_3 = \lambda_4 = \frac{1}{8},$$

da cui $\rho(G) = \dfrac{1}{4}$.

c) Si noti che la sola condizione di predominanza diagonale non è sufficiente per la convergenza. Si consideri infatti la matrice

$$A = \begin{bmatrix} \text{-}4 & \text{-}1 & 1 & 1 \\ 0 & \text{-}4 & 0 & \text{-}4 \\ 1 & 1 & 4 & 1 \\ 0 & \text{-}4 & 0 & 4 \end{bmatrix}$$

che ha predominanza diagonale ma è riducibile. Le due matrici di iterazione sono

$$J = \frac{1}{4}\begin{bmatrix} 0 & -1 & 1 & 1 \\ 0 & 0 & 0 & -4 \\ -1 & -1 & 0 & -1 \\ 0 & 4 & 0 & 0 \end{bmatrix}, \qquad G = \frac{1}{16}\begin{bmatrix} 0 & -4 & 4 & 4 \\ 0 & 0 & 0 & -16 \\ 0 & 1 & -1 & -1 \\ 0 & 0 & 0 & -16 \end{bmatrix}.$$

Il polinomio caratteristico della J è

$$p(\lambda) = \lambda^4 + \frac{17}{16}\lambda^2 + \frac{1}{16} = (\lambda^2 + 1)\left(\lambda^2 + \frac{1}{16}\right),$$

per cui gli autovalori di J sono $\lambda_1 = \mathbf{i}$, $\lambda_2 = -\mathbf{i}$, $\lambda_3 = \dfrac{\mathbf{i}}{4}$, $\lambda_4 = -\dfrac{\mathbf{i}}{4}$ e quindi $\rho(J) = 1$. Il polinomio caratteristico della G è

$$p(\lambda) = \lambda^4 + \frac{17}{16}\lambda^3 + \frac{\lambda^2}{16} = \lambda^2\left(\lambda^2 + \frac{17}{16}\lambda + \frac{1}{16}\right),$$

per cui gli autovalori di G sono $\lambda_1 = \lambda_2 = 0$, $\lambda_3 = -\dfrac{1}{16}$, $\lambda_4 = -1$, e quindi $\rho(G) = 1$. In questo caso né il metodo di Jacobi né quello di Gauss-Seidel convergono. ∎

5.14 Teorema. *Sia A una matrice hermitiana non singolare con elementi principali reali e positivi. Allora il metodo di Gauss-Seidel è convergente se e solo se A è definita positiva.*

Dim. Essendo la matrice A hermitiana, è $C = B^H$ e quindi

$$A = D - B - B^H,$$

e la matrice di iterazione del metodo di Gauss-Seidel risulta

$$G = (D - B)^{-1}B^H = I - (D - B)^{-1}A. \tag{28}$$

Per dimostrare che il metodo è convergente, conviene prima dimostrare che la matrice $A - G^H AG$ è definita positiva. Posto per semplicità

$$F = (D - B)^{-1}A,$$

dalla (28) si ha $G = I - F$ e

$$\begin{aligned} A - G^H AG &= A - (I - F)^H A(I - F) = A - A + F^H A + AF - F^H AF \\ &= F^H(AF^{-1} + F^{-H}A - A)F = F^H(D - B + D - B^H - A)F \\ &= F^H DF. \end{aligned}$$

La matrice F è non singolare perché tali sono le due matrici $(D-B)^{-1}$ e A, ed essendo gli elementi di D positivi, la matrice $A - G^H AG$ risulta definita positiva. Infatti per ogni $\mathbf{x} \neq \mathbf{0}$ risulta

$$\mathbf{x}^H A\mathbf{x} - \mathbf{x}^H G^H AG\mathbf{x} = \mathbf{x}^H F^H DF\mathbf{x} > 0. \tag{29}$$

Si supponga ora che la matrice A sia definita positiva e si consideri un autovalore λ di G e un corrispondente autovettore $\mathbf{x}$. Dalla (29) si ha:

$$\mathbf{x}^H A\mathbf{x} - \lambda\overline{\lambda}\,\mathbf{x}^H A\mathbf{x} > 0,$$

e cioè

$$(1 - |\lambda|^2)\,\mathbf{x}^H A\mathbf{x} > 0. \tag{30}$$

Essendo A definita positiva, dalla (30) risulta $|\lambda| < 1$ e quindi, per il teorema 5.6, il metodo di Gauss-Seidel è convergente.

Viceversa, si supponga che il metodo sia convergente e si consideri il vettore $\mathbf{e}^{(k)} = \mathbf{x}^* - \mathbf{x}^{(k)}$. Per la (12) si ha che

$$\mathbf{e}^{(k)} = G\mathbf{e}^{(k-1)},$$

e, sostituendo nella (30) $\mathbf{e}^{(k-1)}$ al posto di $\mathbf{x}$, poiché la matrice $A - G^H AG$ è definita positiva, risulta:

$$[\mathbf{e}^{(k-1)}]^H A\mathbf{e}^{(k-1)} > [\mathbf{e}^{(k)}]^H A\mathbf{e}^{(k)}. \tag{31}$$

Se A non fosse definita positiva, allora esisterebbe un vettore $\mathbf{e}^{(0)} \neq \mathbf{0}$ per cui $[\mathbf{e}^{(0)}]^H A\mathbf{e}^{(0)} \leq 0$ e quindi la successione $[\mathbf{e}^{(k)}]^H A\mathbf{e}^{(k)}$, che per la (31) è monotona decrescente, non potrebbe convergere a zero, ciò che è assurdo perché il metodo di Gauss-Seidel è convergente, cioè

$$\lim_{k\to\infty} \mathbf{e}^{(k)} = \mathbf{0}. \qquad \blacksquare$$

Nella figura 5.5 sono sinteticamente rappresentate le classi delle matrici hermitiane, delle matrici definite positive e delle matrici con predominanza diagonale in senso stretto e la classe della matrici per cui il metodo di Gauss-Seidel è convergente.

Si può dimostrare (si veda l'esercizio 5.15) che per le matrici a predominanza diagonale in senso stretto vale la relazione $\|G\|_\infty \leq \|J\|_\infty < 1$. Però, anche se $\rho(G) \leq \|G\|_\infty$ e $\rho(J) \leq \|J\|_\infty$, non sempre ne segue che $\rho(G) \leq \rho(J)$, cioè non sempre per le matrici a predominanza diagonale in senso stretto il metodo di Gauss-Seidel è asintoticamente più veloce del metodo di Jacobi.

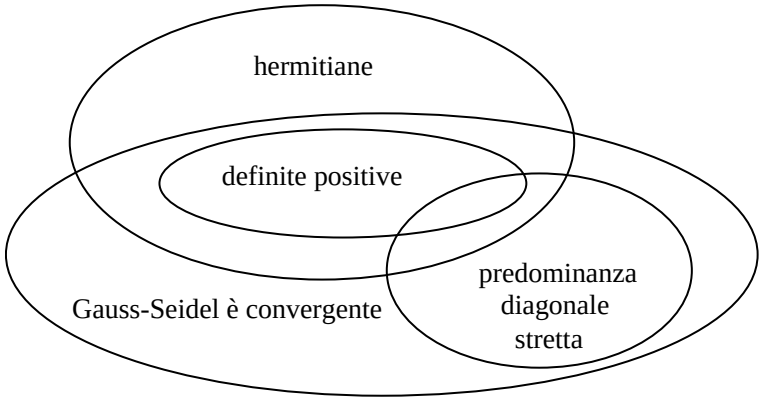


Fig. 5.5 - Classi di matrici per cui il metodo di Gauss-Seidel è convergente.

5.15 Esempio. Per il sistema $A\mathbf{x} = \mathbf{b}$, dove

$$A = \begin{bmatrix} 11 & -5 & -5 \\ 5 & 12 & 6 \\ 6 & -4 & 11 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 1 \\ 23 \\ 13 \end{bmatrix},$$

che ha la soluzione $\mathbf{x}^* = [1,\ 1,\ 1]^T$, risulta

$$J = \begin{bmatrix} 0 & \frac{5}{11} & \frac{5}{11} \\ -\frac{5}{12} & 0 & -\frac{1}{2} \\ -\frac{6}{11} & \frac{4}{11} & 0 \end{bmatrix}, \quad G = \begin{bmatrix} 0 & \frac{5}{11} & \frac{5}{11} \\ 0 & -\frac{25}{132} & -\frac{91}{132} \\ 0 & -\frac{115}{363} & -\frac{181}{363} \end{bmatrix}$$

e $\rho(J) = 0.7917518$, $\rho(G) = 0.8362568$, $\|J\|_\infty = 0.9166667$, $\|G\|_\infty = 0.9090909$. Quindi $\rho(J) < \rho(G)$, mentre $\|G\|_\infty \leq \|J\|_\infty$. Il tasso asintotico di convergenza del metodo di Jacobi è maggiore di quello del metodo di Gauss-Seidel. Assumendo $\mathbf{x}^{(0)} = \mathbf{0}$, e usando il criterio di arresto espresso dalla (16) in norma ∞ con $\epsilon = 10^{-5}$, la successione ottenuta con il metodo di Jacobi si arresta alla 52-esima iterazione, mentre la successione ottenuta con il metodo di Gauss-Seidel si arresta alla 68-esima iterazione. ∎

Il seguente teorema individua un'ampia classe di matrici per cui è possibile stabilire una relazione più precisa fra le velocità di convergenza dei metodi di Gauss-Seidel e di Jacobi.

5.16 Teorema (di Stein-Rosenberg). *Sia $A \in \mathbf{R}^{n\times n}$. Se gli elementi principali di A sono non nulli e gli elementi della matrice di iterazione di Jacobi J sono non negativi, allora vale una e una sola delle seguenti relazioni:*

a) $\rho(G) = \rho(J) = 0;$

b) $\rho(G) < \rho(J) < 1;$

c) $\rho(G) = \rho(J) = 1;$

d) $\rho(G) > \rho(J) > 1;$

(Per la dimostrazione si veda [10]) . ∎

5.17 Esempi. a) Per la matrice

$$A = \begin{bmatrix} 6 & 0 & 0 \\ -7 & 9 & 0 \\ -4 & -1 & 8 \end{bmatrix},$$

si ha

$$J = \begin{bmatrix} 0 & 0 & 0 \\ \frac{7}{9} & 0 & 0 \\ \frac{1}{2} & \frac{1}{8} & 0 \end{bmatrix}, \quad G = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

e $\rho(J) = \rho(G) = 0$.

b) Per la matrice

$$A = \begin{bmatrix} 9 & -3 & -1 \\ -2 & 9 & 0 \\ -2 & 0 & 9 \end{bmatrix},$$

si ha

$$J = \begin{bmatrix} 0 & \frac{1}{3} & \frac{1}{9} \\ \frac{2}{9} & 0 & 0 \\ \frac{2}{9} & 0 & 0 \end{bmatrix}, \quad G = \begin{bmatrix} 0 & \frac{1}{3} & \frac{1}{9} \\ 0 & \frac{2}{27} & \frac{2}{81} \\ 0 & \frac{2}{27} & \frac{2}{81} \end{bmatrix},$$

e $\rho(J) = 0.3142697,\ \rho(G) = 0.09876543$.

c) Per la matrice

$$A = \begin{bmatrix} 1 & 0 & 0 \\ -8 & 1 & -2 \\ -6 & -3 & 6 \end{bmatrix},$$

si ha

$$J = \begin{bmatrix} 0 & 0 & 0 \\ 8 & 0 & 2 \\ 1 & \frac{1}{2} & 0 \end{bmatrix}, \quad G = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 2 \\ 0 & 0 & 1 \end{bmatrix},$$

e $\rho(J) = \rho(G) = 1$.

d) Per la matrice

$$A = \begin{bmatrix} 8 & -6 & -8 \\ -6 & 7 & 0 \\ 0 & -8 & 7 \end{bmatrix},$$

si ha

$$J = \begin{bmatrix} 0 & \frac{3}{4} & 1 \\ \frac{6}{7} & 0 & 0 \\ 0 & \frac{8}{7} & 0 \end{bmatrix}, \quad G = \begin{bmatrix} 0 & \frac{3}{4} & 1 \\ 0 & \frac{9}{14} & \frac{6}{7} \\ 0 & \frac{36}{49} & \frac{48}{49} \end{bmatrix},$$

e $\rho(J) = 1.206222,\ \rho(G) = 1.6224490$. ∎

Molte delle matrici che si ottengono risolvendo numericamente problemi differenziali di tipo ellittico hanno predominanza diagonale e soddisfano alle condizioni del teorema di Stein-Rosenberg: in tal caso è conveniente usare il metodo di Gauss-Seidel. Nel caso delle matrici tridiagonali è possibile stabilire esattamente di quanto il metodo di Gauss-Seidel è più veloce del metodo di Jacobi.

5.18 Teorema. *Sia $A \in \mathbf{C}^{n\times n}$ la matrice tridiagonale*

$$A = \begin{bmatrix} a_1 & c_1 & & & \\ b_1 & a_2 & c_2 & & \\ & b_2 & a_3 & \ddots & \\ & & \ddots & \ddots & c_{n-1} \\ & & & b_{n-1} & a_n \end{bmatrix},$$

in cui $a_i \neq 0$ per $i = 1, \ldots, n$. Valgono le seguenti relazioni:

a) se μ è autovalore di J, allora μ^2 è autovalore di G;

b) se λ è autovalore non nullo di G, allora le radici quadrate di λ sono autovalori di J.

Dim. Sia $S \in \mathbf{C}^{n\times n}$ la matrice diagonale

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

in cui $\alpha \in \mathbf{C}$ è una costante non nulla. Si ha

$$SJS^{-1} = \begin{bmatrix} 0 & -\frac{c_1}{\alpha a_1} & & & \\ -\alpha\,\frac{b_1}{a_2} & 0 & -\frac{c_2}{\alpha a_2} & & \\ & -\alpha\,\frac{b_2}{a_3} & 0 & \ddots & \\ & & \ddots & \ddots & -\frac{c_{n-1}}{\alpha a_{n-1}} \\ & & & -\alpha\,\frac{b_{n-1}}{a_n} & 0 \end{bmatrix}$$

$$= \alpha\, D^{-1}B + \frac{1}{\alpha}\, D^{-1}C,$$

e quindi le matrici J e $\alpha\ D^{-1}B + \frac{1}{\alpha}\ D^{-1}C$ hanno lo stesso polinomio caratteristico qualunque sia $\alpha \neq 0$, cioè se μ è autovalore di J allora

$$\det(\alpha^2\ D^{-1}B + D^{-1}C - \alpha\mu I) = 0, \tag{32}$$

per ogni $\alpha \neq 0$ e viceversa, se esiste un $\alpha \neq 0$ per cui μ soddisfa la (32), allora μ è autovalore di J. Si ha

$$\begin{aligned} G - \lambda I &= (D-B)^{-1}C - \lambda I = (D-B)^{-1}[C - \lambda(D-B)] \\ &= (I - D^{-1}B)^{-1}(\lambda D^{-1}B + D^{-1}C - \lambda I), \end{aligned}$$

e quindi, se μ è autovalore di J, posto $\lambda = \alpha\mu$ e $\alpha^2 = \lambda$, dalla (32) segue che $\det(G - \lambda I) = 0$, per cui i λ tali che $\mu^2 = \lambda$ sono autovalori di G. Viceversa, se $\lambda \neq 0$ è un autovalore di G, siano $\alpha \neq 0$ e μ tali che $\alpha^2 = \lambda$ e $\alpha\mu = \lambda$. Allora

$$0 = \det(\lambda D^{-1}B + D^{-1}C - \lambda I) = \det(\alpha^2 D^{-1}B + D^{-1}C - \alpha\mu I),$$

e per la (32) μ è autovalore di J. ∎

Quindi per le matrici tridiagonali il metodo di Gauss-Seidel è convergente se e solo se lo è il metodo di Jacobi e vale

$$\rho(G) = \rho^2(J).$$

Perciò il tasso asintotico di convergenza del metodo di Gauss-Seidel è doppio di quello del metodo di Jacobi e, asintoticamente, sono necessarie metà iterazioni del metodo di Gauss-Seidel per ottenere la stessa precisione che con il metodo di Jacobi.

5.19 Esempio. Sia $A \in \mathbf{R}^{6\times 6}$, la matrice tridiagonale

$$a_{ij} = \begin{cases} 2 & \text{per } i = j, \\ \text{-1} & \text{per } |i-j| = 1, \\ 0 & \text{altrimenti.} \end{cases}$$

Essendo A simmetrica, si ha

$$A = 2I - U - U^T,$$

dove U è la matrice

$$u_{ij} = \begin{cases} 1 & \text{per } j = i-1, \\ 0 & \text{altrimenti,} \end{cases}$$

e quindi

$$J = \frac{1}{2}(U + U^T) \quad \text{e} \quad G = (2I - U)^{-1}U^T = \left[\sum_{i=0}^{5} \left(\frac{1}{2}\right)^{i+1} U^i\right] U^T$$

(si veda l'esercizio 1.52). I rispettivi polinomi caratteristici sono dati da

$$p_J(\mu) = \mu^6 - \frac{5}{4}\mu^4 + \frac{3}{8}\mu^2 - \frac{1}{64}$$
$$p_G(\lambda) = \lambda^3 \left(\lambda^3 - \frac{5}{4}\lambda^2 + \frac{3}{8}\lambda - \frac{1}{64}\right),$$

da cui si ricavano gli autovalori (dall'esercizio 2.40 si ha che gli autovalori di J sono dati da $\pm\cos\frac{\pi}{7}$, $\pm\cos\frac{2\pi}{7}$, $\pm\cos\frac{3\pi}{7}$)

$$\mu_1 = -\mu_6 = 0.9009688,\ \mu_2 = -\mu_5 = 0.6234898,\ \mu_3 = -\mu_4 = 0.2225209,$$
$$\lambda_1 = \lambda_2 = \lambda_3 = 0,$$
$$\lambda_4 = \mu_1^2 = 0.8117447,\ \lambda_5 = \mu_2^2 = 0.3887395,\ \lambda_6 = \mu_3^2 = 0.04951555.$$

Risulta pertanto che

$$\rho(J) = 0.9009688 \quad \text{e} \quad \rho(G) = \rho^2(J) = 0.8117447. \qquad \blacksquare$$

5. Metodi di Jacobi e di Gauss-Seidel a blocchi

Nel trattamento numerico delle equazioni differenziali intervengono spesso matrici a blocchi. In tal caso risulta naturale estendere i metodi iterativi di Jacobi e di Gauss-Seidel in termini di blocchi.

Sia allora A una matrice $n \times n$ a blocchi $A_{ij} \in \mathbf{C}^{m\times m}$, $i, j = 1, 2, \ldots, n$, tale che i blocchi diagonali A_{ii} siano matrici quadrate non singolari. Partizionando a blocchi i vettori $\mathbf{b}, \mathbf{x}^{(k-1)}$ e $\mathbf{x}^{(k)}$ compatibilmente con la partizione di A:

$$\mathbf{b} = \begin{bmatrix} \mathbf{b}_1 \\ \vdots \\ \mathbf{b}_n \end{bmatrix}, \quad \mathbf{x}^{(k-1)} = \begin{bmatrix} \mathbf{x}_1^{(k-1)} \\ \vdots \\ \mathbf{x}_n^{(k-1)} \end{bmatrix}, \quad \mathbf{x}^{(k)} = \begin{bmatrix} \mathbf{x}_1^{(k)} \\ \vdots \\ \mathbf{x}_n^{(k)} \end{bmatrix},$$

e procedendo come nel caso delle matrici ad elementi scalari, si ha per il metodo di Jacobi

$$\mathbf{x}_i^{(k)} = A_{ii}^{-1}\left[\mathbf{b}_i - \sum_{\substack{j=1 \\ j\neq i}}^{n} A_{ij}\mathbf{x}_j^{(k-1)}\right], \quad i = 1, 2, \ldots, n,$$

e per il metodo di Gauss-Seidel

$$\mathbf{x}_i^{(k)} = A_{ii}^{-1}\left[\mathbf{b}_i - \sum_{j=1}^{i-1} A_{ij}\mathbf{x}_j^{(k)} - \sum_{j=i+1}^{n} A_{ij}\mathbf{x}_j^{(k-1)}\right], \quad i = 1, 2, \ldots, n.$$

Non è facile estendere i criteri di convergenza visti nel paragrafo precedente al caso a blocchi, né si può dire in generale che se il metodo iterativo, applicato scalarmente, è convergente, allora anche il metodo a blocchi è convergente. Vale però il seguente teorema, per la cui dimostrazione si rimanda a [10]. Si indicano con J_B e G_B rispettivamente le matrici di iterazione dei metodi di Jacobi e di Gauss-Seidel a blocchi, e con J e G le corrispondenti matrici dei metodi applicati scalarmente.

5.20 Teorema. *Sia A una matrice $n \times n$ a blocchi $A_{ij} \in \mathbf{R}^{m\times m}$, i, $j = 1, 2, \ldots, n$, tale che i blocchi diagonali A_{ii} siano matrici non singolari. Se $a_{ij} \leq 0$ per ogni $i \neq j$, ed inoltre $A^{-1} > O$, allora i metodi di Jacobi e di Gauss-Seidel applicati scalarmente e a blocchi sono convergenti e si ha*

$$\rho(G_B) < \rho(J_B) < 1,$$

$$\rho(G_B) \leq \rho(G) < 1,$$

$$\rho(J_B) \leq \rho(J) < 1,$$

dove il segno di uguaglianza vale nella seconda (terza) relazione solo se $G_B = G$ (rispettivamente $J_B = J$). ∎

Una classe particolarmente importante di matrici che soddisfano alle condizioni del teorema 5.20 è individuata dal seguente teorema.

5.21 Teorema. *Sia $A \in \mathbf{R}^{n\times n}$ tale che*

(1) A è a predominanza diagonale ed è irriducibile,

(2) $a_{ii} > 0$ per $i = 1, \ldots, n$ e $a_{ij} \leq 0$ per $i \neq j,\ i, j = 1, \ldots, n$.

Allora $A^{-1} > O$.

Dim. Sia $\alpha > 0$ tale che

$$\alpha > \max_{i=1,\ldots,n} a_{ii}.$$

Allora la matrice

$$B = I - \frac{1}{\alpha}A \geq O \tag{33}$$

ha gli elementi principali positivi e gli altri elementi non negativi. Poiché la matrice $\dfrac{1}{\alpha}A$ ha gli elementi principali compresi fra 0 e 1 ed è a predominanza diagonale, i suoi cerchi di Gerschgorin sono tutti interni ad un cerchio di centro 1 e raggio 1. Dalla (33) segue che $\rho(B) < 1$.

Si dimostra ora che per ogni vettore $\mathbf{e}_i$ della base canonica è $B^{n-1}\mathbf{e}_i > \mathbf{0}$, e quindi $B^{n-1} > O$. Infatti, per $i = 1, \ldots, n$, sia $\mathbf{y}_0 = \mathbf{e}_i$ e si considerino i vettori $\mathbf{y}_{k+1} = B\mathbf{y}_k$, $k = 0, \ldots, n-2$. Posto

$$B = D + (B - D),$$

dove D è la matrice diagonale i cui elementi principali coincidono con quelli di B, è $B - D \geq O$ e $D \geq O$ e non singolare. Quindi dalla relazione

$$\mathbf{y}_{k+1} = D\mathbf{y}_k + (B - D)\mathbf{y}_k$$

segue che il numero delle componenti nulle di $\mathbf{y}_{k+1}$ è minore o uguale al numero delle componenti nulle di $\mathbf{y}_k$. Inoltre se i due vettori avessero lo stesso numero r di componenti nulle, esisterebbe una matrice di permutazione Π tale che

$$\Pi\mathbf{y}_k = \begin{bmatrix} \mathbf{x}_k \\ \\ \mathbf{0} \end{bmatrix}, \quad \mathbf{x}_k \in \mathbf{R}^{n-r},\ \mathbf{x}_k > \mathbf{0},$$

e sarebbe

$$\begin{bmatrix} \mathbf{x}_{k+1} \\ \\ \mathbf{0} \end{bmatrix} = \Pi B \Pi^T \begin{bmatrix} \mathbf{x}_k \\ \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} B_{11} & B_{12} \\ \\ B_{21} & B_{22} \end{bmatrix} \begin{bmatrix} \mathbf{x}_k \\ \\ \mathbf{0} \end{bmatrix},$$

da cui $B_{21} = O$, essendo $\mathbf{x}_k > \mathbf{0}$, e ciò è assurdo perché la matrice B è irriducibile, per l'ipotesi di irriducibilità di A. Ne segue che $\mathbf{y}_{k+1}$ ha meno componenti nulle di $\mathbf{y}_k$ e quindi

$$\mathbf{y}_{n-1} = B^{n-1}\mathbf{y}_0 = B^{n-1}\mathbf{e}_i > \mathbf{0} \tag{34}$$

per ogni $i = 1, \ldots, n$.

Poiché $\rho(B) < 1$, per il teorema 5.4 è

$$(I - B)^{-1} = \sum_{k=0}^{\infty} B^k,$$

dove le matrici della sommatoria hanno elementi non negativi e per la (34) hanno elementi positivi per $k \geq n - 1$. Quindi $(I - B)^{-1} > O$, da cui

$$A^{-1} = \frac{1}{\alpha}\,(I - B)^{-1} > O. \qquad \blacksquare$$

Le matrici simmetriche che verificano le ipotesi del teorema 5.21 vengono dette *S-matrici* e per il teorema 2.41 sono definite positive. Sono S-matrici molte delle matrici che si ottengono risolvendo numericamente problemi differenziali di tipo ellittico. Per il teorema 5.20, se la matrice A è una S-matrice, risulta quindi conveniente utilizzare il metodo di Gauss-Seidel a blocchi.

Un risultato analogo al teorema 5.18 vale nel caso delle matrici tridiagonali a blocchi (per la dimostrazione si veda [10]).

5.22 Teorema. *Se A è una matrice $n \times n$ tridiagonale a blocchi $A_{ij} \in \mathbf{C}^{m\times m}$, $i, j = 1, 2, \ldots, n$, tale che i blocchi diagonali A_{ii} siano matrici quadrate non singolari, valgono le seguenti relazioni:*

a) se μ è autovalore di J_B, allora μ^2 è autovalore di G_B;

b) se λ è autovalore non nullo di G_B, allora le radici quadrate di λ sono autovalori di J_B. $\blacksquare$

Quindi per le matrici tridiagonali a blocchi il metodo di Gauss-Seidel a blocchi è convergente se e solo se lo è il metodo di Jacobi a blocchi e inoltre vale

$$\rho(G_B) = \rho^2(J_B).$$

5.23 Esempio. Sia A la matrice

$$A = \begin{bmatrix} B & -I \\ -I & B \end{bmatrix}, \quad \text{dove} \quad B = \begin{bmatrix} 4 & -1 \\ -1 & 4 \end{bmatrix},$$

cioè

$$A = \begin{bmatrix} 4 & -1 & -1 & 0 \\ -1 & 4 & 0 & -1 \\ -1 & 0 & 4 & -1 \\ 0 & -1 & -1 & 4 \end{bmatrix}.$$

Per il teorema 5.21 è $A^{-1} > O$, per cui A verifica le ipotesi del teorema 5.20. Risulta infatti

$$J = \frac{1}{4}\begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}, \quad G = \frac{1}{4}\begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & \frac{1}{4} & \frac{1}{4} & 1 \\ 0 & \frac{1}{4} & \frac{1}{4} & 1 \\ 0 & \frac{1}{8} & \frac{1}{8} & \frac{1}{2} \end{bmatrix},$$

$$J_B = \begin{bmatrix} O & H \\ H & O \end{bmatrix}, \; G_B = \begin{bmatrix} O & H \\ O & H^2 \end{bmatrix}, \; \text{dove} \quad H = B^{-1} = \frac{1}{15}\begin{bmatrix} 4 & 1 \\ 1 & 4 \end{bmatrix}.$$

Gli autovalori sono

per J: $\lambda_1 = \lambda_2 = 0, \; \lambda_3 = \frac{1}{2}, \; \lambda_4 = -\frac{1}{2}$;

per G: $\lambda_1 = \lambda_2 = \lambda_3 = 0, \; \lambda_4 = \frac{1}{4}$;

per J_B: $\lambda_1 = \frac{1}{5}, \; \lambda_2 = -\frac{1}{5}, \; \lambda_3 = \frac{1}{3}, \; \lambda_4 = -\frac{1}{3}$;

per G_B: $\lambda_1 = \lambda_2 = 0, \; \lambda_3 = \frac{1}{25}, \; \lambda_4 = \frac{1}{9}$;

quindi per i raggi spettrali si ha:

$$\rho(J) = \frac{1}{2}, \; \rho(G) = \rho^2(J) = \frac{1}{4}, \; \rho(J_B) = \frac{1}{3}, \; \rho(G_B) = \rho^2(J_B) = \frac{1}{9}. \qquad \blacksquare$$

6. Metodi di rilassamento

Nella risoluzione di alcuni problemi dell'analisi numerica l'introduzione di un parametro permette di migliorare l'efficienza dei metodi usati. Un esempio di questo tipo si incontra nel caso dei metodi iterativi, dove un'opportuna determinazione del parametro permette di ottenere una velocità di convergenza sostanzialmente maggiore di quella dei metodi di Jacobi e di Gauss-Seidel.

Si considera il sistema

$$\omega A\mathbf{x} = \omega\mathbf{b}, \quad \omega \neq 0,$$

equivalente al (6) e si effettua la seguente decomposizione della matrice ωA:

$$\omega A = M - N, \quad \text{dove} \quad M = D - \omega B, \quad N = (1-\omega)D + \omega C. \tag{35}$$

Se $\det M \neq 0$, si ricava il seguente metodo iterativo, detto *di rilassamento*

$$\mathbf{x}^{(k)} = (D - \omega B)^{-1}[(1-\omega)\ D + \omega C]\mathbf{x}^{(k-1)} + \omega(D - \omega B)^{-1}\mathbf{b}. \tag{36}$$

La matrice di iterazione di tale metodo è allora

$$H(\omega) = (D - \omega B)^{-1}[(1-\omega)\ D + \omega C]. \tag{37}$$

Dalla (36) si ottiene

$$\mathbf{x}^{(k)} = (1-\omega)\mathbf{x}^{(k-1)} + \omega D^{-1}[B\mathbf{x}^{(k)} + C\mathbf{x}^{(k-1)} + \mathbf{b}],$$

che in termini di componenti si scrive

$$x_i^{(k)} = (1-\omega)x_i^{(k-1)} + \frac{\omega}{a_{ii}}\left[b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k-1)}\right],$$
$$i = 1, 2, \ldots, n.$$

Si noti che questa espressione coincide con quella del metodo di Gauss-Seidel per $\omega = 1$, e che anche per $\omega \neq 1$ il numero di operazioni richieste per effettuare un'iterazione è, a meno di termini di ordine inferiore, lo stesso che per il metodo di Gauss-Seidel.

Il metodo di rilassamento viene in particolare detto di *sottorilassamento* se $\omega < 1$ e di *sovrarilassamento* se $\omega > 1$. Quest'ultimo è anche detto *metodo SOR*, dalle iniziali dei corrispondenti termini inglesi *Successive Over-Relaxation*.

Nella applicazione di un metodo di rilassamento è importante scegliere, se possibile, un valore ω_o del parametro ω che, oltre ad assicurare la convergenza, renda minimo il raggio spettrale della matrice di iterazione $H(\omega)$, in modo da ottenere la massima velocità di convergenza possibile.

5.24 Teorema (di Kahan). *Per la matrice di iterazione di un metodo di rilassamento, risulta*

$$\rho[H(\omega)] \geq |\omega - 1|.$$

Quindi condizione necessaria per la convergenza è che

$$|\omega - 1| < 1,$$

e se ω è reale, è che

$$0 < \omega < 2.$$

Dim. La matrice $D - \omega B$ è triangolare inferiore, e poiché la matrice B ha nulli gli elementi principali, ne segue che

$$\det(D - \omega B) = \det D.$$

Analogamente è

$$\det[(1-\omega)D + \omega C] = \det[(1-\omega)D] = (1-\omega)^n \det D,$$

e quindi per la (37) è

$$\det[H(\omega)] = (1-\omega)^n.$$

Poiché il determinante di una matrice è uguale al prodotto degli n autovalori, ne segue che

$$\rho[H(\omega)] \geq |\sqrt[n]{\det[H(\omega)]}| = |1-\omega|. \qquad \blacksquare$$

Il teorema di Kahan fornisce una condizione necessaria ma non sufficiente per la convergenza, come si può vedere anche dal seguente esempio.

5.25 Esempio. La figura 5.6 riporta il grafico di $\rho[H(\omega)]$, al variare di ω nell'intervallo [0,1.4] per la matrice

$$A = \begin{bmatrix} -3 & 3 & -6 \\ -4 & 7 & -8 \\ 5 & 7 & -9 \end{bmatrix}, \tag{38}$$

per la quale si è visto nell'esempio 5.11 b) che il metodo di Gauss-Seidel non è convergente. Scegliendo ω compreso fra 0.1 e 0.9 il metodo di rilassamento è convergente. Il valore di ω per cui $\rho[H(\omega)]$ è minimo è $\omega_o = 0.8864098$, a cui corrisponde per il raggio spettrale il valore $\rho[H(\omega_o)] = 0.6101880$.

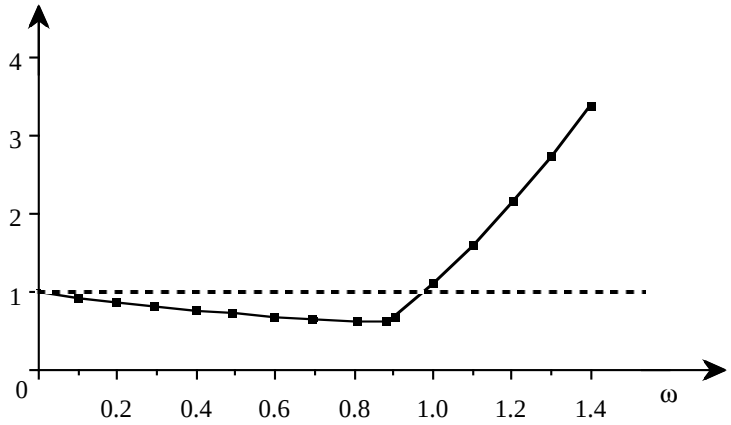


Fig. 5.6 - Grafico di $\rho[H(\omega)]$ per la matrice (38).

Applicando il metodo di rilassamento con tale valore di ω_o per risolvere il sistema $A\mathbf{x} = \mathbf{b}$, dove $\mathbf{b} = [-6, -5, 3]^T$, a partire dal punto iniziale $\mathbf{x}^{(0)} = \mathbf{0}$ e usando gli stessi criteri di arresto degli esempi 5.11, la successione si arresta alla 25-esima iterazione. ∎

La condizione $0 < \omega < 2$ del teorema di Kahan risulta anche sufficiente per la convergenza dei metodi di rilassamento se la matrice A è definita positiva. Vale infatti il seguente

5.26 Teorema (di Ostrowski-Reich). *Se A è definita positiva e ω è un numero reale tale che $0 < \omega < 2$, allora il metodo di rilassamento è convergente.*

Dim. Dalla (37), tenendo conto che $A = D - B - B^H$, si ha che

$$\begin{aligned} H(\omega) &= (D - \omega B)^{-1}[(1-\omega)D + \omega B^H] \\ &= (D - \omega B)^{-1}(D - \omega B - \omega A) \\ &= I - \omega\,(D - \omega B)^{-1}A \end{aligned}$$

Posto per semplicità

$$H(\omega) = I - F,$$

dove

$$F = \omega(D - \omega B)^{-1}A,$$

procedendo come nella dimostrazione del teorema 5.14 si ha

$$\begin{aligned} A - [H(\omega)]^H A H(\omega) &= F^H(AF^{-1} + F^{-H}A - A)F \\ &= F^H[\frac{2}{\omega}D - B - B^H - A]F \\ &= (\frac{2}{\omega} - 1)F^H DF, \end{aligned}$$

e quindi, poiché $0 < \omega < 2$, tale matrice è definita positiva. Indicato con λ un autovalore di $H(\omega)$ e con $\mathbf{x}$ il corrispondente autovettore si ha

$$\mathbf{x}^H A\mathbf{x} - \mathbf{x}^H [H(\omega)]^H A H(\omega)\mathbf{x} = \mathbf{x}^H A\mathbf{x} - \lambda\overline{\lambda}\mathbf{x}^H A\mathbf{x} = (1 - |\lambda|^2)\mathbf{x}^H A\mathbf{x} > 0,$$

ed essendo A definita positiva, ne segue che $|\lambda| < 1$. ∎

5.27 Esempio. La figura 5.7 riporta il grafico di $\rho[H(\omega)]$, al variare di ω nell'intervallo [0,2] per la matrice

$$A = \begin{bmatrix} 7 & 4 & -7 \\ 4 & 5 & -3 \\ -7 & -3 & 8 \end{bmatrix}, \tag{39}$$

definita positiva. Per il teorema 5.14 il metodo di Gauss-Seidel è convergente. In accordo con il teorema 5.26 per i valori di ω compresi fra 0 e 2, estremi esclusi, anche il metodo di rilassamento è convergente. Il valore di ω per cui $\rho[H(\omega)]$ è minimo è $\omega_o = 1.531281$, a cui corrisponde per il raggio spettrale il valore $\rho[H(\omega_o)] = 0.6614684$.

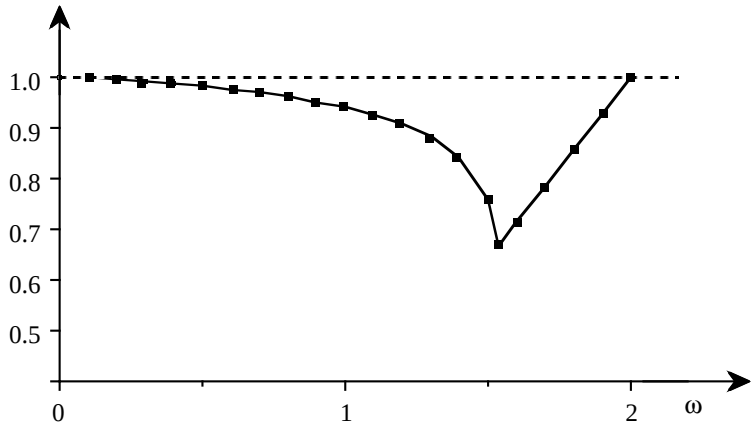


Fig. 5.7 - Grafico di $\rho[H(\omega)]$ per la matrice (39).

Si applica il metodo di rilassamento con tale valore di ω_o per risolvere il sistema lineare $A\mathbf{x} = \mathbf{b}$, dove $\mathbf{b} = [4, 6, -2]^T$, che ha la soluzione $\mathbf{x}^* = [1, 1, 1]^T$. A partire dal vettore iniziale $\mathbf{x}^{(0)} = \mathbf{0}$, usando la condizione di arresto (16) con $\epsilon = 10^{-5}$, la soluzione viene approssimata in 35 iterazioni. Con lo stesso vettore iniziale e con la stessa condizione di arresto il metodo di Gauss-Seidel richiede 119 iterazioni, mentre il metodo di Jacobi non risulta convergente. ∎

Nel seguente caso particolare, importante nelle applicazioni, è possibile dare una relazione esplicita fra gli autovalori della matrice di iterazione del metodo di Jacobi e quella del metodo di rilassamento e indicare il valore ω_o per cui si ha la massima velocità di convergenza e il corrispondente $\rho[H(\omega_o)]$.

5.28 Teorema. *Sia A la matrice tridiagonale del teorema 5.18 e sia $0 < \omega < 2$. Valgono le seguenti relazioni:*

a) se μ è autovalore di J, ogni λ tale che

$$(\lambda + \omega - 1)^2 = \lambda\omega^2\mu^2 \tag{40}$$

è autovalore di $H(\omega)$;

b) se λ è autovalore non nullo di $H(\omega)$, allora ogni μ per cui vale la (40) è autovalore di J;

c) se gli autovalori della matrice di iterazione J sono reali e tali che $\rho(J) < 1$, esiste uno e un solo valore ω_o per cui

$$\rho[H(\omega_o)] = \min_{0<\omega<2} \rho[H(\omega)],$$

ed è

$$\omega_o = \frac{2}{1 + \sqrt{1 - \rho^2(J)}}\ ; \tag{41}$$

d) per $0 < \omega \leq \omega_o$ è

$$\rho[H(\omega)] = 1 - \omega + \frac{1}{2}\omega^2\rho^2(J) + \omega\rho(J)\sqrt{1 - \omega + \frac{1}{4}\omega^2\rho^2(J)}$$

e per $\omega_o \leq \omega < 2$ è

$$\rho[H(\omega)] = \omega - 1,$$

e

$$\rho[H(\omega_o)] = \omega_o - 1 = \left[\frac{\rho(J)}{1 + \sqrt{1 - \rho^2(J)}}\right]^2.$$

Dim. Per quanto riguarda i punti a) e b), si procede come per la dimostrazione del teorema 5.18. Si ha

$$\begin{aligned} H(\omega) - \lambda I &= (D - \omega B)^{-1}[(1 - \omega)\ D + \omega C] - \lambda I \\ &= \omega(I - \omega D^{-1}B)^{-1}(\lambda D^{-1}B + D^{-1}C - \frac{\lambda + \omega - 1}{\omega}\ I). \end{aligned}$$

Se μ è autovalore di J, posto $\alpha^2 = \lambda$ e $\alpha\mu = \dfrac{\lambda + \omega - 1}{\omega}$, dalla (32) segue che $\det[H(\omega) - \lambda I] = 0$, e quindi i numeri λ per cui vale la (40) sono autovalori

di $H(\omega)$. Viceversa, se $\lambda \neq 0$ è autovalore di $H(\omega)$, siano $\alpha \neq 0$ e μ tali che $\alpha^2 = \lambda$ e $\alpha\mu = \dfrac{\lambda + \omega - 1}{\omega}$. Allora

$$0 = \det(\lambda D^{-1}B + D^{-1}C - \frac{\lambda + \omega - 1}{\omega}\, I) = \det(\alpha^2 D^{-1}B + D^{-1}C - \alpha\mu I),$$

e per la (32) μ è autovalore di J.

Per i punti c) e d) si consideri un autovalore μ di J: al variare di ω nell'intervallo (0,2) a μ, che per ipotesi è reale e tale che $|\mu| < 1$, corrispondono due autovalori λ_1 e λ_2 di $H(\omega)$ che soddisfano la (40), cioè soluzioni dell'equazione di secondo grado a coefficienti reali

$$\lambda^2 + [2(\omega - 1) - \omega^2\mu^2]\lambda + (\omega - 1)^2 = 0.$$

Si vuole dapprima determinare il valore di ω per cui è minima la funzione

$$m_\mu(\omega) = \max\, \{|\, \lambda_1|, |\lambda_2|\, \}.$$

Se $\mu = 0$ è $\lambda_1 = \lambda_2 = 1 - \omega$.
Se $\mu \neq 0$, posto

$$\Delta = \omega^2\mu^2\big(1 - \omega + \frac{1}{4}\omega^2\mu^2\big)$$

e

$$\omega_\mu = \frac{2}{1 + \sqrt{1 - \mu^2}}\ ,$$

risulta $\omega_\mu > 1$ e

$$\begin{cases} \text{per} \quad \omega > \omega_\mu \quad \text{è} \quad \Delta < 0,\ \lambda_1 \text{ e } \lambda_2 \text{ complessi e } |\lambda_1| = |\lambda_2| = \omega - 1, \\ \text{per} \quad \omega = \omega_\mu \quad \text{è} \quad \Delta = 0,\ \lambda_1 = \lambda_2 = \omega_\mu - 1, \\ \text{per} \quad \omega < \omega_\mu \quad \text{è} \quad \Delta > 0,\ \lambda_1 \text{ e } \lambda_2 \text{ reali.} \end{cases}$$

Poiché per $\omega < \omega_\mu$ è

$$1 - \omega + \frac{1}{2}\omega^2\mu^2 = \frac{\Delta}{\omega^2\mu^2} + \frac{\omega^2\mu^2}{4} > 0,$$

il massimo fra $|\lambda_1|$ e $|\lambda_2|$ è dato da

$$|\lambda_2| = \lambda_2 = 1 - \omega + \frac{1}{2}\omega^2\mu^2 + \sqrt{\Delta},$$

ed è una funzione decrescente di ω. Riassumendo si ha

$$m_\mu(\omega) = \begin{cases} |1-\omega| & \text{se } \mu = 0, \\ \left\{ \begin{array}{ll} \omega - 1 & \text{se } \omega > \omega_\mu \\ 1 - \omega + \frac{1}{2}\omega^2\mu^2 + \omega\mu\sqrt{1-\omega+\frac{1}{4}\omega^2\mu^2} & \text{se } \omega \le \omega_\mu \end{array} \right\} & \text{se } \mu \neq 0 \end{cases} \tag{42}$$

e il minimo della funzione $m_\mu(\omega)$ è dato da

$$m_\mu(\omega_\mu) = \omega_\mu - 1.$$

Nella figura 5.8 è riportato il grafico della funzione $m_\mu(\omega)$ per diversi valori di μ ($\mu = 0,\ 0.5,\ 0.8,\ 0.95$).

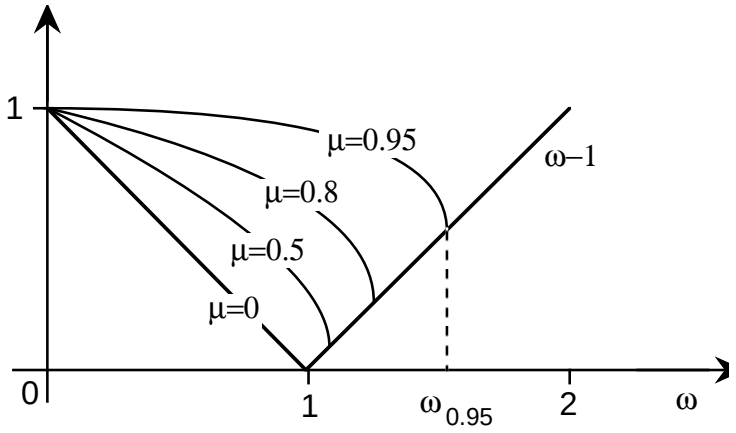


Fig. 5.8 - Grafico della funzione $m_\mu(\omega)$ per $\mu = 0,\ 0.5,\ 0.8,\ 0.95$.

Poiché per i punti a) e b) è

$$\rho[H(\omega)] = \max_\mu\ m_\mu(\omega),$$

dove μ varia sull'insieme degli autovalori di J, dalla (42) segue che

$$\rho[H(\omega)] = m_{\mu_o}(\omega), \quad \text{dove} \quad \mu_o = \max\ \mu,$$

come si vede anche dalla figura 5.8. Inoltre J per ipotesi ha autovalori reali e per il teorema 5.18 se μ è autovalore di J, anche $-\mu$ lo è, per cui $\mu_o = \rho(J)$. Ne segue che

$$\rho[H(\omega)] = m_{\rho(J)}(\omega) \quad \text{e} \quad \omega_o = \omega_{\rho(J)}. \qquad \blacksquare$$

$\rho[H(\omega)]$ ha in ω_o un punto di cuspide con tangente sinistra verticale e per $\omega > \omega_o$ ha un andamento rettilineo con coefficiente angolare 1. È quindi conveniente, se non si riesce a determinare un'approssimazione adeguata di $\rho(J)$, orientarsi verso una approssimazione per eccesso.

5.29 Esempio. La figura 5.9 riporta il grafico della funzione $\rho[H(\omega)]$ calcolata per alcuni valori di ω nell'intervallo (0,2) per la matrice $A \in \mathbf{R}^{6\times 6}$, tridiagonale e definita positiva, i cui elementi sono dati da

$$a_{ij} = \begin{cases} 2 & \text{per } i = j, \\ \text{-1} & \text{per } |i-j| = 1, \\ 0 & \text{altrimenti.} \end{cases}$$

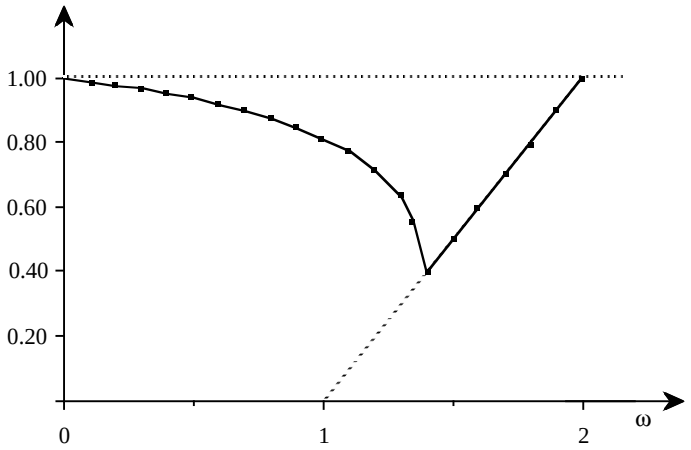


Fig. 5.9 - Grafico di $\rho[H(\omega)]$ per una matrice tridiagonale.

Per tale matrice, come risulta dall'esempio 5.19, è $\rho(J) = 0.9009688$. Per la (41) è $\omega_o = 1.394812$, a cui corrisponde per il raggio spettrale il valore $\rho[H(\omega_o)] = 0.3949117$, e poiché $\rho(G) = 0.8117447$, risulta

$$\rho^2(J) = \rho(G) \quad \text{e} \quad \rho^{4.5}(G) \approx \rho[H(\omega_o)].$$

Perciò, asintoticamente, per ottenere un risultato con la stessa precisione il metodo di Jacobi richiederebbe circa nove volte il numero di iterazioni richieste dal metodo di rilassamento con il parametro ottimo ω_o e il metodo di Gauss-Seidel circa quattro volte e mezzo. Poiché in pratica il calcolo viene interrotto dopo un numero finito di iterazioni, il numero di iterazioni effettivamente richieste per ottenere una determinata precisione può differire dai valori asintotici.

Per il sistema lineare $A\mathbf{x} = \mathbf{b}$, dove $\mathbf{b} = [1, 0, 0, 0, 0, 1]^T$, che ha come soluzione il vettore $\mathbf{x}^* = [1, 1, 1, 1, 1, 1]^T$, sono stati utilizzati i metodi di

Jacobi, Gauss-Seidel e di rilassamento con il parametro ottimo ω_o, a partire dallo stesso vettore iniziale $\mathbf{x}^{(0)} = \mathbf{0}$. Imponendo come condizione di arresto la (16) con $\epsilon = 10^{-5}$, la soluzione viene approssimata in 92 iterazioni dal metodo di Jacobi, in 50 iterazioni dal metodo di Gauss-Seidel, in 18 iterazioni dal metodo di rilassamento. La figura 5.10 riporta, al variare dell'indice k di iterazione, i grafici degli errori $\|\mathbf{e}^{(k)}\|_\infty$, generati dal metodo di Jacobi (quadratini vuoti), dal metodo di Gauss-Seidel (triangolini), e dal metodo di rilassamento (quadratini pieni). ∎

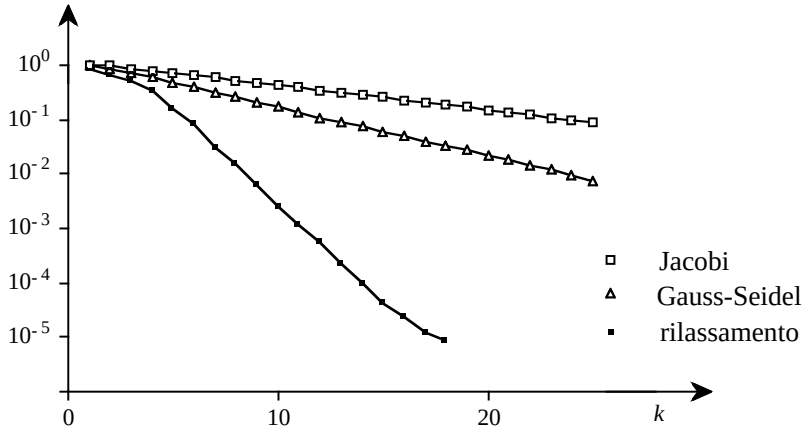


Fig. 5.10 - Andamento degli errori dei metodi iterativi di Jacobi, Gauss-Seidel e rilassamento.

Anche il metodo di rilassamento può essere esteso, in modo analogo a quanto fatto per i metodi di Jacobi e di Gauss-Seidel, al caso di matrici a blocchi. Si esamina in particolare il caso delle matrici tridiagonali a blocchi.

5.30 Teorema. *Sia A una matrice tridiagonale a blocchi, i cui blocchi diagonali siano quadrati e non singolari, e sia $0 < \omega < 2$. Indicate con J_B la matrice di iterazione del metodo di Jacobi a blocchi e con $H_B(\omega)$ la matrice di iterazione del metodo di rilassamento a blocchi, gli autovalori μ di J_B e gli autovalori non nulli λ di $H_B(\omega)$ verificano ancora la relazione (40). Inoltre se gli autovalori di J_B sono reali e tali che $\rho(J_B) < 1$, allora il valore ottimo ω_o del parametro del metodo di rilassamento a blocchi è dato da*

$$\omega_o = \frac{2}{1 + \sqrt{1 - \rho^2(J_B)}} \; ; \tag{43}$$

e

$$\rho[H_B(\omega_o)] = \omega_o - 1 = \left[\frac{\rho(J_B)}{1 + \sqrt{1 - \rho^2(J_B)}} \right]^2 .$$

Per la dimostrazione si veda [10]. ∎

Poiché $\rho[H_B(\omega_o)]$ è una funzione crescente di $\rho(J_B)$, da un confronto fra la (41) e la (43) risulta che nei casi in cui $\rho(J_B) < \rho(J)$, anche il metodo di rilassamento a blocchi ha un tasso asintotico di convergenza maggiore dello stesso metodo applicato scalarmente. Questo è vero in particolare nel caso delle S-matrici che soddisfano alle ipotesi del teorema 5.20.

5.31 Esempio. Sia A la matrice $n \times n$ a blocchi

$$A = \begin{bmatrix} B & -I & & \\ -I & B & \ddots & \\ & \ddots & \ddots & -I \\ & & -I & B \end{bmatrix}, \quad \text{dove} \quad B = \begin{bmatrix} 4 & -1 & & \\ -1 & 4 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 4 \end{bmatrix} \in \mathbf{R}^{n\times n}.$$

Se si applicano scalarmente i metodi iterativi di Jacobi, Gauss-Seidel e rilassamento a elementi, la matrice J è data da

$$J = \frac{1}{4}\begin{bmatrix} H & I & & \\ I & H & \ddots & \\ & \ddots & \ddots & I \\ & & I & H \end{bmatrix}, \quad \text{dove} \quad H = \begin{bmatrix} 0 & 1 & & \\ 1 & 0 & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & 0 \end{bmatrix} \in \mathbf{R}^{n\times n}.$$

Poiché $J = \frac{1}{4}(H \otimes I + I \otimes H)$ (si vedano gli esercizi 1.60 e 2.41) e gli autovalori di H sono (si veda l'esercizio 2.40)

$$\alpha_k = 2\ \cos\frac{k\pi}{n+1}, \quad k = 1, \ldots, n,$$

gli autovalori di J sono

$$\beta_{ij} = \alpha_i + \alpha_j, \quad i,\ j = 1, \ldots, n,$$

per cui risulta

$$\rho(J) = \cos\frac{\pi}{n+1}.$$

Gli autovalori della matrice G sono i quadrati di quelli di J (si veda l'esercizio 5.21), per cui

$$\rho(G) = \rho^2(J) = \cos^2\frac{\pi}{n+1}.$$

Per il metodo di rilassamento si ha (si veda l'esercizio 5.21)

$$\omega_o = \frac{2}{1+\sqrt{1-\rho^2(J)}} = \frac{2}{1+\sin\dfrac{\pi}{n+1}}$$

e

$$\rho[H(\omega_o)] = \omega_o - 1 = \frac{1 - \sin\dfrac{\pi}{n+1}}{1 + \sin\dfrac{\pi}{n+1}}.$$

Se si applicano a blocchi i metodi iterativi di Jacobi, Gauss-Seidel e rilassamento, la matrice J_B è data da

$$J_B = \begin{bmatrix} O & B^{-1} & & \\ B^{-1} & O & \ddots & \\ & \ddots & \ddots & B^{-1} \\ & & B^{-1} & O \end{bmatrix} = H \otimes B^{-1}.$$

Poiché gli autovalori di B^{-1} sono (si veda l'esercizio 2.40)

$$\gamma_k = \left(4 - 2\cos\frac{k\pi}{n+1}\right)^{-1}, \quad k = 1, \ldots, n,$$

gli autovalori di J_B sono (si veda l'esercizio 2.41)

$$\delta_{ij} = \alpha_i\gamma_j = \frac{\cos\dfrac{i\pi}{n+1}}{2 - \cos\dfrac{j\pi}{n+1}}, \quad i,\ j = 1, \ldots, n,$$

per cui risulta

$$\rho(J_B) = \frac{\cos\dfrac{\pi}{n+1}}{2 - \cos\dfrac{\pi}{n+1}}.$$

Per il teorema 5.22 gli autovalori di G_B sono i quadrati degli autovalori di J_B e quindi

$$\rho(G_B) = \rho^2(J_B),$$

e per il teorema 5.30 è

$$\omega_o = \frac{2}{1 + \sqrt{1 - \rho^2(J_B)}} = \frac{4 - 2\cos\dfrac{\pi}{n+1}}{\left(1 + \sqrt{1 - \cos\dfrac{\pi}{n+1}}\right)^2}$$

e

$$\rho[H_B(\omega_o)] = \omega_o - 1 = \frac{\left(1 - \sqrt{1 - \cos\dfrac{\pi}{n+1}}\right)^2}{\left(1 + \sqrt{1 - \cos\dfrac{\pi}{n+1}}\right)^2}.$$

Nei casi particolari $n = 5, 10, 20$ risulta

	$n = 5$	$n = 10$	$n = 20$
$\rho(J)$	0.8660254	0.9594930	0.9888308
$\rho(G)$	0.7500000	0.9206268	0.9777864
$\rho[H(\omega_o)]$	0.3333333	0.5603879	0.7405800
$\rho(J_B)$	0.7637079	0.9221398	0.9779084
$\rho(G_B)$	0.5832498	0.8503418	0.9563048
$\rho[H_B(\omega_o)]$	0.2153903	0.4421100	0.6542134

Per n grande si possono dare le seguenti valutazioni approssimate

$$\rho(J) \approx 1 - \frac{\pi^2}{2(n+1)^2}, \quad \rho(G) \approx 1 - \frac{\pi^2}{(n+1)^2}, \quad \rho[H(\omega_o)] \approx 1 - 2\frac{\pi}{n+1},$$

$$\rho(J_B) \approx 1 - \frac{\pi^2}{(n+1)^2}, \quad \rho(G_B) \approx 1 - \frac{2\pi^2}{(n+1)^2}, \quad \rho[H_B(\omega_o)] \approx 1 - 2\sqrt{2}\frac{\pi}{n+1}.$$

Perciò asintoticamente per ottenere un risultato con la stessa precisione il metodo di Jacobi richiede un numero di iterazioni doppio di quello richiesto dai metodi di Jacobi a blocchi e di Gauss-Seidel e pari a quattro volte il numero di iterazioni richiesto dal metodo di Gauss-Seidel a blocchi. Applicando il metodo di rilassamento la riduzione del numero di iterazioni rispetto al metodo di Jacobi è proporzionale ad n. ∎

7. Metodo del gradiente coniugato

Se la matrice A è reale e definita positiva, il sistema lineare $A\mathbf{x} = \mathbf{b}$ può essere risolto con il metodo del *gradiente coniugato*. Questo metodo, anche se in teoria è un metodo diretto, in quanto viene costruita una successione $\{\mathbf{x}^{(k)}\}_{k=0,1,\ldots}$ di vettori tali che $\mathbf{x}^{(m)} = \mathbf{x}^* = A^{-1}\mathbf{b}$, per un qualche indice $m \leq n$, in pratica però, per la presenza degli errori di arrotondamento, non termina all'm-esimo passo e viene utilizzato come metodo iterativo. In molti casi significativi il numero di iterazioni che occorrono per raggiungere la precisione richiesta è di gran lunga inferiore alla dimensione del sistema, e ciò rende il metodo molto conveniente per trattare problemi di grosse dimensioni, anche rispetto al metodo di rilassamento, poiché non richiede, fra l'altro, la determinazione preliminare di alcun parametro. Del metodo del gradiente coniugato esistono diverse varianti, che all'atto pratico generano successioni confrontabili. L'algoritmo che viene descritto in questo paragrafo è quello originario dovuto a Hestenes e Stiefel [6].

Sia $A \in \mathbf{R}^{n\times n}$, definita positiva, e $\mathbf{b} \in \mathbf{R}^n$, e si consideri il problema di minimizzare su $\mathbf{R}^n$ il funzionale

$$\Phi(\mathbf{x}) = \tfrac{1}{2}\ \mathbf{x}^T A\mathbf{x} - \mathbf{b}^T\mathbf{x}. \tag{44}$$

Tale problema ha una e una sola soluzione che, come si vedrà nel capitolo 7, è data da $\mathbf{x}^* = A^{-1}\mathbf{b}$. Quindi per calcolare la soluzione del sistema lineare $A\mathbf{x} = \mathbf{b}$ possono essere utilizzati dei metodi che minimizzano il funzionale (44). I metodi del *gradiente* sono metodi iterativi che minimizzano il funzionale (44) sfruttando il gradiente negativo di $\Phi(\mathbf{x})$, cioè il vettore

$$-\nabla\Phi(\mathbf{x}) = -\Big[\frac{\partial\Phi}{\partial x_1}\ (\mathbf{x}), \frac{\partial\Phi}{\partial x_2}\ (\mathbf{x}), \ldots, \frac{\partial\Phi}{\partial x_n}\ (\mathbf{x})\Big]^T = \mathbf{b} - A\mathbf{x} = \mathbf{r}(\mathbf{x}).$$

Il vettore $\mathbf{r} = \mathbf{r}(\mathbf{x})$ viene detto *residuo* del sistema $A\mathbf{x} = \mathbf{b}$.

Un metodo del gradiente procede nel modo seguente (per semplificare le notazioni, si scriverà in basso l'indice di iterazione): sia al k-esimo passo $\mathbf{x}_k \neq \mathbf{x}^*$, scelto un vettore direzione $\mathbf{p}_k \neq \mathbf{0}$ di decrescita per $\Phi(\mathbf{x})$, cioè tale che $\mathbf{p}_k^T\nabla\Phi(\mathbf{x}_k) < 0$, si determina il punto $\mathbf{x}_{k+1}$ di minimo del funzionale (44) sulla retta passante per $\mathbf{x}_k$ e di direzione $\mathbf{p}_k$:

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k\mathbf{p}_k, \tag{45}$$

dove $\alpha_k \in \mathbf{R}$ è tale che

$$\Phi(\mathbf{x}_{k+1}) = \min_{\alpha\in\mathbf{R}} \Phi(\mathbf{x}_k + \alpha\mathbf{p}_k).$$

Derivando la funzione $\Phi(\mathbf{x}_k + \alpha\mathbf{p}_k)$ rispetto ad α si ottiene

$$\frac{\partial\Phi}{\partial\alpha} = (\mathbf{x}_k + \alpha\mathbf{p}_k)^T A\mathbf{p}_k - \mathbf{b}^T\mathbf{p}_k,$$

da cui, imponendo che $\dfrac{\partial\Phi}{\partial\alpha} = 0$, si ricava

$$\alpha_k = \frac{(\mathbf{b} - A\mathbf{x}_k)^T\mathbf{p}_k}{\mathbf{p}_k^T A\mathbf{p}_k} = \frac{\mathbf{r}_k^T\mathbf{p}_k}{\mathbf{p}_k^T A\mathbf{p}_k}, \tag{46}$$

in cui $\mathbf{r}_k = \mathbf{r}(\mathbf{x}_k)$ è il residuo in $\mathbf{x}_k$. Poiché

$$\mathbf{r}_k^T\mathbf{p}_k = -\mathbf{p}_k^T\nabla\Phi(\mathbf{x}_k) > 0,$$

ne segue che $\alpha_k > 0$. Così procedendo si ottiene una successione $\{\mathbf{x}_k\}$ che converge al punto $\mathbf{x}^*$.

Dalla (45) segue per $k = 0, 1, \ldots$

$$\mathbf{b} - A\mathbf{x}_{k+1} = \mathbf{b} - A\mathbf{x}_k - \alpha_k A\mathbf{p}_k,$$

e quindi

$$\mathbf{r}_{k+1} = \mathbf{r}_k - \alpha_k A\mathbf{p}_k, \tag{47}$$

da cui per la (46)

$$\mathbf{r}_{k+1}^T \mathbf{p}_k = (\mathbf{r}_k - \alpha_k A\mathbf{p}_k)^T \mathbf{p}_k = \mathbf{r}_k^T \mathbf{p}_k - \alpha_k \mathbf{p}_k^T A\mathbf{p}_k = 0, \tag{48}$$

e quindi ad ogni passo il residuo $\mathbf{r}_{k+1}$ è ortogonale al vettore direzione $\mathbf{p}_k$ del passo precedente.

I diversi metodi del gradiente si distinguono per la diversa scelta del vettore $\mathbf{p}_k$: un metodo classico è quello dello *steepest descent*, in cui si sceglie $\mathbf{p}_k = \mathbf{r}_k = -\nabla\Phi(\mathbf{x}_k)$, cioè ad ogni passo il vettore $\mathbf{p}_k$ coincide con la direzione di massima pendenza per $\Phi(\mathbf{x})$. Questa strategia, anche se in ciascun punto $\mathbf{x}_k$ sfrutta la direzione della massima pendenza, può non essere la migliore, in particolare quando la matrice A è mal condizionata.

Nella figura 5.11, ad esempio, è illustrato il comportamento del metodo dello steepest descent nel caso di una matrice A di ordine 2. In ogni punto $\mathbf{x}_k$ si individua nel piano $\mathbf{R}^2$ la direzione $\mathbf{p}_k$, lungo la quale il funzionale $\Phi(\mathbf{x})$ decresce con la massima pendenza: il punto $\mathbf{x}_{k+1}$ è quello in cui il funzionale $\Phi(\mathbf{x})$ ha il valore minimo e in $\mathbf{x}_{k+1}$ la direzione $\mathbf{p}_k$ è tangente alla curva di livello $\Phi(\mathbf{x}) = \Phi(\mathbf{x}_{k+1})$. Il nuovo vettore direzione $\mathbf{p}_{k+1}$ è ortogonale al precedente vettore $\mathbf{p}_k$: infatti dalla (48) risulta

$$\mathbf{p}_{k+1}^T \mathbf{p}_k = 0, \quad \text{per} \quad k = 0, 1, \ldots \ .$$

La velocità di convergenza di questo procedimento dipende dalla eccentricità delle ellissi che rappresentano le curve di livello $\Phi(\mathbf{x}) = c$, dove c è una costante. L'eccentricità è tanto maggiore quanto maggiore è il rapporto λ_1/λ_2 degli autovalori λ_1 e λ_2, con $\lambda_1 > \lambda_2$, della matrice A, e quindi quanto più la matrice A è mal condizionata.

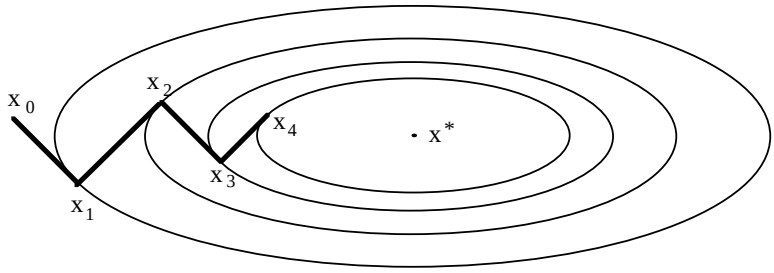


Fig. 5.11 - Il metodo dello steepest descent.

In generale se $\lambda_{\max}$ e $\lambda_{\min}$ sono il massimo e il minimo autovalore della matrice A di ordine n, si può dimostrare [3] che, indicato con $\mathbf{e}_k = \mathbf{x}^* - \mathbf{x}_k$ l'errore al k-esimo passo risulta (si veda l'esercizio 5.28)

$$\mathbf{e}_{k+1}^T A \mathbf{e}_{k+1} \leq \Big(\frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}}\Big)^2 \mathbf{e}_k^T A \mathbf{e}_k.$$

Introducendo la norma vettoriale (si veda l'esercizio 3.6)

$$\|\mathbf{x}\|_A = \sqrt{\mathbf{x}^T A \mathbf{x}}, \tag{49}$$

e notando che $\mu_2(A) = \dfrac{\lambda_{\max}}{\lambda_{\min}}$, si ottiene la limitazione dell'errore al k-esimo passo per il metodo dello steepest descent

$$\|\mathbf{e}_k\|_A \leq \Big(\frac{\mu_2(A) - 1}{\mu_2(A) + 1}\Big)^k \|\mathbf{e}_0\|_A. \tag{50}$$

È possibile ottenere una migliore strategia per la minimizzazione del funzionale $\Phi(\mathbf{x})$, con una scelta di $\mathbf{p}_k$ che tiene conto anche delle direzioni $\mathbf{p}_j,\ j = 1, 2, \ldots, k-1$, calcolate ai passi precedenti. Un metodo che utilizza questa strategia è il metodo del *gradiente coniugato*, in cui il vettore direzione $\mathbf{p}_k$ viene scelto nel modo seguente

$$\mathbf{p}_k = \begin{cases} \mathbf{r}_0 & \text{se } k = 0, \\ \\ \mathbf{r}_k + \beta_k \mathbf{p}_{k-1} & \text{se } k \geq 1, \end{cases} \tag{51}$$

dove β_k è tale che

$$\mathbf{p}_k^T A \mathbf{p}_{k-1} = 0. \tag{52}$$

Il vettore $\mathbf{p}_k$ viene detto *A-coniugato* con il vettore $\mathbf{p}_{k-1}$. Sostituendo nella (52) l'espressione di $\mathbf{p}_k$ data dalla (51), si ricava

$$\beta_k = -\ \frac{\mathbf{r}_k^T A \mathbf{p}_{k-1}}{\mathbf{p}_{k-1}^T A \mathbf{p}_{k-1}}\ , \quad k \geq 1. \tag{53}$$

La direzione $\mathbf{p}_k$ così scelta è una direzione di decrescita del funzionale $\Phi(\mathbf{x})$. Si ha infatti da (51) e (48)

$$-\mathbf{p}_k^T \nabla\Phi(\mathbf{x}_k) = \mathbf{p}_k^T \mathbf{r}_k = \mathbf{r}_k^T \mathbf{r}_k + \beta_k \mathbf{p}_{k-1}^T \mathbf{r}_k = \mathbf{r}_k^T \mathbf{r}_k > 0 \tag{54}$$

se $\mathbf{r}_k \neq \mathbf{0}$, cioè $\mathbf{x}_k \neq \mathbf{x}^*$.

Sostituendo la (54) nella (46) si ha che per il metodo del gradiente coniugato

$$\alpha_k = \frac{\mathbf{r}_k^T \mathbf{r}_k}{\mathbf{p}_k^T A \mathbf{p}_k} \ . \tag{55}$$

Da (51) e (48) si ha

$$\mathbf{r}_k^T \mathbf{r}_{k-1} = \mathbf{r}_k^T \mathbf{p}_{k-1} - \beta_{k-1} \mathbf{r}_k^T \mathbf{p}_{k-2} = -\beta_{k-1} \mathbf{r}_k^T \mathbf{p}_{k-2},$$

e poiché per la (47), la (48) e la (52) è

$$\mathbf{r}_k^T \mathbf{p}_{k-2} = \mathbf{r}_{k-1}^T \mathbf{p}_{k-2} - \alpha_k \mathbf{p}_{k-1}^T A \mathbf{p}_{k-2} = 0,$$

ne segue che

$$\mathbf{r}_k^T \mathbf{r}_{k-1} = 0, \tag{56}$$

cioè ogni residuo è ortogonale al precedente. Inoltre da (51), (56) e (54) si ha

$$\mathbf{p}_k^T \mathbf{r}_{k-1} = \mathbf{r}_k^T \mathbf{r}_{k-1} + \beta_k \mathbf{p}_{k-1}^T \mathbf{r}_{k-1} = \beta_k \mathbf{p}_{k-1}^T \mathbf{r}_{k-1} = \beta_k \mathbf{r}_{k-1}^T \mathbf{r}_{k-1},$$

e da (47), (52) e (54)

$$\mathbf{p}_k^T \mathbf{r}_{k-1} = \mathbf{p}_k^T \mathbf{r}_k + \alpha_{k-1} \mathbf{p}_k^T A \mathbf{p}_{k-1} = \mathbf{p}_k^T \mathbf{r}_k = \mathbf{r}_k^T \mathbf{r}_k,$$

da cui si ottiene un'altra relazione per β_k

$$\beta_k = \frac{\mathbf{r}_k^T \mathbf{r}_k}{\mathbf{r}_{k-1}^T \mathbf{r}_{k-1}} \ . \tag{57}$$

Indicato con S_k lo spazio generato dai k vettori $\mathbf{p}_0, \ldots, \mathbf{p}_{k-1}$, si può dimostrare (si veda l'esercizio 5.33) che le direzioni $\mathbf{p}_i$, $i = 0, \ldots, k-1$, ottenute con la (51) sono tali che

$$\Phi(\mathbf{x}_k) = \min_{\mathbf{x} \in S_k} \Phi(\mathbf{x}),$$

e quindi il metodo del gradiente coniugato determina la soluzione $\mathbf{x}^*$ in al più n passi, cioè esiste un $m \leq n$ tale che

$$\mathbf{r}_m = \mathbf{0}. \tag{58}$$

Questa proprietà si può ricavare direttamente dal seguente teorema.

5.32 Teorema. *Siano* $\mathbf{r}_0 \neq \mathbf{0}$ *ed* $h \geq 1$ *tale che* $\mathbf{r}_k \neq \mathbf{0}$ *per ogni* $k \leq h$. *Allora*

$$\left.\begin{aligned} \mathbf{r}_k^T \mathbf{r}_j &= 0 \\ \mathbf{p}_k^T A \mathbf{p}_j &= 0 \end{aligned}\right\} \quad \text{per } k \neq j,\ k, j = 0, \ldots, h, \tag{59}$$

cioè i primi h *residui costituiscono un insieme di vettori ortogonali e i vettori* $\mathbf{p}_k$ *costituiscono un insieme di vettori* A*-coniugati.*

Dim. Si procede per induzione su h.
Per $h = 1$, la (59) vale per $k = 1$ e $j = 0$, essendo da (56) e (52)

$$\mathbf{r}_1^T \mathbf{r}_0 = 0, \quad \text{e} \quad \mathbf{p}_1^T A \mathbf{p}_0 = 0.$$

Per $h > 1$, si suppone che valgano le (59) e si dimostra che

$$\left.\begin{aligned} \mathbf{r}_{h+1}^T \mathbf{r}_j &= 0 \\ \mathbf{p}_{h+1}^T A \mathbf{p}_j &= 0 \end{aligned}\right\} \quad \text{per} \quad j = 0, \ldots, h-1,$$

in quanto per $j = h$ l'ortogonalità dei residui è già stata dimostrata con la (56) e $\mathbf{p}_{h+1}$ è A-coniugato con $\mathbf{p}_h$ per la (52). Si ha dalla (47) per $j = 0, \ldots, h-1$

$$\mathbf{r}_{h+1}^T \mathbf{r}_j = \mathbf{r}_h^T \mathbf{r}_j - \alpha_h \mathbf{p}_h^T A \mathbf{r}_j = -\alpha_h \mathbf{p}_h^T A \mathbf{r}_j$$

per l'ipotesi induttiva, e per (51) è

$$\mathbf{p}_h^T A \mathbf{r}_j = \mathbf{p}_h^T A \mathbf{p}_j - \beta_j \mathbf{p}_h^T A \mathbf{p}_{j-1} = 0$$

per l'ipotesi induttiva. Quindi

$$\mathbf{r}_{h+1}^T \mathbf{r}_j = 0. \tag{60}$$

Inoltre per $j = 0, \ldots, h-1$ dalla (47) è

$$A \mathbf{p}_j = \frac{1}{\alpha_j}(\mathbf{r}_j - \mathbf{r}_{j+1}), \tag{61}$$

e quindi dalla (51) e dall'ipotesi induttiva

$$\mathbf{p}_{h+1}^T A \mathbf{p}_j = \mathbf{r}_{h+1}^T A \mathbf{p}_j + \beta_{h+1} \mathbf{p}_h^T A \mathbf{p}_j = \mathbf{r}_{h+1}^T A \mathbf{p}_j,$$

e da (61)

$$\mathbf{p}_{h+1}^T A \mathbf{p}_j = \frac{1}{\alpha_j}(\mathbf{r}_{h+1}^T \mathbf{r}_j - \mathbf{r}_{h+1}^T \mathbf{r}_{j+1}) = -\frac{1}{\alpha_j} \mathbf{r}_{h+1}^T \mathbf{r}_{j+1} = 0$$

per la (60) se $j = 0, \ldots, h-2$, e per la (56) se $j = h-1$. ∎

Dal teorema 5.32 segue che, poiché l'insieme dei primi h residui è formato da vettori ortogonali, non vi possono essere più di n vettori $\mathbf{r}_k \neq 0$ e quindi esiste un $m \leq n$ tale che vale la (58). Inoltre $\mathbf{r}_k$ appartiene al sottospazio generato dai vettori $\mathbf{r}_0, A\mathbf{r}_0, A^2\mathbf{r}_0, \ldots, A^k\mathbf{r}_0$, come si può vedere per induzione utilizzando la (47) e la (51). Se la matrice A ha al più s autovalori distinti, $s \leq n$, allora $\mathbf{r}_m = 0$ per qualche $m \leq s$. Infatti lo spazio generato da $\mathbf{r}_0, A\mathbf{r}_0, \ldots, A^{n-1}\mathbf{r}_0$, ha dimensione al più s (si veda l'esercizio 2.21) e quindi in esso non può esistere un vettore $\mathbf{r}_s$ non nullo che sia ortogonale a $\mathbf{r}_0, \mathbf{r}_1, \ldots, \mathbf{r}_{s-1}$. Ne segue che $\mathbf{r}_s = 0$.

Riassumendo, il metodo del gradiente coniugato può essere così descritto:

1. $k = 0,\ \mathbf{x}_0$ arbitrario, $\mathbf{r}_0 = \mathbf{b} - A\mathbf{x}_0$
2. se $\mathbf{r}_k = \mathbf{0}$, stop
3. altrimenti si calcoli

$$\beta_k = \frac{\mathbf{r}_k^T \mathbf{r}_k}{\mathbf{r}_{k-1}^T \mathbf{r}_{k-1}} \quad (\beta_0 = 0 \text{ per } k = 0),$$

$$\mathbf{p}_k = \mathbf{r}_k + \beta_k \mathbf{p}_{k-1} \quad (\mathbf{p}_0 = \mathbf{r}_0 \text{ per } k = 0),$$

$$\alpha_k = \frac{\mathbf{r}_k^T \mathbf{r}_k}{\mathbf{p}_k^T A \mathbf{p}_k},$$

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k,$$

$$\mathbf{r}_{k+1} = \mathbf{r}_k - \alpha_k A \mathbf{p}_k,$$

$k = k + 1$ e si vada al punto 2.

In questo procedimento per α_k e β_k si usano la (55) e la (57) che hanno un minor costo computazionale rispetto alle (46) e (53).

Per gli errori di arrotondamento il metodo può non terminare in n passi e viene di solito usato come metodo iterativo. Il calcolo si arresta quando il residuo $\mathbf{r}_k$ diventa sufficientemente piccolo. Poiché la quantità $\mathbf{r}_k^T \mathbf{r}_k$ viene già calcolata nel corso dell'algoritmo, conviene usare la seguente condizione di arresto:

$$\|\mathbf{r}_k\|_2 < \epsilon \, \|\mathbf{b}\|_2. \tag{62}$$

Un aspetto delicato dell'algoritmo dal punto di vista della stabilità è il calcolo del residuo $\mathbf{r}_{k+1}$ con la relazione ricorrente (47): allo scopo di contenere gli errori che si accumulano nel calcolo di $\mathbf{r}_{k+1}$, è opportuno dopo un certo numero di passi calcolare il residuo con la relazione $\mathbf{r}_{k+1} = \mathbf{b} - A\mathbf{x}_{k+1}$. In [11] si suggerisce di fare questa correzione ogni m passi, dove m è dell'ordine di $\sqrt{n}$.

L'operazione che presenta in generale un maggior costo computazionale è quella relativa alla moltiplicazione della matrice A per il vettore $\mathbf{p}_k$. Quindi la complessità del metodo, se questo richiedesse un numero di passi dell'ordine di n e se la matrice A non fosse sparsa, sarebbe dell'ordine di n^3, superiore a quella del metodo di Cholesky. Però nella risoluzione di sistemi di equazioni lineari che scaturiscono dalla discretizzazione di problemi differenziali, il numero di iterazioni richieste è di solito molto inferiore alla dimensione della matrice e la matrice A è sparsa (si veda a questo proposito l'esempio 5.36).

5.33 Esempio. Si applichi il metodo del gradiente coniugato al sistema lineare $A\mathbf{x} = \mathbf{b}$, dove

$$A = \begin{bmatrix} 7 & 4 & -7 \\ 4 & 5 & -3 \\ -7 & -3 & 8 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 4 \\ 6 \\ -2 \end{bmatrix},$$

la cui soluzione è $\mathbf{x}^* = [1, 1, 1]^T$. Tale sistema è stato già studiato nell'esempio 5.27. Con il metodo del gradiente coniugato, partendo dal vettore iniziale $\mathbf{x}_0 = \mathbf{0}$, si ottiene una successione dei residui $\mathbf{r}_k$ le cui norme sono

$$\|\mathbf{r}_0\|_2 = 7.483314$$
$$\|\mathbf{r}_1\|_2 = 3.712894$$
$$\|\mathbf{r}_2\|_2 = 0.2249498$$
$$\|\mathbf{r}_3\|_2 = 0.1500715\ 10^{-3}$$
$$\|\mathbf{r}_4\|_2 = 0.6938835\ 10^{-5}.$$

Se $\epsilon = 10^{-5}$, la condizione di arresto espressa dalla (62) risulta verificata dopo 4 iterazioni. ∎

Una limitazione dell'errore al k-esimo passo del metodo del gradiente coniugato, è data in [4] per la norma definita in (49)

$$\|\mathbf{e}_k\|_A \leq 2 \left(\frac{\sqrt{\mu_2(A)} - 1}{\sqrt{\mu_2(A)} + 1} \right)^k \|\mathbf{e}_0\|_A. \tag{63}$$

Si confronti questa relazione con la (50) relativa al metodo dello steepest descent.

Dalla relazione (63) segue che se $\mu_2(A)$ è un numero prossimo ad 1, sono sufficienti pochi passi per ottenere una buona riduzione dell'errore iniziale, mentre se $\mu_2(A)$ è elevato, è possibile che siano necessari fino ad n

passi per ottenere una approssimazione accettabile della soluzione, e se n è molto grande è possibile che non si riesca ad ottenere una approssimazione accettabile per la presenza degli errori di arrotondamento.

5.34 Esempio. Sia $n = 1000$ e siano

$$A = \begin{bmatrix} \alpha & 1 & & \\ 1 & \alpha & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & \alpha \end{bmatrix} \in \mathbf{R}^{n\times n},\ \alpha \geq 2, \quad \text{e} \quad \mathbf{b} = \begin{bmatrix} \alpha+1 \\ \alpha+2 \\ \vdots \\ \alpha+2 \\ \alpha+1 \end{bmatrix}.$$

La soluzione del sistema lineare $A\mathbf{x} = \mathbf{b}$ è data da $\mathbf{x}^* = [1, 1, \ldots, 1]^T$. Se $\mathbf{x}_0 = \mathbf{0}$ risulta

$$\|\mathbf{e}_0\|_A = \sqrt{\mathbf{x}^{*T} A \mathbf{x}^*} = \sqrt{n(\alpha+2) - 2}.$$

Gli autovalori della matrice A sono

$$\lambda_i = \alpha + 2\cos\frac{i\pi}{n+1}\ , \quad i = 1, 2, \ldots, n,$$

(si veda l'esercizio 2.40), per cui

$$\mu_2(A) = \frac{\alpha + 2\cos\dfrac{\pi}{n+1}}{\alpha - 2\cos\dfrac{\pi}{n+1}}.$$

Se $\alpha = 20$, allora $\mu_2(A) = 1.105541$, $\|\mathbf{e}_0\|_A = 148.3037$ e dalla (63) si ha

$$\|\mathbf{e}_k\|_A = \sigma^k \|\mathbf{e}_0\|_A, \quad \text{dove} \quad \sigma = 0.05012535,$$

e bastano 7 iterazioni del metodo del gradiente coniugato per ridurre l'errore di un fattore dell'ordine di 10^{-6}. Se invece $\alpha = 2.5$, allora $\mu_2(A) = 2.999968$, $\|\mathbf{e}_0\|_A = 67.03731$ e dalla (63) si ha

$$\|\mathbf{e}_k\|_A = \sigma^k \|\mathbf{e}_0\|_A, \quad \text{dove} \quad \sigma = 0.4999960,$$

e bastano 26 iterazioni del metodo del gradiente coniugato per ridurre l'errore di un fattore dell'ordine di 10^{-6}. In pratica, a causa degli errori di arrotondamento, non è possibile ottenere una soluzione affetta da un errore dell'ordine di 10^{-6} se si usa la relazione ricorrente (47) per il calcolo del residuo $\mathbf{r}_{k+1}$: infatti per $\alpha = 20$ dopo 6 iterazioni l'errore è dell'ordine di $\frac{1}{2}10^{-5}$ e non diminuisce più, mentre per $\alpha = 2.5$ dopo 19 iterazioni l'errore

è dell'ordine di $\frac{1}{2}10^{-4}$ e non diminuisce più. Se invece il residuo viene calcolato con la relazione $\mathbf{r}_{k+1} = \mathbf{b} - A\mathbf{x}_{k+1}$, per $\alpha = 20$ l'errore si riduce dell'ordine di $\frac{1}{2}10^{-6}$ in 5 passi per $\alpha = 20$ e in 18 passi per $\alpha = 2.5$, come è illustrato nella figura 5.12, in cui con i quadratini vuoti sono indicati gli errori generati nel caso $\alpha = 20$ e con i quadratini pieni gli errori per $\alpha = 2.5$.

■

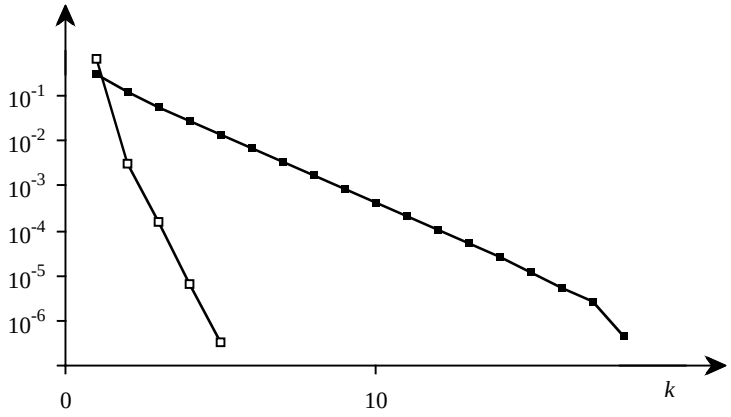


Fig. 5.12 - Errori generati dal metodo del gradiente coniugato applicato al sistema dell'esempio 5.34.

In alcuni casi è possibile ottenere una migliore convergenza utilizzando tecniche di *precondizionamento*, che trasformano il problema originale in un problema equivalente meglio condizionato.

Le tecniche di precondizionamento, che hanno avuto recentemente consistenti sviluppi, consistono essenzialmente nell'individuare una matrice C reale e non singolare, in modo che la matrice

$$B = C^{-1}AC^{-T}$$

sia tale che $\mu_2(B) < \mu_2(A)$. Il sistema a cui il metodo viene applicato è

$$B\mathbf{y} = \mathbf{c},$$

dove $\mathbf{y} = C^T\mathbf{x}$ e $\mathbf{c} = C^{-1}\mathbf{b}$. Naturalmente, per non aumentare eccessivamente il costo computazionale, la matrice C deve essere scelta di forma opportuna. La matrice

$$M = CC^T,$$

reale e definita positiva, è detta *precondizionatore* ed è quella che interviene nel seguente algoritmo del metodo del *gradiente coniugato con precondizionamento*. Infatti, poiché il residuo $\mathbf{s}_k$ del punto $\mathbf{y}_k$ nel metodo con

precondizionamento è legato al residuo $\mathbf{r}_k$ del punto $\mathbf{x}_k$ nel metodo senza precondizionamento dalla relazione

$$\mathbf{s}_k = C^{-1}\mathbf{r}_k,$$

allora

$$\mathbf{s}_k^T\mathbf{s}_k = \mathbf{r}_k^T M^{-1}\mathbf{r}_k.$$

Definito il vettore $\mathbf{z}_k$ tale che

$$M\mathbf{z}_k = \mathbf{r}_k,$$

si ha

$$\mathbf{s}_k^T\mathbf{s}_k = \mathbf{z}_k^T\mathbf{r}_k.$$

L'algoritmo del metodo del gradiente coniugato con precondizionamento risulta allora il seguente:

1. $k = 0,\ \mathbf{x}_0$ arbitrario, $\mathbf{r}_0 = \mathbf{b} - A\mathbf{x}_0$
2. se $\mathbf{r}_k = \mathbf{0}$, stop
3. altrimenti si calcoli

   $\mathbf{z}_k$ tale che $M\mathbf{z}_k = \mathbf{r}_k$,

   $\beta_k = \dfrac{\mathbf{z}_k^T\mathbf{r}_k}{\mathbf{z}_{k-1}^T\mathbf{r}_{k-1}} \quad (\beta_0 = 0)$,

   $\mathbf{p}_k = \mathbf{z}_k + \beta_k\mathbf{p}_{k-1} \quad (\mathbf{p}_0 = \mathbf{z}_0)$,

   $\alpha_k = \dfrac{\mathbf{z}_k^T\mathbf{r}_k}{\mathbf{p}_k^T A\mathbf{p}_k}$,

   $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k\mathbf{p}_k$,

   $\mathbf{r}_{k+1} = \mathbf{r}_k - \alpha_k A\mathbf{p}_k$,

   $k = k + 1$ e si vada al punto 2.

Per la (63) la velocità di convergenza del metodo con precondizionamento può risultare tanto maggiore quanto più la matrice B è "vicina" alla matrice identica, cioè quanto più la matrice M è "vicina" alla matrice A. Varie sono le tecniche di precondizionamento: di notevole interesse fra di esse quelle che si applicano a matrici con strutture particolari, come ad esempio le tecniche che si applicano alle matrici tridiagonali a blocchi trattate in [2].

Semplici tecniche di precondizionamento utilizzano la matrice

$$M = \begin{bmatrix} a_{11} & & & \\ & a_{22} & & \\ & & \ddots & \\ & & & a_{nn} \end{bmatrix},$$

che ha come elementi principali quelli di A o la matrice

$$M = \begin{bmatrix} A_{11} & & & \\ & A_{22} & & \\ & & \ddots & \\ & & & A_{nn} \end{bmatrix},$$

diagonale a blocchi, dove i blocchi diagonali quadrati A_{ii} sono i corrispondenti blocchi di A. Una tecnica più raffinata, frequentemente utilizzata e che è particolarmente conveniente se A è sparsa, è quella basata sulla *fattorizzazione incompleta di Cholesky*, in cui $M = LL^T$, dove L è una matrice triangolare inferiore, ottenuta nel modo seguente

$$l_{ii} = \sqrt{a_{ii} - \sum_{r=1}^{i-1} l_{ir}^2}, \quad i = 1, \ldots, n,$$

$$l_{ij} = \left\{ \begin{array}{ll} 0 & \text{se } a_{ij} = 0, \\ \dfrac{1}{l_{jj}} \left[a_{ij} - \displaystyle\sum_{r=1}^{j-1} l_{ir} l_{jr} \right] & \text{se } a_{ij} \neq 0, \end{array} \right\}, \; j = 1, \ldots, i-1, \; i = 2, \ldots, n.$$

La matrice L così ottenuta non è la matrice che si ottiene con fattorizzazione di Cholesky di A, in quanto in essa vengono posti a zero gli elementi che corrispondono a elementi nulli di A. In questo modo, se la matrice A è sparsa, anche L risulta una matrice sparsa e la risoluzione del sistema $M\mathbf{z}_k = \mathbf{r}_k$ viene ricondotta alla risoluzione di due sistemi con matrice triangolare sparsa.

5.35 Esempio Applicando il metodo del gradiente coniugato con il precondizionamento a blocchi di ordine 2 al sistema $A\mathbf{x} = \mathbf{b}$ dell'esempio 5.34 e calcolando il residuo esatto ogni 3 iterazioni, la soluzione viene calcolata con un errore dell'ordine di 10^{-6} con 3 iterazioni se $\alpha = 20$ e con 15 iterazioni se $\alpha = 2.5$. ∎

5.36 Esempio (problema di Dirichlet). Siano Ω il quadrato

$$\Omega = \{ \; (x, y) \in \mathbf{R}^2, 0 < x, y < 1 \; \}$$

e $\partial\Omega$ la sua frontiera e siano $f(x,y)$ e $g(x,y)$ due funzioni assegnate, definite rispettivamente su Ω e su $\partial\Omega$, tali che il problema

$$\begin{cases} \dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial^2 u}{\partial y^2} = f(x,y) & \text{per } (x,y) \in \Omega, \\ u(x,y) = g(x,y) & \text{per } (x,y) \in \partial\Omega, \end{cases} \tag{64}$$

abbia una e una sola soluzione sufficientemente regolare $u(x,y)$. La soluzione $u(x,y)$ può essere approssimata nel modo seguente: si costruisce un reticolo formato di linee parallele agli assi, di solito ugualmente distanti fra di loro, e si considera un problema discreto che approssima il problema (64) nei nodi del reticolo. Più esattamente, fissato un intero n si considerano i punti

$$(x_i, y_j), \quad \text{tali che } x_i = ih,\ y_j = jh,\ h = \frac{1}{n+1},\ i,j = 0,\ldots,n+1.$$

Il reticolo corrispondente è illustrato nella figura 5.13.

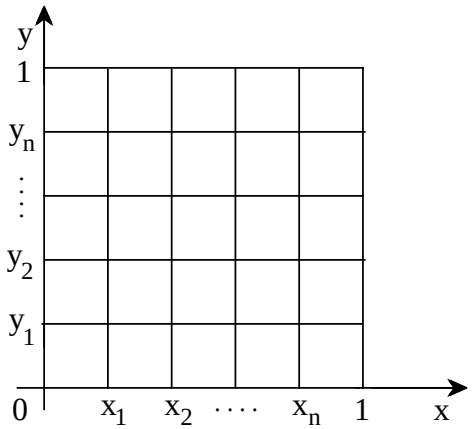


Fig. 5.13 - Reticolo per l'approssimazione della soluzione del problema di Dirichlet.

Per calcolare un'approssimazione della derivata seconda di una funzione $F(x)$ che si suppone derivabile 4 volte con continuità, si utilizzano le due espressioni ottenute con la formula di Taylor

$$F(x_0 - h) = F(x_0) - hF'(x_0) + \frac{h^2}{2}F''(x_0) - \frac{h^3}{3!}F'''(x_0) + \frac{h^4}{4!}F^{(4)}(\xi_1),$$

$$F(x_0 + h) = F(x_0) + hF'(x_0) + \frac{h^2}{2}F''(x_0) + \frac{h^3}{3!}F'''(x_0) + \frac{h^4}{4!}F^{(4)}(\xi_2),$$

da cui si ricava che

$$F''(x_0) = \frac{1}{h^2}[F(x_0 - h) - 2F(x_0) + F(x_0 + h)] + O(h^2).$$

Utilizzando questa relazione, la restrizione della prima equazione del problema (64) ai nodi del reticolo è data da

$$u(x_{i-1}, y_j) - 2u(x_i, y_j) + u(x_{i+1}, y_j) + u(x_i, y_{j-1}) - 2u(x_i, y_j) + u(x_i, y_{j+1})$$

$$= h^2[f(x_i, y_j) + O(h^2)], \qquad i,\ j = 1, \ldots, n. \tag{65}$$

La restrizione della seconda equazione ai nodi del reticolo è data da

$$\begin{aligned} u(x_i, y_j) = g(x_i, y_j), \quad &\text{per } i = 0 \text{ e } i = n+1,\ j = 1, \ldots, n, \\ &\text{e per } i = 1, \ldots, n,\ j = 0 \text{ e } j = n+1. \end{aligned} \tag{66}$$

Trascurando i termini in $O(h^2)$, le relazioni (65) e (66) si riducono ad un sistema di equazioni lineari che consente di calcolare le approssimazioni $u_{i,j}$ della funzione $u(x, y)$ nei punti (x_i, y_j), $i,\ j = 1, \ldots, n$. Il sistema che si ottiene è il seguente

$$-u_{i,j-1} - u_{i-1,j} + 4u_{i,j} - u_{i+1,j} - u_{i,j+1} = -h^2 f(x_i, y_j), \quad i,\ j = 1, \ldots, n,$$

$$\begin{aligned} u_{i,j} = g(x_i, y_j), \quad &\text{per } i = 0 \text{ e } i = n+1,\ j = 1, \ldots, n, \\ &\text{e per } i = 1, \ldots, n,\ j = 0 \text{ e } j = n+1. \end{aligned}$$

Per semplicità si studia il caso in cui $f(x, y) = 0$ per $(x, y) \in \Omega$ e $g(x, y) = 1$ per $(x, y) \in \partial\Omega$, la cui soluzione è $u(x, y) = 1$. Il sistema lineare $A\mathbf{u} = \mathbf{b}$ ottenuto è

$$\begin{bmatrix} B & -I & & \\ -I & B & \ddots & \\ & \ddots & \ddots & -I \\ & & -I & B \end{bmatrix} \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_n \end{bmatrix} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_n \end{bmatrix}, \tag{67}$$

dove

$$B = \begin{bmatrix} 4 & -1 & & \\ -1 & 4 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 4 \end{bmatrix} \in \mathbf{R}^{n \times n},\ \mathbf{u}_j = \begin{bmatrix} u_{1,j} \\ u_{2,j} \\ \vdots \\ u_{n,j} \end{bmatrix} \in \mathbf{R}^n,\ \text{per } j = 1, \ldots, n,$$

$$\mathbf{b}_1 = \mathbf{b}_n = \begin{bmatrix} 2 \\ 1 \\ \vdots \\ 1 \\ 2 \end{bmatrix} \in \mathbf{R}^n, \quad \mathbf{b}_j = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \in \mathbf{R}^n, \text{ per } j = 2, \ldots, n-1.$$

La matrice A ha predominanza diagonale ed è irriducibile: per il teorema 2.41 A è non singolare e quindi il sistema (67) ha una e una sola soluzione $u_{i,j},\ i, j = 1, \ldots, m$, che si assume come approssimazione della soluzione del problema (64). In generale è possibile dimostrare [7] che, sotto opportune ipotesi di regolarità della funzione $u(x, y)$, per $h \to 0$, cioè al tendere a zero dell'ampiezza delle maglie del reticolo, la soluzione del sistema (67) tende alla soluzione $u(x, y)$ del problema (64), nel senso che

$$\lim_{h \to 0} \max_{i,j=1,\ldots,n} |u(x_i, y_j) - u_{ij}| = 0.$$

I metodi esposti in questo testo per la risoluzione dei sistemi lineari sono stati utilizzati per risolvere il sistema (67) nei casi $n = 5$, 10 e 20, cioè per sistemi di ordine rispettivamente 25, 100 e 400. È opportuno rilevare che i sistemi lineari che scaturiscono da problemi reali hanno dimensioni molto maggiori, dell'ordine di decine di migliaia di equazioni, ed è possibile risolverli efficientemente solo perché la matrice ha specifiche proprietà di struttura, quale quella di essere sparsa. Il problema dell'esempio, pur con le sue ridotte dimensioni, è comunque significativo, per confrontare e mettere in evidenza le caratteristiche dei metodi proposti.

Si sono utilizzati i seguenti metodi

G	metodo di Gauss	metodi diretti
GB	metodo di Gauss a blocchi	
C	metodo di Cholesky	
H	metodo di Householder	

J	metodo di Jacobi	metodi iterativi
JB	metodo di Jacobi a blocchi	
GS	metodo di Gauss-Seidel	
GSB	metodo di Gauss-Seidel a blocchi	
S	metodo di rilassamento	
SB	metodo di rilassamento a blocchi	

GC	metodo del gradiente coniugato	metodi del gradiente coniugato
GCB	metodo del gradiente coniugato con il precondizionamento a blocchi	
GCC	metodo del gradiente coniugato con la fattor. incompleta di Cholesky	

Nella tabella 5.1 sono riportati i risultati ottenuti utilizzando i metodi diretti. t indica il tempo impiegato, misurato in millesimi di secondo, ed $\|\mathbf{e}\|_2$ indica la norma 2 dell'errore assoluto della soluzione calcolata.

Nella tabella 5.2 sono riportati i risultati ottenuti utilizzando i metodi iterativi. È stata usata la condizione di arresto (16) $\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|_\infty < \epsilon$ con la tolleranza $\epsilon = 10^{-5}$ e k indica il numero delle iterazioni richieste. Come si nota, i risultati ottenuti sono del tutto in accordo con i risultati dell'esempio 5.31, e in particolare per il caso n=20 con le stime asintotiche della velocità di convergenza.

Nella tabella 5.3 sono riportati i risultati ottenuti utilizzando il metodo del gradiente coniugato: si è usata la condizione di arresto (62) $\|\mathbf{r}_k\|_2 < \epsilon\|\mathbf{b}\|_2$, con la tolleranza $\epsilon = 10^{-5}$. Come si nota, il metodo del gradiente coniugato consente di approssimare la soluzione con una precisione maggiore e in un tempo minore rispetto agli altri metodi e la precisione richiesta viene raggiunta con un numero di iterazioni assai inferiore a n^2, dimensione della matrice A.

Per implementare su calcolatore i vari metodi sono state sfruttate le specifiche proprietà della matrice. In particolare sia il metodo di Gauss che quello di Cholesky, utilizzando il fatto che la matrice è definita positiva e a banda di ampiezza n richiedono un numero di operazioni dell'ordine $n^4/2$ (si veda l'esercizio 4.41). Il metodo di Householder genera una matrice a banda superiore di ampiezza $2n$, e richiede $4n^4$ operazioni (si veda l'esercizio 4.41). I tempi di calcolo sono legati, oltre che al numero delle operazioni, anche ad altri fattori (per esempio l'individuazione e il trasferimento dei dati), che dipendono anche dalle tecniche di programmazione. In particolare per valori piccoli di n, al numero delle operazioni, va aggiunto il numero delle radici quadrate per il metodo di Cholesky e di Householder e per quello del gradiente coniugato con la fattorizzazione incompleta di Cholesky.

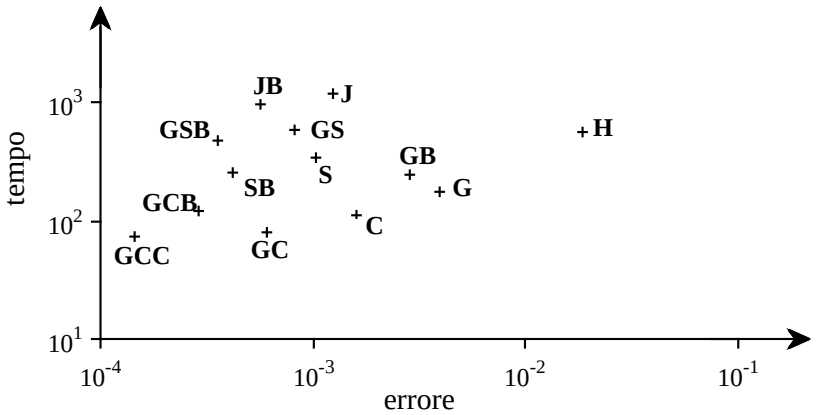


Fig. 5.14 - Tempo (in millisec.) ed errore dei metodi usati per risolvere il sistema (67).

Per il caso $n = 20$ (ordine del sistema 400) i risultati delle tabelle sono sintetizzati nella figura 5.14, in cui in ascissa sono riportati gli errori del risultato generato da ciascun metodo e in ordinata i tempi di esecuzione in millesimi di secondo. ∎

Esercizi proposti

5.1 Sia $A \in \mathbf{C}^{n\times n}$. Senza utilizzare il teorema 5.2,

a) si dimostri per ogni norma matriciale indotta $\| \, . \, \|$ che se

$$\lim_{k\to\infty} \|A^k\| = 0 \quad \text{allora} \quad \rho(A) < 1.$$

Posto $A = UTU^H$, U unitaria e T triangolare superiore, se $\rho(A) < 1$, si dimostri che

b) se T ha n autovettori linearmente indipendenti, allora $\lim_{k\to\infty} T^k = O$;

c) se T non ha n autovettori linearmente indipendenti, si può costruire una matrice $S \in \mathbf{R}^{n\times n}$ tale che $S > |T|$, $\rho(S) < 1$ e S ha n autovettori linearmente indipendenti, e quindi

$$\lim_{k\to\infty} S^k = O;$$

d) se $\rho(A) < 1$, esiste una norma matriciale indotta $\| \, . \, \|$ per cui

$$\lim_{k\to\infty} \|A^k\| = 0.$$

Sfruttando i risultati dei punti a), b) e c), si dia una dimostrazione del teorema 5.2, in cui si faccia riferimento alla forma normale di Schur, anziché alla forma normale di Jordan; sfruttando i risultati dei punti a) e d), si dia un'altra dimostrazione del teorema 5.2.

(Traccia: a) per il teorema 3.10 è $\|A^k\| \geq [\rho(A)]^k$; b) sia D diagonale tale che $T = CDC^{-1}$, allora $T^k = CD^kC^{-1}$, inoltre $\rho(D) = \rho(T) = \rho(A) < 1$ e quindi $\lim_{k\to\infty} D^k = O$; c) poiché $\rho(T) < 1$, gli elementi principali di T sono in modulo minori di 1, basta porre S la matrice i cui elementi sono

$$s_{ij} = \begin{cases} \alpha_i & \text{per } j = i, \\ \beta & \text{per } j > i, \\ 0 & \text{per } j < i, \end{cases} \quad \text{dove} \quad \begin{cases} \alpha_i > |t_{ii}| & \text{per } i = 1, \ldots, n, \\ \beta > |t_{ij}| & \text{per } i, j = 1, \ldots, n, \end{cases}$$

in cui gli elementi α_i sono tutti diversi fra loro e minori di 1; d) per il teorema 3.12 esiste una norma indotta $\|\ .\ \|$ tale che $\|A\| < 1$, e si tenga conto della proprietà $\|A^k\| \leq \|A\|^k$. Per la dimostrazione del teorema 5.2 si ha che se $\lim_{k\to\infty} A^k = O$, allora $\rho(A) < 1$ per il punto a); viceversa, se $\rho(A) < 1$, allora $\lim_{k\to\infty} A^k = \lim_{k\to\infty} T^k = O$ per il punto b) oppure $\lim_{k\to\infty} |A^k| = \lim_{k\to\infty} |T^k| \leq \lim_{k\to\infty} S^k = O$ per il punto c). Alternativamente, il viceversa segue dal punto d). Questa seconda dimostrazione è stata data da Householder.)

5.2 Si determinino due matrici A e $B \in \mathbf{R}^{2\times 2}$ tali che

$$\lim_{k\to\infty} A^k = O, \quad \lim_{k\to\infty} B^k = O,$$

ma il $\lim_{k\to\infty} (AB)^k$ non esista.

(Risposta: ad esempio

$$A = B^T = \begin{bmatrix} \frac{1}{2} & 1 \\ 0 & 0 \end{bmatrix}.)$$

5.3 Sia $A \in \mathbf{R}^{n\times n}$ definita da

$$a_{ij} = \begin{cases} i & \text{per } i = j, \\ \dfrac{1}{2^{j-1}} & \text{per } i < j, \\ \dfrac{1}{2^{i-1}} & \text{per } i > j. \end{cases}$$

Si dimostri che per ogni $\alpha \in \mathbf{R}$ tale che $0 < \alpha < \dfrac{1}{n+1}$ vale

$$A^{-1} = \alpha \sum_{k=0}^{\infty} (I - \alpha A)^k.$$

(Traccia: poiché $A = \dfrac{1}{\alpha}\ (I - (I - \alpha A))$, si imponga la condizione che $\rho(I - \alpha A) < 1$, verificando che A ha predominanza diagonale in senso stretto e quindi è definita positiva, ed inoltre

$$\|A\|_\infty \leq n + \sum_{k=1}^{n-1} \frac{1}{2^k} < n + 1.\ )$$

5.4 Sia

$$A = \frac{1}{2}\begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}.$$

Si dica per quali valori di α la successione di vettori $\{\mathbf{x}^{(k)}\}$ definita da

$$\mathbf{x}^{(k)} = (I + \alpha A + \alpha^2 A^2)\mathbf{x}^{(k-1)}, \quad k = 1, 2, \ldots,$$

dove $\mathbf{x}^{(0)} \in \mathbf{C}^n$ è un vettore qualsiasi, converge al vettore $\mathbf{x}^* = \mathbf{0}$ per $k \to \infty$.

(Risposta: $-\frac{1}{2} < \alpha < 0$.)

5.5 Sia $P \in \mathbf{C}^{n\times n}$ una matrice nilpotente e $\mathbf{q} \in \mathbf{C}^n$. Si dimostri che il metodo iterativo

$$\mathbf{x}^{(k)} = P\mathbf{x}^{(k-1)} + \mathbf{q}$$

è convergente in un numero finito di passi (per la definizione e le proprietà delle matrici nilpotenti si vedano gli esercizi 1.10 e 2.32). Si esamini in particolare il metodo di Jacobi applicato al sistema lineare $A\mathbf{x} = \mathbf{b}$, dove

$$A = \begin{bmatrix} 1 & 2 & -2 \\ 1 & 1 & 1 \\ 2 & 2 & 1 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 1 \\ 3 \\ 5 \end{bmatrix}.$$

Per questo sistema il metodo di Gauss-Seidel è convergente?

(Traccia: sia k tale che $P^k = O$, allora $\mathbf{e}^{(k)} = P^k\mathbf{e}^{(0)} = \mathbf{0}$ per ogni $\mathbf{e}^{(0)} \in \mathbf{C}^n$; nel caso particolare J è nilpotente e $J^3 = O$. Il metodo di Gauss-Seidel non è convergente.)

5.6 Siano

$$P = \begin{bmatrix} \frac{1}{2} & \frac{3}{2} & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}, \quad \mathbf{q} = \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix};$$

si dica se il metodo iterativo

$$\mathbf{x}^{(k)} = P\mathbf{x}^{(k-1)} + \mathbf{q}$$

è convergente e si studino in particolare le successioni ottenute ponendo

$$(1)\ \mathbf{x}^{(0)} = [2, -1, -1]^T \quad \text{e} \quad (2)\ \mathbf{x}^{(0)} = [1, 1, 1]^T.$$

(Risposta: il metodo non è convergente, la successione ottenuta da (1) è convergente, quella ottenuta da (2) no.)

5.7 Sia $A \in \mathbf{C}^{n\times n}$ singolare e $A = M - N$, in cui M è non singolare. Si dimostri che il metodo iterativo

$$\mathbf{x}^{(k)} = M^{-1}N\mathbf{x}^{(k-1)} + M^{-1}\mathbf{b}$$

per la risoluzione del sistema $A\mathbf{x} = \mathbf{b}$ non è convergente.

(Traccia: sia $\mathbf{x} \neq \mathbf{0}$ tale che $(M - N)\mathbf{x} = \mathbf{0}$, allora $\mathbf{x} = M^{-1}N\mathbf{x}$ e quindi $\rho(M^{-1}N) \geq 1$.)

5.8 Siano $A \in \mathbf{C}^{n\times n}$, $\mathbf{b} \in \mathbf{C}^n$, $\omega \in \mathbf{R}$, si consideri il metodo iterativo

$$\mathbf{x}^{(k)} = -\frac{1}{\omega}\,(A - \omega I)\mathbf{x}^{(k-1)} + \frac{\mathbf{b}}{\omega}.$$

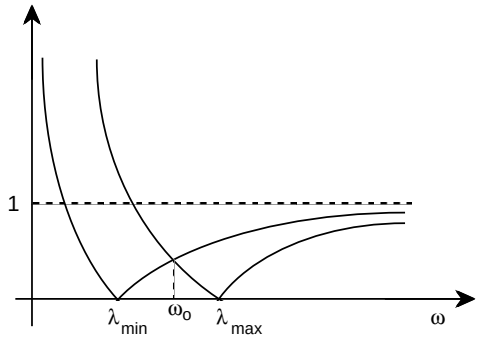
a) Si determini il limite della successione $\{\mathbf{x}^{(k)}\}$, nell'ipotesi che il metodo converga;

b) se A è hermitiana, si dia una condizione necessaria e sufficiente di convergenza e si dica qual è il valore ω_o di ω per cui il raggio spettrale della matrice di iterazione è minimo;

c) si esamini il caso particolare in cui

$$A = \begin{bmatrix} 3 & 2 \\ 1 & 2 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

(Traccia: a) posto $P(\omega) = I - \dfrac{1}{\omega}\,A$, se $\rho[P(\omega)] < 1$, A è non singolare e $\lim\limits_{k\to\infty} = \mathbf{x}^* = A^{-1}\mathbf{b}$; b) A definita positiva o definita negativa, ω dello stesso segno degli autovalori di A e $|\omega| > \dfrac{\rho(A)}{2}$; $\omega_o = \dfrac{\lambda_{\min} + \lambda_{\max}}{2}$, in cui $\lambda_{\min}$ e $\lambda_{\max}$ sono il minimo e il massimo modulo degli autovalori di A, si veda la figura per il caso in cui A è definita positiva.



c) il metodo è convergente per $\omega > 2$ e $\omega_o = \dfrac{5}{2}$.)

5.9 Sia $A \in \mathbf{C}^{n\times n}$.

a) Si diano delle condizioni necessarie e sufficienti per la convergenza delle successioni $\{X^{(k)}\}$ e $\{Y^{(k)}\}$ definite da

$$(1) \qquad X^{(k)} = I + (I - A)X^{(k-1)}, \quad X^{(0)} = I,$$

$$(2) \qquad Y^{(k)} = \omega I + (I - \omega A)Y^{(k-1)}, \quad Y^{(0)} = \omega I, \ \omega \in \mathbf{R};$$

b) si dimostri che se i limiti delle due successioni esistono, essi coincidono e se ne determini il comune valore X^*;

c) si dia una maggiorazione della norma dell'errore al k-esimo passo

$$\|X^* - X^{(k)}\| \quad \text{o} \quad \|X^* - Y^{(k)}\|,$$

nell'ipotesi che

$$\|I - A\| < 1 \quad \text{o} \quad \|I - \omega A\| < 1;$$

d) si esamini il caso particolare in cui

$$A = \begin{bmatrix} 1-3\alpha & 1-3\alpha & 0 \\ -2\alpha & 3 & -3\alpha-2 \\ 0 & 8\alpha & 1-12\alpha \end{bmatrix}, \quad \alpha \in \mathbf{R}.$$

(Traccia: siano λ_i gli autovalori di A. (1) $|1 - \lambda_i| < 1$ per $i = 1, \ldots, n$; (2) Re (λ_i) di segno costante per $i = 1, \ldots, n$, ω dello stesso segno e tale che

$$\omega < \min_{i=1,\ldots,n} \left\{ \frac{2\text{Re } (\lambda_i)}{|\lambda_i|^2} \right\};$$

b) $X^* = A^{-1}$; c) per la successione (1) si ha

$$X^{(k+1)} - X^{(k)} = (I - A)^k (X^{(1)} - X^{(0)})$$

e

$$X^{(k+1)} - X^{(k)} = I - AX^{(k)} = A(A^{-1} - X^{(k)}),$$

e quindi

$$A^{-1} - X^{(k)} = A^{-1}(I - A)^k (X^{(1)} - X^{(0)}),$$

da cui passando alle norme

$$\|A^{-1} - X^{(k)}\| \leq \frac{\|I - A\|^k}{1 - \|I - A\|} \, \|X^{(1)} - X^{(0)}\|;$$

per la successione (2) si proceda in modo analogo; d) la successione (1) converge per $\frac{1}{9} < \alpha < \frac{1}{6}$, la successione (2) converge per $\alpha < \frac{1}{6}$ e $0 < \omega < \dfrac{2}{3(1-3\alpha)}$.)

5.10 Sia $A \in \mathbf{C}^{n\times n}$. Si consideri in $\mathbf{C}^{n\times n}$ la successione di matrici $\{X^{(k)}\}$ così definita

$$X^{(k)} = X^{(k-1)}(2I - AX^{(k-1)}), \quad k = 1, 2, \ldots \tag{68}$$

a) Si dimostri che se $X^{(0)}$ commuta con A, allora $X^{(k)}$ commuta con A per ogni k;

b) si dimostri che la successione, se è convergente, converge a A^{-1} e, indicate con $R^{(k)} = I - AX^{(k)}$ e $E^{(k)} = A^{-1} - X^{(k)}$ le matrici residuo ed errore al k-esimo passo, si verifichi che

$$R^{(k)} = [R^{(k-1)}]^2 \quad \text{e} \quad E^{(k)} = E^{(k-1)}AE^{(k-1)};$$

c) si dia una condizione necessaria e sufficiente affinchè la successione sia convergente;

d) si studi il caso particolare in cui la matrice A è definita positiva e si scelga $X^{(0)} = \alpha I$, con α costante reale non nulla.

La successione (68) definisce quindi un metodo iterativo del "secondo ordine" per calcolare A^{-1} e viene spesso usata, sotto il nome di *raffinamento iterativo*, per migliorare l'approssimazione $\widetilde{B}$ di una matrice inversa $B = A^{-1}$ calcolata con un altro metodo. Si dica qual è il costo computazionale di ogni passo del metodo.

(Traccia: a) si proceda per induzione; c) $\rho(R^{(0)}) < 1$; d) la successione è convergente se $0 < \alpha < \dfrac{2}{\rho(A)}$.)

5.11 Sia $A\mathbf{x} = \mathbf{b}$ un sistema lineare, sia $\widetilde{\mathbf{x}}$ la soluzione calcolata con il metodo di Gauss, e siano $\widetilde{L}$ e $\widetilde{U}$ le matrici effettivamente calcolate della fattorizzazione LU di A. Si consideri il seguente metodo iterativo:

$\mathbf{x}_0 = \widetilde{\mathbf{x}}$,

per $k = 1, 2, \ldots$ si calcoli

$\mathbf{r}_k = \mathbf{b} - A\mathbf{x}_{k-1}$,

$\mathbf{y}_k$ come soluzione del sistema $\widetilde{L}\mathbf{y} = \mathbf{r}_k$,

$\mathbf{z}_k$ come soluzione del sistema $\widetilde{U}\mathbf{z} = \mathbf{y}_k$,

$\mathbf{x}_k = \mathbf{x}_{k-1} + \mathbf{z}_k$.

Questo algoritmo, indicato come metodo di *correzione post-iterativa*, viene utilizzato per migliorare l'approssimazione della soluzione del sistema lineare calcolata con il metodo di Gauss. Si dica sotto quale ipotesi il metodo è convergente, con quale precisione devono essere eseguiti i calcoli e il costo computazionale.

(Risposta: $\rho(I-\widetilde{A}^{-1}A) < 1$, dove $\widetilde{A} = \widetilde{L}\widetilde{U}$; nel calcolo di $\mathbf{r}_k$ si deve utilizzare una precisione maggiore di quella usata per il calcolo di $\widetilde{\mathbf{x}}$, ad esempio la doppia precisione se il calcolo di $\widetilde{\mathbf{x}}$ è stato fatto con la precisione semplice; ogni passo di iterazione richiede $2n^2$ operazioni moltiplicative.)

5.12 Siano

$$P = \begin{bmatrix} \frac{1}{2} & 0 & 0 & 1 \\ 0 & 0 & \frac{1}{3} & 0 \\ 0 & \frac{1}{4} & 0 & 0 \\ 1 & 0 & 0 & \frac{1}{5} \end{bmatrix}, \quad \mathbf{q} \in \mathbf{R}^4.$$

a) Si dica se il metodo iterativo

$$\mathbf{x}^{(k)} = P\mathbf{x}^{(k-1)} + \mathbf{q}$$

è convergente;

b) si verifichi che il metodo di Jacobi applicato al sistema $(I-P)\mathbf{x} = \mathbf{q}$ non è convergente, ma esiste una permutazione delle equazioni del sistema per cui il metodo di Jacobi è convergente.

(Risposta: a) no; b) è

$$J = \begin{bmatrix} 0 & 0 & 0 & 2 \\ 0 & 0 & \frac{1}{3} & 0 \\ 0 & \frac{1}{4} & 0 & 0 \\ \frac{5}{4} & 0 & 0 & 0 \end{bmatrix}$$

e $\rho(J) = \sqrt{\frac{5}{2}}$, per cui il metodo non è convergente; scambiando la prima e l'ultima equazione del sistema, si ottiene una matrice dei coefficienti con predominanza diagonale in senso stretto.)

5.13 Si determini una condizione necessaria e sufficiente per la convergenza del metodo di Jacobi applicato al sistema lineare $A\mathbf{x} = \mathbf{b}$, dove $A \in \mathbf{R}^{n\times n}$ è la matrice i cui elementi sono dati da

$$a_{ij} = \begin{cases} 1 & \text{per } i = j, \\ \\ k & \text{per } i \neq j, \end{cases}$$

e $\mathbf{b} \in \mathbf{R}^n$.

(Traccia: è

$$J = -k \begin{bmatrix} 0 & 1 & \dots & 1 \\ 1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 1 & \dots & 1 & 0 \end{bmatrix}$$

e gli autovalori di J sono $\lambda_1 = k$ (di molteplicità $n-1$) e $\lambda_2 = -(n-1)k$ (si veda l'esercizio 2.37). Quindi il metodo di Jacobi è convergente se e solo se $|k| < \dfrac{1}{n-1}$.)

5.14 È dato il sistema lineare $A\mathbf{x} = \mathbf{b}$, dove

$$A = \begin{bmatrix} 1 & \alpha & -\alpha \\ 1 & 1 & 1 \\ \alpha & \alpha & 1 \end{bmatrix}, \quad \alpha \in \mathbf{R}.$$

a) Si determinino i valori di α per i quali i metodi di Jacobi e di Gauss-Seidel sono convergenti;

b) per i valori di α per i quali entrambi i metodi sono convergenti, si dica quale dei due ha il maggiore tasso asintotico di convergenza.

(Risposta: a) $\rho(J) = \sqrt{|2\alpha - \alpha^2|}$, $\rho(G) = |\alpha|$, il metodo di Jacobi è convergente per $|\alpha - 1| < \sqrt{2}$ e $\alpha \neq 1$, il metodo di Gauss-Seidel è convergente per $|\alpha| < 1$; b) il metodo di Gauss-Seidel.)

5.15 Sia $A \in \mathbf{C}^{n\times n}$ una matrice a predominanza diagonale in senso stretto, e siano

$$J = L + U \quad \text{e} \quad G = (I - L)^{-1}U,$$

dove $L = D^{-1}B$ e $U = D^{-1}C$, le matrici di iterazione di Jacobi e di Gauss-Seidel di $A = D - B - C$.

a) Si dimostri che

$$(I - L)^{-1} = \sum_{i=0}^{n-1} L^i \quad \text{e} \quad I \leq |(I - L)^{-1}| \leq (I - |L|)^{-1},$$

e quindi

$$|J| = |L| + |U| \quad \text{e} \quad |G| \leq (I - |L|)^{-1}|U|;$$

b) se $\mathbf{e} = [1, 1, \dots, 1]^T$, si dimostri che

$$|G|\mathbf{e} \leq |J|\mathbf{e}$$

e quindi

$$\|G\|_\infty \leq \|J\|_\infty.$$

(Traccia: a) poiché L è triangolare inferiore con elementi principali nulli, è $L^i = 0$ per $i \geq n$; b)

$$\begin{aligned} |G| &\leq |(I-L)^{-1}|\ |U| \leq (I-|L|)^{-1}|U| \\ &= I-(I-|L|)^{-1}(I-|L|-|U|) = I-(I-|L|)^{-1}(I-|J|); \end{aligned}$$

per la predominanza diagonale in senso stretto di A, il vettore $(I-|J|)\mathbf{e}$ ha tutte componenti positive, e poiché $(I-|L|)^{-1} \geq I$, risulta

$$|G|\ \mathbf{e} \leq \mathbf{e} - (I-|J|)\ \mathbf{e} = |J|\ \mathbf{e}.\ )$$

5.16 Sia $A \in \mathbf{C}^{2\times 2}$. Si dimostri che $\rho^2(J) = \rho(G)$, per cui

a) i metodi di Jacobi e di Gauss-Seidel sono entrambi convergenti o entrambi divergenti,

b) se A è definita positiva entrambi i metodi sono convergenti,

c) se i metodi sono convergenti, il tasso asintotico di convergenza del metodo di Gauss-Seidel è doppio di quello di Jacobi,

d) si esamini in particolare il caso

$$A = \begin{bmatrix} \alpha & 1 \\ 1 & \alpha \end{bmatrix}, \quad \alpha \in \mathbf{R}.$$

(Traccia: a) segue dal teorema 5.18 o, direttamente, sia $A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$, allora

$$\rho^2(J) = \left|\frac{a_{12}a_{21}}{a_{11}a_{22}}\right| = \rho(G);$$

d) $\rho^2(J) = \rho(G) = \dfrac{1}{\alpha^2}$, quindi i metodi sono convergenti per $|\alpha| > 1$.)

5.17 Sia

$$A = \begin{bmatrix} \alpha & 1 & 1 \\ 1 & \alpha & 1 \\ 1 & 1 & \alpha \end{bmatrix}, \quad \alpha \in \mathbf{R}.$$

Si determini un valore di α per il quale A sia definita positiva, ma per cui il metodo di Jacobi non sia convergente.

(Risposta: α tale che $1 < \alpha < 2$.)

5.18 Sia $A \in \mathbf{C}^{n\times n}$, con $A = M - N$, dove

$$m_{ij} = \begin{cases} a_{ij} & \text{per } i = j \text{ oppure } i = j+1, \\ \\ 0 & \text{altrimenti.} \end{cases}$$

Per risolvere il sistema lineare $A\mathbf{x} = \mathbf{b}$, $\mathbf{b} \in \mathbf{C}^n$, si consideri il metodo iterativo

$$\mathbf{x}^{(k)} = P\mathbf{x}^{(k-1)} + \mathbf{q}, \quad k = 1, 2, \ldots,$$

dove $P = M^{-1}N$ e $\mathbf{q} = M^{-1}\mathbf{b}$.

a) Si descriva il metodo iterativo in termini di componenti;

b) si dimostri che la predominanza diagonale in senso stretto di A è condizione sufficiente per la convergenza.

(Traccia: a) se $a_{ii} \neq 0$ per $i = 1, 2, \ldots, n$, è

$$x_i^{(k)} = \frac{1}{a_{ii}}\left[b_i - \sum_{\substack{j=1 \\ j\neq i \\ j\neq i-1}}^{n} a_{ij}x_j^{(k-1)} - a_{i,i-1}x_{i-1}^{(k)}\right], \quad i = 1, 2, \ldots, n,$$

in cui l'ultimo termine nella parentesi non compare se $i = 1$; b) si proceda in modo analogo a quanto fatto nella dimostrazione del teorema 5.12.)

5.19 Dato il sistema $A\mathbf{x} = \mathbf{b}$, dove

$$A = \begin{bmatrix} 2 & 0 & 1 \\ 0 & 2 & 0 \\ -1 & 0 & 2 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix},$$

a) si determini il raggio spettrale delle matrici di iterazione dei metodi di Jacobi, di Gauss-Seidel e di rilassamento;

b) per il metodo di rilassamento si determini per quali valori di ω vi è convergenza e il valore ottimo ω_o;

c) si dica quante iterazioni occorrono per i tre metodi affinché le componenti dell'errore diventino in modulo minori di 1 se si sceglie $\mathbf{x}^{(0)} = [1025, 1025, 1025]^T$.

(Risposta: a) è

$$J = \begin{bmatrix} 0 & 0 & -\frac{1}{2} \\ 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 \end{bmatrix}, \quad G = \begin{bmatrix} 0 & 0 & -\frac{1}{2} \\ 0 & 0 & 0 \\ 0 & 0 & -\frac{1}{4} \end{bmatrix},$$

per cui $\rho(J) = \frac{1}{2}$, $\rho(G) = \frac{1}{4}$,

$$H(\omega) = \begin{bmatrix} 1-\omega & 0 & -\dfrac{\omega}{2} \\ 0 & 1-\omega & 0 \\ \dfrac{\omega}{2}(1-\omega) & 0 & 1-\omega-\dfrac{\omega^2}{4} \end{bmatrix},$$

per cui

$$\rho[H(\omega)] = \begin{cases} |1-\omega| & \text{per } 0 < \omega \leq \omega_o, \\ \dfrac{\omega}{8}[\omega + \sqrt{\omega^2 + 16(\omega-1)} + 8] - 1 & \text{per } \omega > \omega_o; \end{cases}$$

b) il metodo di rilassamento è convergente per $0 < \omega < \frac{4}{3}$, $\omega_o = 4\sqrt{5} - 8$; c) per il metodo di Jacobi occorrono 10 iterazioni, circa la metà per il metodo di Gauss-Seidel, circa un quarto per il metodo di rilassamento con $\omega = \omega_o$.)

5.20 Siano $A,\ B \in \mathbf{C}^{n\times n}$, $\mathbf{x}, \mathbf{b} \in \mathbf{C}^{2n}$ e si supponga che A sia non singolare e che la matrice $A^{-1}B$ abbia autovalori reali.

a) Si studi la convergenza del metodo di Jacobi a blocchi, di Gauss-Seidel a blocchi e di rilassamento a blocchi per la risoluzione del sistema

$$\begin{bmatrix} A & B \\ B & A \end{bmatrix} \mathbf{x} = \mathbf{b};$$

b) si esamini il caso particolare in cui $A = \alpha I_n,\ \alpha \in \mathbf{C}$.

(Traccia: a) si verifichi che $\det(J_B - \lambda I_{2n}) = \det[\lambda^2 I_n - (A^{-1}B)^2]$ e si applichi il teorema 5.30.)

5.21 Una matrice $A \in \mathbf{C}^{n\times n}$ si dice *consistentemente ordinata* se la sua matrice di iterazione di Jacobi $J = D^{-1}(B+C)$ ha gli stessi autovalori della matrice $J_\alpha = D^{-1}(\alpha B + \dfrac{1}{\alpha}C)$ per ogni costante $\alpha \in \mathbf{C}$, $\alpha \neq 0$.

a) Si dimostri che sono consistentemente ordinate le matrici appartenenti alle seguenti classi:

(1) matrici tridiagonali,

(2) matrici ad albero (per la definizione si veda l'esercizio 1.59).

b) Sia A la matrice $n \times n$ tridiagonale a blocchi di ordine m

$$A = \begin{bmatrix} B & -I_m & & \\ -I_m & B & \ddots & \\ & \ddots & \ddots & -I_m \\ & & -I_m & B \end{bmatrix}, \quad B = \begin{bmatrix} 4 & -1 & & \\ -1 & 4 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 4 \end{bmatrix}.$$

Si dimostri che A è consistentemente ordinata.

c) Si dimostri che se una matrice A è consistentemente ordinata, allora valgono le tesi dei teoremi 5.18 e 5.28.

d) Si esamini in particolare il caso

$$A = \begin{bmatrix} 9 & 8 & 10 \\ 1 & 4 & 0 \\ 3 & 0 & 5 \end{bmatrix}.$$

Anche la matrice A dell'esempio 5.19 è ad albero. Si dica perché in questo caso il valore ω_o trovato differisce da quello fornito dal teorema 5.28.

(Traccia: a) (1) per le matrici tridiagonali si veda la prima parte della dimostrazione del teorema 5.18, oppure si sfrutti la relazione ricorrente ottenuta all'esercizio 2.40; (2) per le matrici ad albero, si verifichi che $J_\alpha = D_\alpha J D_\alpha^{-1}$, dove D_α è la matrice diagonale i cui elementi principali sono $1, \alpha, \ldots, \alpha$.
b) Sia $H_k \in \mathbf{R}^{k \times k}$ la matrice bidiagonale simmetrica

$$H_k = \begin{bmatrix} 0 & -1 & & \\ -1 & 0 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 0 \end{bmatrix}.$$

Allora $J = \dfrac{1}{4}(I_n \otimes H_m + H_n \otimes I_m)$ (per la definizione e le proprietà del prodotto tensoriale si veda l'esercizio 1.60). Posto

$$S_k = \begin{bmatrix} 1 & & & \\ & \alpha & & \\ & & \ddots & \\ & & & \alpha^{k-1} \end{bmatrix} \in \mathbf{C}^{k \times k},$$

si ha $(S_n \otimes S_m)J(S_n \otimes S_m)^{-1}$

$$= \frac{1}{4}(S_n \otimes S_m)(I_n \otimes H_m + H_n \otimes I_m)(S_n \otimes S_m)^{-1}$$
$$= \frac{1}{4}[I_n \otimes (S_m H_m S_m^{-1}) + (S_n H_n S_n^{-1}) \otimes I_m)] = J_\alpha.$$

c) Nella dimostrazione dei teoremi 5.18 e 5.28 si usa in effetti solo l'ipotesi che A sia consistentemente ordinata; d) gli autovalori della matrice J sono $\mu_1 = 0$, $\mu_{2,3} = \pm\dfrac{\sqrt{8}}{3}$, quindi $\rho(J) = \dfrac{\sqrt{8}}{3}$; gli autovalori della matrice G sono $\lambda_{1,2} = 0$, $\lambda_3 = \dfrac{8}{9}$, quindi $\rho(G) = \dfrac{8}{9}$; è $\omega_o = \dfrac{3}{2}$ e $\rho[H(\omega_o)] = \dfrac{1}{2}$; per quanto riguarda la matrice dell'esempio 5.19, il valore di ω_o trovato non corrisponde a quello del teorema 5.28 perché gli autovalori di J non sono reali.)

5.22 Sia $A \in \mathbf{C}^{n\times n}$

$$A = \begin{bmatrix} a_1 & & & b_1 \\ b_2 & a_2 & & \\ & \ddots & \ddots & \\ & & b_n & a_n \end{bmatrix}.$$

a) Si dimostri che $\rho(G) = \rho^n(J)$;

b) si dimostri che per la matrice trasposta vale invece $\rho^{n-1}(G) = \rho^n(J)$;

c) seguendo la dimostrazione del teorema 5.18, si dimostri che se μ è autovalore non nullo di J e λ è autovalore non nullo di G, allora $\lambda = \mu^n$, ritrovando così il risultato ottenuto al punto a);

d) si scriva la relazione che lega gli autovalori non nulli di J agli autovalori non nulli λ della matrice di iterazione $H(\omega)$ del metodo di rilassamento.

(Traccia: a) si verifichi che

$$\rho(J) = \sqrt[n]{\prod_{i=1}^{n} \left|\frac{b_i}{a_i}\right|}.$$

e che $G = [O \mid \mathbf{u}]$, dove

$$\mathbf{u} = -b_1 \begin{bmatrix} a_1 & & & \\ b_2 & a_2 & & \\ & \ddots & \ddots & \\ & & b_n & a_n \end{bmatrix}^{-1} \mathbf{e}_1$$

e

$$\rho(G) = |u_n| = \prod_{i=1}^{n} \left| \frac{b_i}{a_i} \right|;$$

b) per il metodo di Jacobi è come per il punto a), per il metodo di Gauss-Seidel si verifichi che

$$G = - \begin{bmatrix} 0 & \frac{b_2}{a_1} & & & \\ 0 & 0 & \frac{b_3}{a_2} & & \\ \vdots & & \ddots & \ddots & \\ \vdots & & & \ddots & \frac{b_n}{a_{n-1}} \\ 0 & \frac{-b_1 b_2}{a_1 a_n} & & & 0 \end{bmatrix},$$

per cui

$$\rho(G) = \sqrt[n-1]{\prod_{i=1}^{n} \left| \frac{b_i}{a_i} \right|};$$

c) si verifichi che $SJS^{-1} = \alpha D^{-1}B + \frac{1}{\alpha^{n-1}} D^{-1}C$, per $\alpha \in \mathbf{C}$, $\alpha \neq 0$, dove S è la matrice definita nel teorema 5.18; d) seguendo la dimostrazione del teorema 5.28 si dimostri che

$$(\lambda + \omega - 1)^n = \lambda^{n-1} \mu^n \omega^n. \ )$$

5.23 Sia $A \in \mathbf{C}^{n \times n}$. Vale il seguente *teorema di Stein*:
$\rho(A) < 1$ se e solo se esiste una matrice B definita positiva tale che $B - A^H B A$ sia definita positiva.
(Traccia: sia $A\mathbf{x} = \lambda \mathbf{x}$, $\mathbf{x} \neq \mathbf{0}$. Se B e $B - A^H B A$ sono definite positive, allora

$$0 < \mathbf{x}^H (B - A^H B A)\mathbf{x} = (1 - |\lambda|^2)\mathbf{x}^H B \mathbf{x},$$

da cui $|\lambda| < 1$. Viceversa, se $\rho(A) < 1$ posto $B = \sum_{i=0}^{k} (A^i)^H A^i$, risulta $B - A^H B A = I - (A^{k+1})^H A^{k+1}$ che è definita positiva per k sufficientemente grande.)

5.24 Sia

$$\mathbf{x}^{(k)} = D^{-1}(B + C)\mathbf{x}^{(k-1)} + D^{-1}\mathbf{b}$$

il metodo di Jacobi per il sistema lineare $A\mathbf{x} = \mathbf{b}$. Si consideri il metodo di rilassamento applicato al metodo di Jacobi nel modo seguente

$$\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} + \omega(\mathbf{y}^{(k)} - \mathbf{x}^{(k-1)}), \quad \omega \in \mathbf{R},$$

dove $\mathbf{y}^{(k)} = D^{-1}(B + C)\mathbf{x}^{(k-1)} + D^{-1}\mathbf{b}$.

a) Si diano delle condizioni di convergenza del metodo;

b) se gli autovalori della matrice di Jacobi sono reali, si determini il valore ω_o di ω per cui il raggio spettrale della matrice di iterazione $H(\omega)$ è minimo e il corrispondente valore $H(\omega_o)$).

(Traccia: a) la matrice di iterazione è

$$H(\omega) = I - \omega D^{-1}A,$$

si ha convergenza se $|1 - \omega\lambda| < 1$, per ogni autovalore λ di $D^{-1}A$. Perciò condizione necessaria affinché esista un ω per cui il metodo è convergente è che i λ abbiano parte reale tutti dello stesso segno. Condizione sufficiente di convergenza è che ω abbia lo stesso segno e $|\omega| < 2 \min \dfrac{|Re\lambda|}{|\lambda|^2}$; b) si proceda in modo analogo a quanto fatto nell'esercizio 5.8, risulta $\omega_o = \dfrac{2}{\lambda_{\max} + \lambda_{\min}}$, dove $\lambda_{\max}$ e $\lambda_{\min}$ sono il massimo e il minimo modulo degli autovalori di $D^{-1}A$ e $\rho[(H(\omega_o)] = \dfrac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}}$.)

5.25 Sia $A \in \mathbf{R}^{n\times n}$. Vale il seguente teorema di *Perron-Frobenius* (per la dimostrazione si veda [10]). Se $A \geq O$ è una matrice irriducibile, allora

1) $\rho(A) > 0$ ed esiste un autovalore positivo di A di molteplicità algebrica 1 uguale a $\rho(A)$;
2) esiste $\mathbf{x} \in \mathbf{R}^n$, $\mathbf{x} > \mathbf{0}$, tale che $A\mathbf{x} = \rho(A)\mathbf{x}$;
3) se $B \in \mathbf{R}^{n\times n}$ è tale che $B \geq A$ e $B \neq A$, allora $\rho(B) > \rho(A)$;
4) se $A > O$, allora $\rho(A)$ è l'unico autovalore di modulo massimo.

Se $A \geq O$ non è irriducibile, allora la tesi si indebolisce, cioè

1) $\rho(A) \geq 0$ ed esiste un autovalore non negativo di A uguale a $\rho(A)$;
2) esiste $\mathbf{x} \in \mathbf{R}^n$, $\mathbf{x} \geq \mathbf{0}$, tale che $A\mathbf{x} = \rho(A)\mathbf{x}$;
3) se $B \in \mathbf{R}^{n\times n}$ è tale che $B \geq A$, allora $\rho(B) \geq \rho(A)$.

Inoltre per ogni matrice A vale la relazione $\rho(A) \leq \rho(|A|)$.

Sfruttando il teorema di Perron-Frobenius si dimostri che:

a) se la matrice A ha predominanza diagonale ed è irriducibile, allora il metodo di Gauss-Seidel è convergente.

b) Sia $A \geq O$ e irriducibile.

Se $\sum_{j=1}^{n} a_{ij} = \sigma$, per $i = 1, \ldots, n$, allora $\rho(A) = \sigma$; altrimenti posto

$$\alpha = \min_{i=1,\ldots,n} \sum_{j=1}^{n} a_{ij} \quad \text{e} \quad \beta = \max_{i=1,\ldots,n} \sum_{j=1}^{n} a_{ij},$$

è

$$\alpha < \rho(A) < \beta.$$

c) Sia $A^{(n)}$ la matrice di Hilbert di ordine n (si veda l'esempio 4.2). Si dimostri che $\rho(A^{(n)}) < \rho(A^{(n+1)})$.

(Traccia: a) dalla relazione $|G| \leq (I - |L|)^{-1}|U|$ (esercizio 5.15) segue per il teorema di Perron-Frobenius che $\rho(G) \leq \rho(|G|) \leq \rho[(I - |L|)^{-1}|U|]$; inoltre la matrice $|J| = |L| + |U|$ soddisfa al teorema di Stein-Rosenberg e poiché $\rho(|J|) < 1$ per l'ipotesi di predominanza e irriducibilità della matrice A, ne segue che $\rho[(I - |L|)^{-1}|U|] < \rho(|J|) < 1$. b) Per il primo caso si veda l'esercizio 3.16, per il secondo si considerino due matrici B e C tali che $O \leq B \leq A$, $C \geq A$, $B,\ C \neq A$, e

$$\sum_{j=1}^{n} b_{ij} = \alpha, \quad \sum_{j=1}^{n} c_{ij} = \beta, \quad \text{per} \quad i = 1, \ldots, n,$$

dove ad elementi nulli di B e di C corrispondono elementi nulli di A. B e C risultano quindi irriducibili e non negative. Si ha $\rho(B) = \alpha$, $\rho(C) = \beta$ e per il teorema di Perron-Frobenius applicato ad A e a B risulta $\rho(B) < \rho(A) < \rho(C)$. c) Si consideri la matrice $B_\epsilon \in \mathbf{R}^{(n+1)\times(n+1)}$ i cui elementi sono

$$b_{ij}^{(\epsilon)} = \begin{cases} a_{ij}^{(n)} & \text{se } i,\ j = 1, \ldots, n, \\ \dfrac{1}{2n+1} & \text{se } i = j = n+1, \\ \epsilon & \text{altrimenti.} \end{cases}$$

Allora se $0 < \epsilon < \dfrac{1}{2n+1}$, risulta $B_\epsilon \leq A^{(n+1)}$, $B_\epsilon \neq A^{(n+1)}$ e per il teorema di Perron-Frobenius applicato a B_ϵ è $\rho(B_\epsilon) < \rho(A^{(n+1)})$; vale poi $\rho(A^{(n)}) \leq \rho(B_0) \leq \rho(B_\epsilon)$.)

5.26 La decomposizione (5) della matrice A non singolare nella forma

$$A = M - N, \quad \det M \neq 0,$$

viene detta *partizionamento regolare* se $M^{-1} \geq O$ e $N \geq O$. Si dimostri che

a) se $A^{-1} \geq O$ e se la decomposizione (5) è un partizionamento regolare, allora la matrice di iterazione $P = M^{-1}N$ del metodo iterativo (9) è tale che
$$\rho(P) = \frac{\rho(A^{-1}N)}{1 + \rho(A^{-1}N)} < 1,$$
e quindi il metodo iterativo è convergente;

b) se $A^{-1} \geq O$, e se $A = M_1 - N_1$ e $A = M_2 - N_2$ sono due partizionamenti regolari della matrice A, tali che $N_2 \geq N_1$, allora
$$\rho(M_2^{-1}N_2) \geq \rho(M_1^{-1}N_1),$$
e quindi il metodo iterativo corrispondente al primo partizionamento ha una velocità asintotica maggiore od uguale a quello corrispondente al secondo partizionamento. Si esaminino in particolare i metodi di Jacobi a blocchi e a elementi e di Gauss-Seidel a blocchi e a elementi per il sistema (67) dell'esempio 5.36.

(Traccia: a) poiché
$$P = M^{-1}N = (I + A^{-1}N)^{-1}A^{-1}N,$$
gli autovalori di P sono della forma $\lambda = \dfrac{\mu}{1+\mu}$, in cui μ è un autovalore di $A^{-1}N$; inoltre per ipotesi P e $A^{-1}N$ sono matrici non negative e quindi per il teorema di Perron-Frobenius (esercizio 5.25) hanno entrambe un autovalore non negativo (positivo se $A^{-1}N$ è irriducibile) uguale al loro raggio spettrale; b) è $A^{-1}N_2 \geq A^{-1}N_1$, e per il teorema di Perron-Frobenius è $\rho(A^{-1}N_2) \geq \rho(A^{-1}N_1)$ e quindi se μ_1 e μ_2 sono autovalori non negativi di $A^{-1}N_1$ e $A^{-1}N_2$, da $\mu_2 \geq \mu_1$ segue che
$$\lambda_2 = \frac{\mu_2}{1+\mu_2} \geq \frac{\mu_1}{1+\mu_1} = \lambda_1. \ )$$

5.27 Sia $B \in \mathbf{R}^{n\times n}$, $B \geq O$. Si dimostri che le seguenti condizioni sono equivalenti:

(a) $\mu > \rho(B)$,

(b) $\det(\mu I - B) \neq 0$ e $(\mu I - B)^{-1} \geq O$,

(c) esiste $\mathbf{x} \in \mathbf{R}^n$, $\mathbf{x} > \mathbf{0}$ tale che $\mu\mathbf{x} > B\mathbf{x}$,

(d) gli autovalori delle sottomatrici principali di $\mu I - B$ hanno parte reale positiva,

(e) gli autovalori di $\mu I - B$ hanno parte reale positiva,

(f) i determinanti delle sottomatrici principali di $\mu I - B$ sono positivi,

(g) i determinanti delle sottomatrici principali di testa di $\mu I - B$ sono positivi.

Le matrici A tali che $A = \mu I - B$, $B \geq O$, sono dette *M-matrici* e intervengono nella risoluzione numerica di certe equazioni differenziali ellittiche. I metodi di Jacobi e di Gauss-Seidel per tali matrici corrispondono a partizionamenti regolari (si veda l'esercizio 5.26).

(Traccia: $(a) \Rightarrow (b)$ poiché è $\mu > \rho(B)$ è $\det(\mu I - B) \neq 0$ e vale

$$\mu I - B = \mu\Big[I - \frac{1}{\mu}B\Big], \quad \rho\Big(\frac{1}{\mu}B\Big) < 1,$$

quindi per il teorema 5.4

$$(\mu I - B)^{-1} = \frac{1}{\mu}\sum_{i=0}^{\infty}\Big(\frac{1}{\mu}B\Big)^i \geq O;$$

$(b) \Rightarrow (c)$ essendo $(\mu I - B)^{-1} \geq O$, è $\mathbf{x} = (\mu I - B)^{-1}\mathbf{y} > \mathbf{0}$, dove $\mathbf{y} > \mathbf{0}$. Risulta allora $\mathbf{x} > \mathbf{0}$ e $(\mu I - B)\mathbf{x} = \mathbf{y} > \mathbf{0}$;
$(c) \Rightarrow (d)$ sia D la matrice diagonale con elementi principali $x_1, \ldots, x_n$; ogni sottomatrice principale di $D^{-1}(\mu I - B)D$ è simile alla corrispondente sottomatrice principale di $\mu I - B$, e vale $D^{-1}(\mu I - B)D\mathbf{e} > \mathbf{0}$, dove $\mathbf{e} = [1, 1, \ldots, 1]^T$. Cioè risulta $\mu > b_{ii}$ e $D^{-1}(\mu I - B)D$ ha predominanza diagonale in senso stretto, insieme con tutte le sue sottomatrici principali. Si applichi il teorema 2.41;
$(d) \Rightarrow (e)$ ovvio;
$(e) \Rightarrow (a)$ poiché $Re(\mu - \rho(B)) > 0$, risulta $\mu > \rho(B)$;
$(d) \Rightarrow (f)$ segue dal fatto che il segno del determinante di una matrice reale è dato dal segno del prodotto degli autovalori reali;
$(f) \Rightarrow (g)$ ovvio;
$(g) \Rightarrow (b)$ si dimostri per induzione su n utilizzando l'espressione dell'inversa e del determinante di B in termini del complemento di Schur (si veda l'esercizio 1.43).)

5.28 a) Si verifichi che la funzione

$$f(x) = \alpha x + \frac{\beta}{x} + \gamma, \quad 0 < m < x < M, \quad \alpha,\ \beta > 0,$$

non ha punti di massimo interni all'intervallo (m, M);

b) si verifichi che la funzione

$$f(x_1, \ldots, x_n) = \sum_{i=1}^{n} a_i x_i \ \sum_{i=1}^{n} \frac{a_i}{x_i}, \quad 0 < m < x_i < M, \quad a_i > 0,$$

ha come punti di massimo solo punti le cui componenti sono uguali a m oppure a M e che

$$f(x_1, \ldots, x_n) \leq \left(\sum_{i=1}^{n} a_i \right)^2 \frac{(m+M)^2}{4mM} .$$

c) Sia $A \in \mathbf{C}^{n \times n}$ definita positiva e siano $\lambda_{\max}$ e $\lambda_{\min}$ il massimo e il minimo dei suoi autovalori. Si dimostri la seguente *disuguaglianza di Kantorovich*

$$\frac{(\mathbf{x}^H \mathbf{x})^2}{(\mathbf{x}^H A \mathbf{x})\ (\mathbf{x}^H A^{-1} \mathbf{x})} \geq \frac{4\lambda_{\max}\lambda_{\min}}{(\lambda_{\max} + \lambda_{\min})^2}, \quad \text{per ogni } \mathbf{x} \in \mathbf{C}^n.$$

d) Siano $A \in \mathbf{R}^{n \times n}$ definita positiva, $\mathbf{x}^*$ soluzione del sistema $A\mathbf{x} = \mathbf{b}$, $\{\mathbf{x}_k\}$, $k = 1, 2, \ldots$, una successione di punti ottenuta con il metodo dello steepest descent. Si dimostri che, indicato con $\mathbf{e}_k = \mathbf{x}^* - \mathbf{x}_k$, risulta

$$\mathbf{e}_{k+1}^T A \mathbf{e}_{k+1} \leq \left(\frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}} \right)^2 \mathbf{e}_k^T A \mathbf{e}_k,$$

dove $\lambda_{\max}$ e $\lambda_{\min}$ sono il massimo e il minimo degli autovalori di A.

(Traccia: b) Fissato i, $1 \leq i \leq n$, si consideri $f(x_1, \ldots, x_n)$ come funzione della sola x_i e si applichi il punto a); quindi i punti di massimo della funzione $f(x_1, \ldots, x_n)$ possono avere solo componenti uguali a m o M, si supponga allora che un punto di massimo sia tale che $x_1 = \ldots = x_k = m$ e $x_{k+1} = \ldots = x_n = M$. Posto

$$s_1 = \sum_{i=1}^{k} a_i, \quad s_2 = \sum_{i=k+1}^{n} a_i,$$

si ha

$$\begin{aligned} f(x_1, \ldots, x_n) &\leq (s_1 m + s_2 M)\Big(\frac{s_1}{m} + \frac{s_2}{M}\Big) = (s_1 + s_2)^2 + \frac{(m-M)^2}{mM}\ s_1 s_2 \\ &\leq (s_1 + s_2)^2 \Big(1 + \frac{(m-M)^2}{4mM}\Big); \end{aligned}$$

c) siano $\lambda_1, \ldots, \lambda_n$ gli autovalori di A e $\mathbf{u}_1, \ldots, \mathbf{u}_n$ i corrispondenti autovettori ortonormali, si può esprimere

$$\mathbf{x} = \sum_{i=1}^{n} c_i \mathbf{u}_i,$$

e quindi

$$\mathbf{x}^H\mathbf{x} = \sum_{i=1}^{n} |c_i|^2, \quad A\mathbf{x} = \sum_{i=1}^{n} c_i\lambda_i\mathbf{u}_i,$$

$$\mathbf{x}^H A\mathbf{x} = \sum_{i=1}^{n} |c_i|^2\lambda_i, \quad \mathbf{x}^H A^{-1}\mathbf{x} = \sum_{i=1}^{n} \frac{|c_i|^2}{\lambda_i}.$$

Ne segue

$$\frac{(\mathbf{x}^H\mathbf{x})^2}{(\mathbf{x}^H A\mathbf{x})\ (\mathbf{x}^H A^{-1}\mathbf{x})} = \frac{\Big(\sum\limits_{i=1}^{n} |c_i|^2\Big)^2}{\sum\limits_{i=1}^{n} |c_i|^2\lambda_i \ \sum\limits_{i=1}^{n} \dfrac{|c_i|^2}{\lambda_i}}.$$

La disuguaglianza segue dal punto b), perché $\lambda_{\min} \leq \lambda_i \leq \lambda_{\max}$, $i = 1, \ldots, n$;

d) è $\quad \mathbf{e}_{k+1} = \mathbf{e}_k - \alpha_k\mathbf{r}_k = \mathbf{e}_k - \dfrac{\mathbf{r}_k^T\mathbf{r}_k}{\mathbf{r}_k^T A\mathbf{r}_k}\ \mathbf{r}_k, \quad A\mathbf{e}_k = A\mathbf{x}^* - A\mathbf{x}_k = \mathbf{r}_k,$

e quindi

$$\mathbf{e}_{k+1}^T A\mathbf{e}_{k+1} = \mathbf{e}_k^T A\mathbf{e}_k - \frac{(\mathbf{r}_k^T\mathbf{r}_k)^2}{\mathbf{r}_k^T A\mathbf{r}_k}.$$

Inoltre $\mathbf{e}_k = A^{-1}\mathbf{r}_k$, per cui

$$\mathbf{e}_k^T A\mathbf{e}_k = \mathbf{r}_k^T A^{-1}\mathbf{r}_k,$$

e

$$\mathbf{e}_{k+1}^T A\mathbf{e}_{k+1} = \mathbf{e}_k^T A\mathbf{e}_k \left[1 - \frac{(\mathbf{r}_k^T\mathbf{r}_k)^2}{(\mathbf{r}_k^T A\mathbf{r}_k)(\mathbf{r}_k^T A^{-1}\mathbf{r}_k)}\right].$$

Si applichi poi la disuguaglianza di Kantorovich.)

5.29 Sia $A \in \mathbf{R}^{n\times n}$ definita positiva e siano $\mathbf{p}_1, \ldots, \mathbf{p}_n$, n autovettori ortonormali di A. Si dimostri che

a) i vettori $\mathbf{p}_1, \ldots, \mathbf{p}_n$ sono A-coniugati;

b) se λ_i è l'autovalore corrispondente all'autovettore $\mathbf{p}_i$, è $\lambda_i = \mathbf{p}_i^T A\mathbf{p}_i$;

c) se P è la matrice le cui colonne sono i vettori $\mathbf{p}_1, \ldots, \mathbf{p}_n$, allora le colonne della matrice AP sono vettori A^{-1}-coniugati;

d) se $A = LDL^T$, in cui L è una matrice triangolare inferiore con elementi principali uguali a 1, allora le colonne di L sono vettori A^{-1}-coniugati e le colonne di L^{-T} sono vettori A-coniugati;

e) si determini una n-upla di vettori A-coniugati della matrice tridiagonale di ordine n

$$A = \begin{bmatrix} 2 & -1 & & \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{bmatrix}.$$

(Traccia: d) poiché $L^{-1}AL^{-T} = D$, è $\mathbf{l}_i^T A \mathbf{l}_j = d_{ij}$, dove $\mathbf{l}_i$ e $\mathbf{l}_j$ sono la i-esima e la j-esima colonna di L^{-T}; e) A è definita positiva, $A = LDL^T$, dove

$$L = \begin{bmatrix} 1 & & & & \\ -\frac{1}{2} & 1 & & & \\ & -\frac{2}{3} & 1 & & \\ & & \ddots & \ddots & \\ & & & -\frac{n-1}{n} & 1 \end{bmatrix}, \quad D = \begin{bmatrix} 2 & & & & \\ & \frac{3}{2} & & & \\ & & \frac{4}{3} & & \\ & & & \ddots & \\ & & & & \frac{n+1}{n} \end{bmatrix},$$

i vettori cercati sono le colonne L^{-T}.)

5.30 Sia $A \in \mathbf{R}^{n\times n}$ definita positiva. Si consideri la successione di matrici

$$B_0 = O, \quad B_{k+1} = B_k + \frac{\mathbf{p}_k \mathbf{p}_k^T}{\mathbf{p}_k^T A \mathbf{p}_k}, \quad k = 0, \ldots, n-1,$$

in cui i vettori $\mathbf{p}_k$ sono A-coniugati. Si dimostri che $B_n = A^{-1}$, e quindi

$$A^{-1} = \sum_{k=0}^{n-1} \frac{\mathbf{p}_k \mathbf{p}_k^T}{\mathbf{p}_k^T A \mathbf{p}_k}.$$

Si applichi questa relazione alla matrice dell'esempio 5.29 e), con $n = 5$.

(Traccia: per $k \geq 1$ è $B_k A \mathbf{p}_j = \mathbf{p}_j$ per $j = 0, \ldots, k-1$, e quindi $B_n A = I$.)

5.31 Sia $A \in \mathbf{R}^{n\times n}$ definita positiva e siano $\mathbf{p}_k$ e $\mathbf{r}_k$ le direzioni e i residui generati dal metodo del gradiente coniugato a partire da un vettore $\mathbf{x}_0$ arbitrario. Si dimostri che per ogni $k \geq 1$ per cui $\mathbf{r}_k \neq \mathbf{0}$ è

a) $\|\mathbf{p}_k\|_2^2 > \|\mathbf{r}_k\|_2^2$; b) $\mathbf{p}_k^T A \mathbf{p}_k < \mathbf{r}_k^T A \mathbf{r}_k$;

c) $\lambda_{\min} < \dfrac{1}{\alpha_k} < \lambda_{\max}$, dove $\lambda_{\min}$ e $\lambda_{\max}$ sono il minimo e il massimo autovalore di A;

d) l'angolo fra le direzioni $\mathbf{p}_k$ e $\mathbf{p}_j$, per $j < k$, è acuto.

(traccia: a) si utilizzi la (51), tenendo conto che $\mathbf{r}_k^T \mathbf{p}_{k-1} = 0$; b) utilizzando la (51) e la (53) si ottiene

$$\mathbf{p}_k^T A \mathbf{p}_k = \mathbf{r}_k^T A \mathbf{r}_k - \beta_k^2 \mathbf{p}_{k-1}^T A \mathbf{p}_{k-1};$$

c) si dimostri che per ogni vettore $\mathbf{x} \neq \mathbf{0}$ è $\lambda_{\min} \leq \dfrac{\mathbf{x}^T A \mathbf{x}}{\mathbf{x}^T \mathbf{x}} \leq \lambda_{\max}$, quindi

$$\frac{1}{\alpha_k} = \frac{\mathbf{p}_k^T A \mathbf{p}_k}{\mathbf{r}_k^T \mathbf{r}_k} > \frac{\mathbf{p}_k^T A \mathbf{p}_k}{\mathbf{p}_k^T \mathbf{p}_k} \geq \lambda_{\min}.$$

D'altra parte

$$\frac{1}{\alpha_k} < \frac{\mathbf{r}_k^T A \mathbf{r}_k}{\mathbf{r}_k^T \mathbf{r}_k} \leq \lambda_{\max};$$

d) si dimostri che $\mathbf{p}_j^T \mathbf{r}_k = 0$ per $j < k$, quindi per la (51) è

$$\mathbf{p}_j^T \mathbf{p}_k = \beta_k \beta_{k-1} \dots \beta_{j+1} \mathbf{p}_j^T \mathbf{p}_j > 0,$$

essendo per la (57) $\beta_i > 0$ per $i = 1, \dots, k$.)

5.32 Sia $A \in \mathbf{R}^{n \times n}$ definita positiva, con m autovettori distinti $\lambda_1, \dots, \lambda_m$. Si applichi ad A il metodo del gradiente coniugato a partire da un vettore $\mathbf{x}_0$ e si supponga che $m \leq n$ sia il minimo intero tale che $\mathbf{r}_m = \mathbf{0}$. Allora il polinomio $s_m(\lambda)$ costruito con la relazione ricorrente

$$\left.\begin{array}{l} s_0(\lambda) = q_0(\lambda) = 1, \\ s_k(\lambda) = s_{k-1}(\lambda) - \alpha_{k-1} \lambda q_{k-1}(\lambda), \\ q_k(\lambda) = s_k(\lambda) + \beta_k q_{k-1}(\lambda), \end{array}\right\} \quad k = 1, \dots, m,$$

è dato da

$$s_m(\lambda) = (-1)^m \alpha_0 \alpha_1 \dots \alpha_{m-1} p(\lambda),$$

dove $p(\lambda)$ è il polinomio minimo di A.

(Traccia: si dimostri per induzione che

$$\mathbf{r}_k = s_k(A)\mathbf{r}_0 \quad \text{e} \quad \mathbf{p}_k = q_k(A)\mathbf{r}_0,$$

per cui $s_m(A)\mathbf{r}_0 = \mathbf{0}$. Indicati con $\mathbf{u}_1, \dots, \mathbf{u}_n$, n autovettori linearmente indipendenti di A, è

$$\mathbf{r}_0 = \sum_{i=1}^{n} c_i \mathbf{u}_i,$$

da cui, essendo solo m gli autovalori distinti, segue che

$$A^k\mathbf{r}_0 = \sum_{i=1}^{n} \lambda_i^k c_i \mathbf{u}_i = \sum_{j=1}^{m} \lambda_j^k d_j \mathbf{x}_j,$$

in cui i vettori $\mathbf{x}_j$, $j = 1, \ldots, m$, sono linearmente indipendenti, e quindi

$$s_m(A)\mathbf{r}_0 = \sum_{j=1}^{m} s_m(\lambda_j) d_j \mathbf{x}_j,$$

da cui $s_m(\lambda_j) = 0$ per $j = 1, \ldots, m$. Il polinomio $s_m(\lambda)$ ha grado m e primo coefficiente $(-1)^m \alpha_0 \alpha_1 \ldots \alpha_{m-1}$.)

5.33 Sia $A \in \mathbf{R}^{n\times n}$ definita positiva e siano $\mathbf{p}_k$ le direzioni generate dal metodo del gradiente coniugato a partire da un vettore $\mathbf{x}_0$ arbitrario.

a) Si dimostri che i k vettori $\mathbf{p}_0, \ldots, p_{k-1}$ sono linearmente indipendenti;

b) sia S_k lo spazio generato dai k vettori $\mathbf{p}_0, \ldots, \mathbf{p}_{k-1}$. Posto per semplicità $\mathbf{x}_0 = \mathbf{0}$ (se non lo fosse basterebbe considerare il sistema "traslato" $A(\mathbf{x} - \mathbf{x}_0) = \mathbf{b} - A\mathbf{x}_0$), si dimostri che i punti $\mathbf{x}_k$ determinati dal metodo del gradiente coniugato sono tali che

$$\Phi(\mathbf{x}_k) = \min_{\mathbf{x}\in S_k} \Phi(\mathbf{x}).$$

(Traccia: a) se per assurdo fosse $\mathbf{p}_j = \sum_{\substack{i=0 \\ i\neq j}}^{k-1} \gamma_i \mathbf{p}_i$ con γ_i non tutti nulli, sarebbe

$$\mathbf{p}_j^T A \mathbf{p}_j = \sum_{\substack{i=0 \\ i\neq j}}^{k-1} \gamma_i \mathbf{p}_j^T A \mathbf{p}_i^T = 0;$$

b) si indichi con

$$\mathbf{x} = \sum_{i=0}^{k-1} \gamma_i \mathbf{p}_i$$

un punto di S_k, allora

$$\Phi(\mathbf{x}) = \tfrac{1}{2}\Big(\sum_{i=0}^{k-1} \gamma_i \mathbf{p}_i\Big)^T A \Big(\sum_{i=0}^{k-1} \gamma_i \mathbf{p}_i\Big) - \mathbf{b}^T \Big(\sum_{i=0}^{k-1} \gamma_i \mathbf{p}_i\Big)$$

e poiché le direzioni $\mathbf{p}_i$ sono A-coniugate, si ha

$$\Phi(\mathbf{x}) = \tfrac{1}{2}\sum_{i=0}^{k-1}\gamma_i^2\mathbf{p}_i^T A\mathbf{p}_i - \mathbf{b}^T\Big(\sum_{i=0}^{k-1}\gamma_i\mathbf{p}_i\Big)$$

$$= \Big[\tfrac{1}{2}\sum_{i=0}^{k-2}\gamma_i^2\mathbf{p}_i^T A\mathbf{p}_i - \mathbf{b}^T\Big(\sum_{i=0}^{k-2}\gamma_i\mathbf{p}_i\Big)\Big] + \Big[\tfrac{1}{2}\gamma_{k-1}^2\mathbf{p}_{k-1}^T A\mathbf{p}_{k-1} - \gamma_{k-1}\mathbf{b}^T\mathbf{p}_{k-1}\Big].$$

Il problema di trovare il minimo di $\Phi(\mathbf{x})$ su S_k risulta così decomposto in due problemi indipendenti:

$$\min_{\mathbf{x}\in S_k}\Phi(\mathbf{x}) = \min_{\mathbf{x}\in S_{k-1}}\Phi(\mathbf{x}) + \min_{\gamma\in\mathbf{R}}\tfrac{1}{2}\,\gamma^2\mathbf{p}_{k-1}^T A\mathbf{p}_{k-1} - \gamma\mathbf{b}^T\mathbf{p}_{k-1}.$$

Procedendo per induzione, sia $\mathbf{x}_{k-1}$ il punto di minimo su S_{k-1} e γ_{k-1} il punto di minimo del problema

$$\min_{\gamma\in\mathbf{R}}\tfrac{1}{2}\,\gamma^2\mathbf{p}_{k-1}^T A\mathbf{p}_{k-1} - \gamma\mathbf{b}^T\mathbf{p}_{k-1},$$

cioè

$$\gamma_{k-1} = \frac{\mathbf{b}^T\mathbf{p}_{k-1}}{\mathbf{p}_{k-1}^T A\mathbf{p}_{k-1}},$$

allora

$$\mathbf{x}_k = \mathbf{x}_{k-1} + \gamma_{k-1}\mathbf{p}_{k-1}.$$

Poiché

$$\mathbf{b} = \mathbf{r}_{k-1} + A\mathbf{x}_{k-1},$$

risulta

$$\mathbf{b}^T\mathbf{p}_{k-1} = (\mathbf{r}_{k-1} + A\mathbf{x}_{k-1})^T\mathbf{p}_{k-1} = \mathbf{r}_{k-1}^T\mathbf{p}_{k-1} + \Big(\sum_{i=0}^{k-2}\gamma_i A\mathbf{p}_i\Big)^T\mathbf{p}_{k-1},$$

e poiché le direzioni $\mathbf{p}_i$ sono A-coniugate, è

$$\mathbf{b}^T\mathbf{p}_{k-1} = \mathbf{r}_{k-1}^T\mathbf{p}_{k-1}.$$

L'espressione per γ_{k-1} risulta quindi coincidere con l'espressione per α_{k-1} data dalla (46). Ne segue che il punto $\mathbf{x}_k$ è proprio quello determinato dalla (45).)

Commento bibliografico

I metodi iterativi scaturiscono in generale dalla necessità di risolvere sistemi lineari di grosse dimensioni. Inizialmente tali sistemi si presentarono nel campo dell'astronomia e più specificatamente in problemi di minimi quadrati applicati ad osservazioni astronomiche. Gauss nel 1823 ebbe l'idea di scomporre il sistema in due o più sistemi, bloccando temporaneamente alcune variabili, di risolvere questi sottosistemi e poi di sostituire le approssimazioni così ottenute nuovamente nel sistema di partenza. Successivamente Jacobi, interessandosi a problemi legati allo studio di sistemi fisici sottoposti a piccole oscillazioni, riprese i risultati di Gauss sui minimi quadrati e nel 1845 propose il metodo che porta il suo nome. Jacobi interessò all'argomento anche il suo ex-allievo Seidel, che nel 1874 descrisse un metodo derivato da quello usato da Gauss.

In seguito i metodi iterativi, anche prima che vi fossero dei precisi teoremi di convergenza, sono stati estensivamente usati, inizialmente nella scuola tedesca, poi dalle altre. I primi lavori sulla convergenza del metodo di Gauss-Seidel risalgono al 1885 e al 1892, ad opera rispettivamente di Nekrasov e di Memke, mentre è del 1929 il lavoro di Von Mises e Pollazek-Geiringer, in cui si esaminano comparativamente le condizioni sufficienti per la convergenza dei metodi di Jacobi e di Gauss-Seidel. Il teorema 5.2 fu dimostrato da Hensel nel 1926 e riscoperto indipendentemente da Oldenburger nel 1940. Di questo teorema esistono altre dimostrazioni che non utilizzano la forma normale di Jordan di una matrice; nel 1958 Householder ne dette una che utilizza solo le proprietà delle norme di matrici (si veda l'esercizio 5.1). La prima dimostrazione completa del teorema che lega la convergenza del metodo di Gauss-Seidel con le matrici definite positive è quella di Reich del 1949, anche se dimostrazioni parziali erano già state pubblicate prima. La dimostrazione riportata in questo testo si basa su quella di Ostrowski del 1954.

La possibilità di accelerare la convergenza dei metodi iterativi introducendo dei parametri è stata presa in considerazione fin dal 1910 da Richardson e successivamente altri autori hanno applicato la tecnica al metodo di Gauss-Seidel. Nel 1950 Frankel e Young, separatamente, stabilirono relazioni fra le velocità di convergenza dei metodi di Jacobi, Gauss-Seidel e di rilassamento.

Attualmente i metodi iterativi sono usati soprattutto per risolvere sistemi di grosse dimensioni che si incontrano nella risoluzione di problemi alle derivate parziali: si tratta in generale di sistemi con matrici sparse e fortemente strutturate, per cui esiste un'ampia teoria. Nel libro di Varga [10] del 1962, in cui è riportata una trattazione sistematica della materia, la dimostrazione dei teoremi di convergenza dei metodi iterativi si basa sulla teoria delle matrici non negative introdotta da Perron e Frobenius nel

1907 e nel 1912 (si veda l'esercizio 5.25); parte dei teoremi sono dimostrati senza utilizzare il teorema di Perron-Frobenius nel libro di Stoer [9]. Una trattazione più recente e completa sui metodi iterativi e sulle tecniche di accelerazione della convergenza è riportata nel libro di Hageman e Young [5] del 1981, in cui sono esposti molti schemi di algoritmi e vengono esaminati anche i problemi pratici che sorgono nella effettiva implementazione dei metodi. Nel testo di Young [12] del 1971 si trova uno studio più approfondito dei metodi del rilassamento e della scelta dell'ω ottimo.

Il metodo dello steepest descent è dovuto a Cauchy (1847), che lo sviluppò per risolvere sistemi di equazioni non lineari. Il metodo del gradiente coniugato fu descritto nel 1952 da Hestenes e Stiefel [6], ma per quanto destasse subito un certo interesse nell'ambiente matematico, esso non divenne di uso corrente per almeno una ventina di anni, fino a quando in un lavoro di Reid [8] del 1971 si suggerì di usarlo per la risoluzione di sistemi di grosse dimensioni a matrice sparsa, come metodo iterativo, superando così la difficoltà della perdita di ortogonalità dei residui. Lo studio della velocità di convergenza del metodo e delle tecniche di precondizionamento è stato ed è oggetto di molti lavori: per problemi particolari sono state proposte delle tecniche di precondizionamento efficaci (si veda ad esempio il lavoro di Concus, Golub e Meurant [2]).

La risoluzione numerica del problema di Dirichlet, detto *problema modello*, è un esempio classico di approssimazione numerica della soluzione di un problema ellittico. Uno studio dei metodi alle differenze per le equazioni differenziali alle derivate parziali è riportato nel libro di Isaacson e Keller [7]; confronti fra i metodi iterativi utilizzati per risolvere il problema modello si trovano sul libro di Young [12]. Un testo recente sul confronto fra le prestazioni dei metodi diretti, dei metodi iterativi e di molti altri metodi ad hoc per risolvere problemi più generali, che si incontrano nel trattamento numerico delle equazioni alle derivate parziali di tipo ellittico, è quello di Birkhoff e Lynch [1].

Bibliografia

[1] G. Birkhoff, R. E. Lynch, *Numerical Solution of Elliptic Problems*, SIAM Studies in Applied Mathematics, Philadelphia, 1984.

[2] P. Concus, G. H. Golub, G. Meurant, "Block Preconditioning for the Conjugate Gradient Method", *SIAM J. Sci. Stat. Comput.*, 6, 1985, pp. 220-252.

[3] D. K. Faddeev, V. N. Faddeeva, *Computational Methods of Linear Algebra*, Freeman and Co., San Francisco, 1963.

[4] G. H. Golub, C. F. Van Loan, *Matrix Computations*, 2nd Edition, The Johns Hopkins University Press, Baltimore, Maryland, 1989.

[5] L. A Hageman, D. M. Young, *Applied Iterative Methods*, Academic Press, New York, 1981.

[6] M. R. Hestenes, E. Stiefel, "Methods of Conjugate Gradients for Solving Linear Systems", *J. Res. Nat. Bur. Stand.*, 49, 1952, pp. 409-436.

[7] E. Isaacson, H. B. Keller, *Analysis of Numerical Methods*, J. Wiley & Sons, New York, 1966.

[8] J. K. Reid, "On the Method of Conjugate Gradients for the Solution of Large Sparse Systems of Linear Equations", in *Large Sparse Sets of Linear Equations*, ed. J. K. Reid, Academic Press, New York, 1971, pp. 231-254.

[9] J. Stoer, R. Bulirsch, *Introduction to Numerical Analysis*, Springer-Verlag, New York, 1980.

[10] R. S. Varga, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, New Jersey, 1962.

[11] J. H. Wilkinson, C. Reinsch, *Handbook for Automatic Computation, vol. 2, Linear Algebra*, Springer-Verlag, New York, 1971.

[12] D. M. Young, *Iterative Solution of Large Linear Systems*, Academic Press, New York, 1971.

Capitolo 6

METODI PER IL CALCOLO DI AUTOVALORI E AUTOVETTORI

1. Teoremi di localizzazione

Poiché gli autovalori di una matrice A sono gli zeri del suo polinomio caratteristico $P(\lambda) = \det(A - \lambda I)$, il calcolo numerico degli autovalori può essere effettuato applicando un qualsiasi metodo di iterazione funzionale (metodo delle tangenti, delle secanti, ecc.) all'equazione $P(\lambda) = 0$. Questo modo di procedere può essere conveniente se sono disponibili dei metodi efficienti per il calcolo del valore che la funzione $P(\lambda)$ (ed eventualmente la sua derivata prima) assume in un punto, come nel caso che la matrice abbia alcune proprietà di struttura (si veda il paragrafo 6 per il caso in cui la matrice A sia tridiagonale). Un'altra possibilità potrebbe essere quella di calcolare i coefficienti del polinomio caratteristico e poi applicare un metodo numerico per la risoluzione dell'equazione $P(\lambda) = 0$. Anche se il calcolo dei coefficienti del polinomio caratteristico ha lo stesso costo asintotico della moltiplicazione di matrici, cioè $O(n^{2.38})$, questo modo di procedere non è conveniente essenzialmente per due motivi: da una parte il costo computazionale rimane comunque elevato anche per valori grandi di n, dall'altra perché gli errori di arrotondamento generati nel calcolo dei coefficenti di $P(\lambda)$ possono indurre elevate variazioni degli zeri del polinomio. Quest'ultimo inconveniente non si presenta nel caso di matrici hermitiane, considerando gli autovalori come funzioni degli elementi della matrice (si veda il paragrafo 2). È comunque evidente che in generale il calcolo numerico degli autovalori di una matrice può essere fatto solamente con un procedimento iterativo. In questo capitolo verranno descritti i principali metodi numerici per il calcolo degli autovalori di una matrice.

Inizialmente verranno esposti alcuni teoremi di *localizzazione* che permettono di determinare facilmente sottoinsiemi del piano complesso in cui si trovano gli autovalori. Di questi teoremi i più importanti sono i teoremi di Gershgorin 2.35, 2.37 e 2.38, che per la loro generalità e semplicità di applicazione sono stati anticipati nel secondo capitolo.

6.1 Teorema (di Hirsch). *Sia $A \in \mathbf{C}^{n\times n}$ e sia $\| \cdot \|$ una qualsiasi norma matriciale indotta. Allora il cerchio*

$$\{\ z \in \mathbf{C}\ :\ |z| \leq \|A\|\ \}$$

contiene tutti gli autovalori di A.

Dim. La tesi segue dal teorema 3.10. ∎

6.2 Teorema. *Siano $A \in \mathbf{C}^{n\times n}$ normale, $\mathbf{x} \in \mathbf{C}^n, \mathbf{x} \neq \mathbf{0}$ e f una funzione razionale definita su un sottoinsieme del piano complesso contenente gli autovalori di A. Allora esiste almeno un autovalore λ di A tale che*

$$|f(\lambda)| \leq \frac{\|f(A)\mathbf{x}\|_2}{\|\mathbf{x}\|_2}. \tag{1}$$

Dim. Poiché A è normale, per il teorema 2.28 esiste una matrice unitaria U tale che

$$A = UDU^H,$$

dove D è la matrice diagonale il cui i-esimo elemento principale è uguale a λ_i, autovalore di A, per $i = 1, \ldots, n$. Quindi

$$\|A\mathbf{x}\|_2 = \|UDU^H\mathbf{x}\|_2 = \|DU^H\mathbf{x}\|_2,$$

in quanto U è unitaria. Posto $\mathbf{y} = U^H\mathbf{x}$, è $\|\mathbf{y}\|_2 = \|\mathbf{x}\|_2$ e

$$\frac{\|A\mathbf{x}\|_2}{\|\mathbf{x}\|_2} = \frac{\|D\mathbf{y}\|_2}{\|\mathbf{y}\|_2} = \sqrt{\frac{\sum_{j=1}^{n} |d_{jj}y_j|^2}{\sum_{j=1}^{n} |y_j|^2}} \geq \sqrt{\frac{\min_{i=1,\ldots,n} |\lambda_i|^2 \sum_{j=1}^{n} |y_j|^2}{\sum_{j=1}^{n} |y_j|^2}} = \min_{i=1,\ldots,n} |\lambda_i|.$$

Per la (20) del capitolo 2 è

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

e quindi

$$\frac{\|f(A)\mathbf{x}\|_2}{\|\mathbf{x}\|_2} = \frac{\|Uf(D)U^H\mathbf{x}\|_2}{\|\mathbf{x}\|_2} = \frac{\|f(D)\mathbf{y}\|_2}{\|\mathbf{y}\|_2} \geq \min_{i=1,\ldots,n} |f(\lambda_i)|. \quad ∎$$

La (1) può essere utilizzata per determinare delle maggiorazioni a posteriori dell'errore che si commette approssimando gli autovalori di una matrice normale.

Sia ad esempio σ un'approssimazione di un autovalore di una matrice A normale e sia $\mathbf{x}$ un vettore che approssima l'autovettore corrispondente. Allora se

$$f(z) = z - \sigma,$$

dalla (1) si ha che esiste un autovalore λ di A tale che

$$|\lambda - \sigma| \leq \frac{\|(A - \sigma I)\mathbf{x}\|_2}{\|\mathbf{x}\|_2}; \tag{2}$$

se invece

$$f(z) = \frac{z - \sigma}{z},$$

e A è non singolare, dalla (1) si ha che esiste un autovalore λ di A tale che

$$\left|\frac{\lambda - \sigma}{\lambda}\right| \leq \frac{\|(A - \sigma I)A^{-1}\mathbf{x}\|_2}{\|\mathbf{x}\|_2};$$

e ponendo $A^{-1}\mathbf{x} = \mathbf{z}$ è

$$\left|\frac{\lambda - \sigma}{\lambda}\right| \leq \frac{\|(A - \sigma I)\mathbf{z}\|_2}{\|A\mathbf{z}\|_2}. \tag{3}$$

La (2) e la (3) danno una stima facilmente calcolabile dell'errore assoluto e relativo che si commette assumendo σ come approssimazione dell'autovalore λ, e possono essere anche usate come criterio di arresto per i metodi iterativi che approssimano gli autovalori.

Utilizzando il teorema 6.2 si dimostra un teorema di localizzazione in cui interviene il *quoziente di Rayleigh* di una matrice A relativo ad un vettore $\mathbf{x} \neq \mathbf{0}$:

$$r_A(\mathbf{x}) = \frac{\mathbf{x}^H A\mathbf{x}}{\mathbf{x}^H\mathbf{x}}. \tag{4}$$

6.3 Teorema (di Weinstein). *Siano $A \in \mathbf{C}^{n\times n}$ normale e $\mathbf{x} \in \mathbf{C}^n, \mathbf{x} \neq \mathbf{0}$. Allora esiste almeno un autovalore λ di A nel cerchio*

$$\left\{ z \in \mathbf{C} \ : \ |z - r_A(\mathbf{x})| \leq \sqrt{\frac{\mathbf{x}^H A^H A\mathbf{x}}{\mathbf{x}^H\mathbf{x}} - |r_A(\mathbf{x})|^2} \right\}.$$

Dim. Dal teorema 6.2, ponendo

$$f(z) = z - r_A(\mathbf{x}),$$

risulta che esiste un autovalore λ di A tale che

$$|\lambda - r_A(\mathbf{x})| \leq \frac{\|[A - r_A(\mathbf{x})I]\mathbf{x}\|_2}{\|\mathbf{x}\|_2} = \sqrt{\frac{\mathbf{x}^H[A^H - \overline{r_A(\mathbf{x})}\ I]\ [A - r_A(\mathbf{x})\ I]\ \mathbf{x}}{\mathbf{x}^H\mathbf{x}}}$$

$$= \sqrt{\frac{\mathbf{x}^H A^H A\mathbf{x} - r_A(\mathbf{x})\mathbf{x}^H A^H\mathbf{x} - \overline{r_A(\mathbf{x})}\mathbf{x}^H A\mathbf{x} + |r_A(\mathbf{x})|^2\mathbf{x}^H\mathbf{x}}{\mathbf{x}^H\mathbf{x}}},$$

da cui, poiché $\dfrac{\mathbf{x}^H A^H\mathbf{x}}{\mathbf{x}^H\mathbf{x}} = \overline{r_A(\mathbf{x})}$, segue che

$$|\lambda - r_A(\mathbf{x})| \leq \sqrt{\frac{\mathbf{x}^H A^H A\mathbf{x}}{\mathbf{x}^H\mathbf{x}} - |r_A(\mathbf{x})|^2}.$$

∎

2. Teoremi di perturbazione

In questo paragrafo, in modo analogo a quanto fatto nel primo paragrafo del capitolo 4 per il problema della risoluzione dei sistemi lineari, si studia il condizionamento del problema del calcolo degli autovalori di una matrice, cioè si analizza la variazione indotta sugli autovalori da una perturbazione degli elementi della matrice. Tali risultati permettono di valutare l'errore inerente del problema del calcolo degli autovalori, generato dalla rappresentazione dei dati con un numero finito di cifre.

6.4 Teorema (di Bauer-Fike). *Sia* $\| \cdot \|$ *una norma matriciale indotta che verifichi la seguente proprietà*

$$\|D\| = \max_{i=1,\ldots,n} |d_{ii}|$$

per ogni matrice diagonale $D \in \mathbf{C}^{n\times n}$ *(una tale norma viene detta norma assoluta, le norme* $\| \cdot \|_1$, $\| \cdot \|_2$ *e* $\| \cdot \|_\infty$ *sono assolute). Sia* $A \in \mathbf{C}^{n\times n}$ *una matrice diagonalizzabile, cioè tale che*

$$A = TDT^{-1},$$

con D *diagonale e* T *non singolare. Se* $\delta A \in \mathbf{C}^{n\times n}$ *e* ξ *è un autovalore di* $A + \delta A$*, allora esiste almeno un autovalore* λ *di* A *tale che*

$$|\lambda - \xi| \leq \mu(T)\ \|\delta A\|,$$

dove $\mu(T) = \|T\|\ \|T^{-1}\|$.

Dim. Se ξ fosse autovalore di A, la tesi sarebbe verificata. Altrimenti la matrice $A - \xi I$ risulta non singolare e dalla relazione

$$(A + \delta A)\mathbf{y} = \xi\mathbf{y},$$

dove $\mathbf{y}$ è autovettore di $A + \delta A$, si ha

$$\delta A\mathbf{y} = -(A - \xi I)\mathbf{y},$$

da cui

$$(A - \xi I)^{-1}\delta A\mathbf{y} = -\mathbf{y}$$

e quindi

$$\|(A - \xi I)^{-1}\delta A\| \geq 1. \tag{5}$$

Poiché

$$(A - \xi I)^{-1} = T(D - \xi I)^{-1}T^{-1},$$

si ha dalla (5)

$$1 \leq \|T(D - \xi I)^{-1}T^{-1}\delta A\| \leq \|T\| \ \|T^{-1}\| \ \|(D - \xi I)^{-1}\| \ \|\delta A\|,$$

e poiché $\| \ . \ \|$ è una norma assoluta, ne segue

$$1 \leq \mu(T)\frac{1}{\min\limits_{i=1,\ldots,n} |\lambda_i - \xi|} \ \|\delta A\|, \tag{6}$$

in cui i λ_i, $i = 1, \ldots, n$, sono gli autovalori di A e quindi gli elementi principali di D. Dalla (6) segue che

$$\min_{i=1,\ldots,n} |\lambda_i - \xi| \leq \mu(T)\|\delta A\|,$$

da cui la tesi. ∎

Il teorema 6.4 esprime un risultato di perturbazione: perturbando gli elementi di una matrice A, gli autovalori cambiano al più proporzionalmente all'entità della perturbazione δA. Il condizionamento del problema del calcolo degli autovalori di A è legato al numero di condizionamento della matrice T le cui colonne sono gli autovettori di A: quindi il problema del calcolo degli autovalori è tanto meglio condizionato quanto più basso è il numero di condizionamento $\mu(T)$. Se A è una matrice normale, allora T è unitaria, per cui $\mu_2(T) = 1$ e dal teorema 6.4 si ha

$$|\lambda - \xi| \leq \|\delta A\|_2,$$

ossia il problema del calcolo degli autovalori per matrici normali è ben condizionato per tutti gli autovalori.

Per matrici non normali il problema del calcolo degli autovalori può essere ben condizionato o mal condizionato, a seconda delle proprietà dell'autovalore considerato. Per questa ragione è bene analizzare il comportamento del problema per un singolo autovalore, distinguendo il caso di un autovalore di molteplicità algebrica uno dal caso di un autovalore di molteplicità algebrica maggiore di uno.

6.5 Teorema. *Sia $A \in \mathbf{C}^{n\times n}$, λ un autovalore di A di molteplicità algebrica uno, $\mathbf{x}, \mathbf{y} \in \mathbf{C}^n$, $\|\mathbf{x}\|_2 = \|\mathbf{y}\|_2 = 1$, tali che*

$$A\mathbf{x} = \lambda\mathbf{x}$$

$$\mathbf{y}^H A = \lambda\mathbf{y}^H.$$

Allora è $\mathbf{y}^H\mathbf{x} \neq 0$ ed inoltre per ogni $F \in \mathbf{C}^{n\times n}$ esiste nel piano complesso un intorno V dello zero e una funzione $\lambda(\epsilon) : V \to \mathbf{C}$, analitica, tale che

a) $\lambda(\epsilon)$ *è autovalore con molteplicità algebrica uno di* $A + \epsilon F$,

b) $\lambda(0) = \lambda$,

c) $\lambda'(0) = \dfrac{\mathbf{y}^H F\mathbf{x}}{\mathbf{y}^H\mathbf{x}}$,

d) a meno dei termini di ordine superiore in ϵ *è*

$$\lambda(\epsilon) - \lambda = \epsilon\, \frac{\mathbf{y}^H F\mathbf{x}}{\mathbf{y}^H\mathbf{x}}\ .$$

Per la dimostrazione si veda [26] (si veda anche l'esercizio 6.9). ∎

Anche in questo caso risulta che la variazione nell'autovalore dovuta alla perturbazione ϵF di A è proporzionale ad ϵ. Inoltre il condizionamento del problema dipende dalla quantità

$$\left|\frac{\mathbf{y}^H F\mathbf{x}}{\mathbf{y}^H\mathbf{x}}\right|,$$

che, data F, è tanto più grande quanto più piccolo è $|\mathbf{y}^H\mathbf{x}|$. Nel caso delle matrici normali è $\mathbf{y}^H\mathbf{x} = 1$, in accordo con i risultati del teorema 6.4.

Se λ è un autovalore di molteplicità algebrica $\sigma(\lambda) > 1$ e di molteplicità geometrica $\tau(\lambda)$, a cui corrispondono i blocchi di Jordan

$$C^{(1)}, C^{(2)}, \ldots, C^{(\tau(\lambda))},$$

di ordine massimo η, allora si può dimostrare che esiste un intorno V di zero e una costante $\gamma > 0$, tale che per $\epsilon \in V$ la matrice $A + \epsilon F$ ha autovalori $\lambda_i(\epsilon),\ i = 1, \ldots, \sigma(\lambda)$, tali che

$$|\lambda_i(\epsilon) - \lambda| \leq \gamma\, |\epsilon|^{1/\eta}.$$

Se $\eta > 1$, il problema del calcolo dell'autovalore λ può essere fortemente mal condizionato.

6.6 Esempio. Siano

$$A = \begin{bmatrix} \mu & 1 & 0 & 0 \\ 0 & \mu & 1 & 0 \\ 0 & 0 & \mu & 1 \\ 0 & 0 & 0 & \mu \end{bmatrix}, \quad F = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix},$$

allora

$$A + \epsilon F = \begin{bmatrix} \mu & 1 & 0 & 0 \\ 0 & \mu & 1 & 0 \\ 0 & 0 & \mu & 1 \\ \epsilon & 0 & 0 & \mu \end{bmatrix}.$$

Si ha

$$\det(A + \epsilon F - \lambda I) = (\mu - \lambda)^4 - \epsilon,$$

da cui risulta che gli autovalori λ_j di $A + \epsilon F$ soddisfano alla relazione

$$|\lambda_j - \mu| = \sqrt[4]{|\epsilon|}, \quad j = 1, \ldots, 4.$$

Quindi ad una perturbazione ϵ dell'elemento a_{41} corrisponde una variazione di modulo $\sqrt[4]{|\epsilon|}$ negli autovalori. Se ad esempio fosse $\epsilon = 10^{-8}$ (inferiore alla precisione di macchina quando si opera con 6 cifre significative esadecimali), risulterebbe

$$|\lambda_j - \mu| = 10^{-2}, \quad j = 1, \ldots, 4. \qquad \blacksquare$$

3. Caso delle matrici hermitiane

Nel caso delle matrici hermitiane il quoziente di Rayleigh (4) assume una notevole importanza, in quanto è legato a proprietà interessanti, utili anche per il calcolo. È opportuno richiamare alcune proprietà riguardanti i sottospazi introdotti nei paragrafi 2 e 6 del capitolo 1.

1 - Se S è un sottospazio di $\mathbf{C}^n$, allora il sottospazio ortogonale $S^\perp$ ha dimensione

$$\dim S^\perp = n - \dim S. \tag{7}$$

2 - Se S e T sono due sottospazi di $\mathbf{C}^n$, allora per il sottospazio $S \cap T$ vale

$$\dim(S \cap T) \geq \max \{0, \dim S + \dim T - n\}. \tag{8}$$

3 - Se $A \in \mathbf{C}^{m \times n}, m \leq n$, e S è un sottospazio di $\mathbf{C}^n$, allora per le dimensioni dei sottospazi

$$T = \{ \mathbf{x} \in \mathbf{C}^m \text{ tali che } \mathbf{x} = A\mathbf{y}, \mathbf{y} \in S \}$$

e

$$N = \{ \mathbf{x} \in S \text{ tali che } A\mathbf{x} = \mathbf{0} \}$$

vale la relazione

$$\dim T + \dim N = \dim S. \tag{9}$$

6.7 Teorema (di Courant-Fischer o del minimax). *Sia $A \in \mathbf{C}^{n\times n}$ una matrice hermitiana con autovalori*

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n.$$

Allora risulta

$$\lambda_{n-k+1} = \min_{V_k} \max_{\substack{\mathbf{x}\neq\mathbf{0} \\ \mathbf{x}\in V_k}} r_A(\mathbf{x}), \tag{10}$$

$$\lambda_k = \max_{V_k} \min_{\substack{\mathbf{x}\neq\mathbf{0} \\ \mathbf{x}\in V_k}} r_A(\mathbf{x}), \tag{11}$$

dove V_k è un qualunque sottospazio di $\mathbf{C}^n$ di dimensione k, per $k = 1, \ldots, n$.

Dim. Siano $\mathbf{x}_i, i = 1, \ldots, n$, autovettori ortonormali di A corrispondenti agli autovalori λ_i e, fissato un indice k, sia S il sottospazio di dimensione $n-k+1$ generato dagli $n-k+1$ vettori $\mathbf{x}_k, \ldots, \mathbf{x}_n$. Per la (8) è

$$\dim(S \cap V_k) \geq 1$$

e quindi l'intersezione fra S e V_k non può ridursi al solo vettore nullo. Sia allora

$$\mathbf{x} = \sum_{i=k}^{n} \alpha_i \mathbf{x}_i \neq \mathbf{0}$$

elemento di $S \cap V_k$. Poiché i vettori $\mathbf{x}_i$ sono ortonormali e vale

$$A\mathbf{x}_i = \lambda_i \mathbf{x}_i, \quad i = 1, \ldots, n,$$

allora

$$r_A(\mathbf{x}) = \frac{\mathbf{x}^H A\mathbf{x}}{\mathbf{x}^H\mathbf{x}} = \frac{\sum_{j=k}^{n} |\alpha_j|^2 \lambda_j}{\sum_{j=k}^{n} |\alpha_j|^2} \leq \frac{\max_{i=k,\ldots,n} \lambda_i \sum_{j=k}^{n} |\alpha_j|^2}{\sum_{j=k}^{n} |\alpha_j|^2} = \max_{i=k,\ldots,n} \lambda_i = \lambda_k.$$

Quindi si ha

$$\min_{\substack{\mathbf{x}\neq\mathbf{0} \\ \mathbf{x}\in V_k}} r_A(\mathbf{x}) \leq \lambda_k. \tag{12}$$

D'altra parte, se V_k è proprio lo spazio generato da $\mathbf{x}_1, \ldots, \mathbf{x}_k$, il vettore $\mathbf{x}_k$ è elemento di V_k e vale

$$r_A(\mathbf{x}_k) = \lambda_k,$$

quindi nella (12) vale il segno di uguaglianza, da cui segue la (11). Per dimostrare la (10) è sufficiente applicare la (11) alla matrice $-A$. ∎

Si osservi che dal teorema del minimax si ottiene in particolare

$$\lambda_1 = \max_{\mathbf{x}\neq\mathbf{0}} \; r_A(\mathbf{x}),$$

$$\lambda_n = \min_{\mathbf{x}\neq\mathbf{0}} \; r_A(\mathbf{x}).$$

Inoltre è facile verificare che se A è una matrice reale simmetrica, allora la funzione $f \; : \; \mathbf{R}^n \to \mathbf{R}$,

$$f(\mathbf{x}) = r_A(\mathbf{x})$$

è stazionaria nel punto $\mathbf{v} \in \mathbf{R}^n$ se e solo se

$$A\mathbf{v} = \lambda\mathbf{v}, \; \lambda = f(\mathbf{v}).$$

Dal teorema del minimax seguono i seguenti teoremi.

6.8 Teorema. *Sia $A \in \mathbf{C}^{n\times n}$ hermitiana, e $U \in \mathbf{C}^{n\times(n-1)}$, tale che $U^H U = I_{n-1}$. Sia inoltre $B = U^H AU$. Allora per gli autovalori λ_i, $i = 1, \ldots, n$, di A con $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$, e μ_i, $i = 1, \ldots, n-1$, di B con $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_{n-1}$, vale la relazione*

$$\lambda_1 \geq \mu_1 \geq \lambda_2 \geq \mu_2 \geq \ldots \geq \mu_{n-1} \geq \lambda_n.$$

Tale proprietà viene anche espressa dicendo che gli autovalori di B separano gli autovalori di A.

Dim. Si dimostra dapprima che per $k = 2, \ldots, n$ è

$$\lambda_k \leq \mu_{k-1}. \tag{13}$$

Dalla (11) segue che esiste un sottospazio Z_k di $\mathbf{C}^n$ di dimensione k tale che

$$\lambda_k = \min_{\substack{\mathbf{x}\neq\mathbf{0} \\ \mathbf{x}\in Z_k}} \; r_A(\mathbf{x}). \tag{14}$$

Sia S il sottospazio di dimensione $n-1$ generato dalle colonne di U. Per la (8) è

$$\dim(Z_k \cap S) \geq k - 1,$$

per cui, poiché $k \geq 2$, esistono vettori $\mathbf{x} \neq \mathbf{0}$ appartenenti a $Z_k \cap S$ e quindi dalla (14)

$$\lambda_k \leq \min_{\substack{\mathbf{x}\neq\mathbf{0} \\ \mathbf{x}\in Z_k\cap S}} \; r_A(\mathbf{x}).$$

Per ogni vettore $\mathbf{x} \in S$, con $\mathbf{x} \neq \mathbf{0}$, esiste un solo vettore $\mathbf{y} \in \mathbf{C}^{n-1},\ \mathbf{y} \neq \mathbf{0}$, tale che

$$\mathbf{x} = U\mathbf{y},$$

per cui

$$\mathbf{y} = U^H\mathbf{x}.$$

Si considera allora il sottospazio

$$W = \{\ \mathbf{y} \in \mathbf{C}^{n-1} \text{ tali che } \mathbf{y} = U^H\mathbf{x},\ \mathbf{x} \in Z_k \cap S\ \}.$$

Poiché $U^HU = I$, il nucleo di U^H

$$N = \{\ \mathbf{x} \in S \text{ tali che } U^H\mathbf{x} = \mathbf{0}\ \}$$

ha dimensione nulla; per la (9) risulta

$$\dim W = \dim(Z_k \cap S) \geq k - 1$$

e quindi

$$\lambda_k \leq \min_{\substack{\mathbf{x}\neq\mathbf{0}\\ \mathbf{x}\in Z_k\cap S}} r_A(\mathbf{x}) = \min_{\substack{\mathbf{x}\neq\mathbf{0}\\ \mathbf{x}\in Z_k\cap S}} \frac{\mathbf{x}^HA\mathbf{x}}{\mathbf{x}^H\mathbf{x}} = \min_{\substack{\mathbf{y}\neq\mathbf{0}\\ \mathbf{y}\in W}} \frac{\mathbf{y}^HU^HAU\mathbf{y}}{\mathbf{y}^HU^HU\mathbf{y}}$$

$$= \min_{\substack{\mathbf{y}\neq\mathbf{0}\\ \mathbf{y}\in W}} \frac{\mathbf{y}^HB\mathbf{y}}{\mathbf{y}^H\mathbf{y}} \leq \max_{V_{k-1}} \min_{\substack{\mathbf{y}\neq\mathbf{0}\\ \mathbf{y}\in V_{k-1}}} r_B(\mathbf{y}) = \mu_{k-1},$$

dove V_{k-1} è un qualunque sottospazio di $\mathbf{C}^{n-1}$ di dimensione $k-1$, da cui segue la (13).

Applicando la (13) alle matrici $-A$ e $-B$, i cui autovalori sono

$$-\lambda_n \geq -\lambda_{n-1} \geq \ldots \geq -\lambda_1$$

e

$$-\mu_{n-1} \geq -\mu_{n-2} \geq \ldots \geq -\mu_1,$$

si ha

$$-\lambda_{n+1-k} \leq -\mu_{(n-1)+1-(k-1)},\ \text{per } k = 2, \ldots, n,$$

e quindi

$$\lambda_{n+1-k} \geq \mu_{n+1-k},\ \text{per } k = 2, \ldots, n,$$

cioè

$$\lambda_i \geq \mu_i,\ \text{per } i = 1, \ldots, n-1.$$

∎

6.9 Teorema. *Sia $A \in \mathbf{C}^{n\times n}$ hermitiana e per $m \le n$ sia $U_m \in \mathbf{C}^{n\times m}$ tale che $U_m^H U_m = I_m$. Allora indicati con λ_1 e λ_n il massimo e il minimo autovalore di A e con μ_1 e μ_m il massimo e il minimo autovalore di $U_m^H A U_m$, valgono le relazioni*

$$\lambda_1 \ge \mu_1, \qquad \mu_m \ge \lambda_n.$$

Dim. Sia $B = U_m^H A U_m$. Si ha

$$\mu_1 = \max_{\substack{\mathbf{y}\neq\mathbf{0}\\ \mathbf{y}\in\mathbf{C}^m}} r_B(\mathbf{y}) = \max_{\substack{\mathbf{y}\neq\mathbf{0}\\ \mathbf{y}\in\mathbf{C}^m}} \frac{\mathbf{y}^H U_m^H A U_m \mathbf{y}}{\mathbf{y}^H\mathbf{y}} = \max_{\substack{\mathbf{x}\neq\mathbf{0}\\ \mathbf{x}=U_m\mathbf{y}\\ \mathbf{y}\in\mathbf{C}^m}} \frac{\mathbf{x}^H A\mathbf{x}}{\mathbf{x}^H\mathbf{x}}$$

$$\le \max_{\substack{\mathbf{x}\neq\mathbf{0}\\ \mathbf{x}\in\mathbf{C}^n}} r_A(\mathbf{x}) = \lambda_1.$$

Analogamente si procede per μ_m. ∎

6.10 Teorema. *Sia $A \in \mathbf{C}^{n\times n}$ hermitiana, e sia A_k la sottomatrice principale di testa di ordine k di A. Allora gli autovalori di A_k separano gli autovalori di A_{k+1}, per $k = 1, \ldots, n-1$.*

Dim. Si osservi che se

$$U = \begin{bmatrix} I_k \\ \mathbf{0}^H \end{bmatrix} \begin{matrix} \} & k \text{ righe} \\ \} & 1 \text{ riga} \end{matrix}$$

allora $U^H U = I_k$ e $A_k = U^H A_{k+1} U$. La tesi segue dal teorema 6.8. ∎

6.11 Esempio. Come illustrazione del teorema 6.10 si consideri la matrice hermitiana

$$A = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 & 2 \\ 1 & 2 & 3 & 3 & 3 \\ 1 & 2 & 3 & 4 & 4 \\ 1 & 2 & 3 & 4 & 5 \end{bmatrix}.$$

Gli autovalori delle sottomatrici A_k principali di testa di ordine k di A sono

k	λ_1	λ_2	λ_3	λ_4	λ_5
1	1				
2	2.618033	0.3819660			
3	5.048913	0.6431029	0.3079774		
4	8.290849	1.	0.4260219	0.2831172	
5	12.34352	1.448682	0.5829639	0.3532520	0.2715528

Gli autovalori sono stati calcolati con il metodo di Jacobi (si veda il paragrafo 9). ∎

Facendo ancora uso del teorema del minimax si dimostrano i seguenti risultati.

6.12 Teorema. *Sia* $\mathbf{u} \in \mathbf{C}^n, \sigma \in \mathbf{R}, \sigma \geq 0$, *e siano* $A, B \in \mathbf{C}^{n\times n}$ *hermitiane, tali che*

$$B = A + \sigma \mathbf{u}\mathbf{u}^H.$$

Per gli autovalori

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$$

di A, *e*

$$\mu_1 \geq \mu_2 \geq \ldots \geq \mu_n$$

di B, *vale la relazione*

$$\lambda_1 + \sigma \mathbf{u}^H \mathbf{u} \geq \mu_1 \geq \lambda_1 \geq \mu_2 \geq \lambda_2 \geq \ldots \geq \lambda_{n-1} \geq \mu_n \geq \lambda_n.$$

Dim. Se $\mathbf{u} = \mathbf{0}$, la tesi è banale; si suppone allora che $\mathbf{u} \neq \mathbf{0}$ e si dimostra che $\mu_i \geq \lambda_i$, per $i = 1, \ldots, n$. Dalla (10) si ha che per $k = 1, \ldots, n$, esiste un sottospazio Z_k di dimensione k, tale che

$$\mu_{n-k+1} = \max_{\substack{\mathbf{x}\neq\mathbf{0} \\ \mathbf{x}\in Z_k}} r_B(\mathbf{x}) = \max_{\substack{\mathbf{x}\neq\mathbf{0} \\ \mathbf{x}\in Z_k}} \frac{\mathbf{x}^H B\mathbf{x}}{\mathbf{x}^H\mathbf{x}} = \max_{\substack{\mathbf{x}\neq\mathbf{0} \\ \mathbf{x}\in Z_k}} \left[\frac{\mathbf{x}^H A\mathbf{x}}{\mathbf{x}^H\mathbf{x}} + \sigma \frac{|\mathbf{x}^H\mathbf{u}|^2}{\mathbf{x}^H\mathbf{x}} \right]$$

$$\geq \max_{\substack{\mathbf{x}\neq\mathbf{0} \\ \mathbf{x}\in Z_k}} \frac{\mathbf{x}^H A\mathbf{x}}{\mathbf{x}^H\mathbf{x}} \geq \min_{V_k} \max_{\substack{\mathbf{x}\neq\mathbf{0} \\ \mathbf{x}\in V_k}} r_A(\mathbf{x}) = \lambda_{n-k+1}.$$

Inoltre dalla (10) si ha che esiste un sottospazio W_k di dimensione k tale che

$$\lambda_{n-k+1} = \max_{\substack{\mathbf{x}\neq\mathbf{0} \\ \mathbf{x}\in W_k}} r_A(\mathbf{x}).$$

Sia S il sottospazio di $\mathbf{C}^n$ generato dal vettore $\mathbf{u}$ e sia $\mathbf{x}$ un vettore del sottospazio $T = W_k \cap S^\perp$, cioè tale che

$$\mathbf{x}^H\mathbf{u} = 0.$$

Allora è

$$\mathbf{x}^H B\mathbf{x} = \mathbf{x}^H A\mathbf{x} + \sigma|\mathbf{x}^H\mathbf{u}|^2 = \mathbf{x}^H A\mathbf{x}$$

e quindi

$$\lambda_{n-k+1} \geq \max_{\substack{\mathbf{x}\neq\mathbf{0} \\ \mathbf{x}\in T}} r_A(\mathbf{x}) = \max_{\substack{\mathbf{x}\neq\mathbf{0} \\ \mathbf{x}\in T}} r_B(\mathbf{x}). \tag{15}$$

Poiché $\dim S = 1$, per la (7) è $\dim S^{\perp} = n - 1$, e quindi per la (8) è $\dim T \geq k - 1$. Dalla (15) segue che

$$\lambda_{n-k+1} \geq \min_{V_{k-1}} \max_{\substack{\mathbf{x}\neq\mathbf{0} \\ \mathbf{x}\in V_{k-1}}} r_B(\mathbf{x}) = \mu_{n-k+2}.$$

Inoltre dal teorema del minimax si ottiene

$$\mu_1 = \max_{\mathbf{x}\neq\mathbf{0}} r_B(\mathbf{x}) = \max_{\mathbf{x}\neq\mathbf{0}} \frac{\mathbf{x}^H B\mathbf{x}}{\mathbf{x}^H\mathbf{x}} = \max_{\mathbf{x}\neq\mathbf{0}} \left[\frac{\mathbf{x}^H A\mathbf{x}}{\mathbf{x}^H\mathbf{x}} + \sigma \frac{|\mathbf{x}^H\mathbf{u}|^2}{\mathbf{x}^H\mathbf{x}}\right],$$

e poiché per la disuguaglianza di Cauchy-Schwarz (1) cap. 1, è

$$\max_{\mathbf{x}\neq\mathbf{0}} \frac{|\mathbf{x}^H\mathbf{u}|^2}{\mathbf{x}^H\mathbf{x}} = \mathbf{u}^H\mathbf{u},$$

ne segue che

$$\mu_1 \leq \max_{\mathbf{x}\neq\mathbf{0}} r_A(\mathbf{x}) + \sigma\mathbf{u}^H\mathbf{u} = \lambda_1 + \sigma\mathbf{u}^H\mathbf{u}. \qquad \blacksquare$$

6.13 Esempio. La matrice

$$A = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 & 2 \\ 1 & 2 & 3 & 3 & 3 \\ 1 & 2 & 3 & 4 & 4 \\ 1 & 2 & 3 & 4 & 5 \end{bmatrix}$$

ha gli autovalori $\lambda_1 = 12.34352$, $\lambda_2 = 1.448682$, $\lambda_3 = 0.5829639$, $\lambda_4 = 0.3532520$, $\lambda_5 = 0.2715528$ (si veda l'esempio 6.11).

Se $\sigma = 1$ e $\mathbf{u} = [1,\ \text{-1},\ 0,\ \text{-1},\ 0]^T$, la matrice $B = A + \sigma\mathbf{u}\mathbf{u}^T$ ha gli autovalori $\mu_1 = 12.96301$, $\mu_2 = 2.736083$, $\mu_3 = 1.448030$, $\mu_4 = 0.5076391$, $\mu_5 = 0.3451974$, che separano gli autovalori di A e che sono tali che $\mu_i > \lambda_i$, per $i = 1, \ldots, 5$, e $\mu_1 < \lambda_1 + \sigma\mathbf{u}^H\mathbf{u} = \lambda_1 + 3$.

Se $\sigma = -2$, la matrice $C = A + \sigma\mathbf{u}\mathbf{u}^T$ ha gli autovalori $\eta_1 = 11.65525$, $\eta_2 = 1.448380$, $\eta_3 = 0.5175053$, $\eta_4 = 0.3456874$, $\eta_5 = -4.966838$, che separano ancora gli autovalori di A e sono tali che $\eta_i < \lambda_i$, per $i = 1, \ldots, 5$, e $\eta_5 > \lambda_5 + \sigma\mathbf{u}^H\mathbf{u} = \lambda_5 - 6$. $\blacksquare$

6.14 Teorema. *Siano $A, B, C \in \mathbf{C}^{n\times n}$ hermitiane, tali che $C = A + B$. Per gli autovalori*

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$$

di A,

$$\mu_1 \geq \mu_2 \geq \ldots \geq \mu_n$$

di B,

$$\nu_1 \geq \nu_2 \geq \ldots \geq \nu_n$$

di C, vale la relazione

$$\lambda_k + \mu_n \leq \nu_k \leq \lambda_k + \mu_1, \quad k = 1, \ldots, n.$$

Dim. Dalla (11) si ha per $k = 1, \ldots, n$

$$\nu_k = \max_{V_k} \min_{\substack{\mathbf{x}\neq\mathbf{0} \\ \mathbf{x}\in V_k}} r_C(\mathbf{x}) = \max_{V_k} \min_{\substack{\mathbf{x}\neq\mathbf{0} \\ \mathbf{x}\in V_k}} [r_A(\mathbf{x}) + r_B(\mathbf{x})]$$

$$\geq \max_{V_k} \min_{\substack{\mathbf{x}\neq\mathbf{0} \\ \mathbf{x}\in V_k}} [r_A(\mathbf{x}) + \mu_n] = \max_{V_k} \min_{\substack{\mathbf{x}\neq\mathbf{0} \\ \mathbf{x}\in V_k}} r_A(\mathbf{x}) + \mu_n$$

$$= \lambda_k + \mu_n.$$

L'altra relazione si ricava applicando lo stesso procedimento alla matrice $A = C - B$, per cui si ottiene

$$\lambda_k \geq \nu_k + (-\mu_1).$$ ∎

6.15 Esempio. Si considerino le due matrici A e $B \in \mathbf{R}^{5\times 5}$ simmetriche a banda

$$A = \begin{bmatrix} 6 & -4 & 0 & 0 & 0 \\ -4 & 6 & -4 & 0 & 0 \\ 0 & -4 & 6 & -4 & 0 \\ 0 & 0 & -4 & 6 & -4 \\ 0 & 0 & 0 & -4 & 6 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}.$$

Gli autovalori di A sono dati da

$$\lambda_k = 6 + 8\cos\frac{k\pi}{6},\ k = 1, \ldots, 5$$

(si veda l'esercizio 2.40), cioè $\lambda_1 = 6 + 4\sqrt{3} = 12.92820$,

$$\lambda_2 = 10,\ \lambda_3 = 6,\ \lambda_4 = 2,\ \lambda_5 = 6 - 4\sqrt{3} = -0.9282032.$$

La matrice B ha il polinomio caratteristico

$$p(\lambda) = -\lambda\,(\lambda^2 - 1)\,(\lambda^2 - 2)$$

e quindi ha gli autovalori

$$\mu_1 = \sqrt{2} = 1.414214,\ \mu_2 = 1,\ \mu_3 = 0,\ \mu_4 = -1,\ \mu_5 = -\sqrt{2} = -1.414214.$$

La matrice $C = A + B$ è una matrice pentadiagonale i cui autovalori sono

$$\nu_1 = 14.10892, \nu_2 = 9.531118, \nu_3 = 4.678975, \nu_4 = 1.468864, \nu_5 = 0.2120767$$

e soddisfano le disuguaglianze

$$\lambda_k - \sqrt{2} < \nu_k < \lambda_k + \sqrt{2}, \quad k = 1, \ldots, 5. \qquad \blacksquare$$

Il teorema 6.14 fornisce anche un risultato di perturbazione. Se A e $B \in \mathbf{C}^{n\times n}$ sono matrici hermitiane e $C = A + \epsilon B,\ \epsilon > 0$, allora per gli autovalori $\nu_1, \ldots, \nu_n$ di C vale la relazione

$$\lambda_k + \epsilon\mu_n \leq \nu_k \leq \lambda_k + \epsilon\mu_1,$$

dove $\lambda_1, \ldots, \lambda_n$ sono gli autovalori di A e $\mu_1, \ldots, \mu_n$ sono gli autovalori di B, ordinati in ordine non crescente. Cioè la variazione sugli autovalori della matrice perturbata C è proporzionale all'entità della perturbazione ϵB.

6.16 Esempio. Sia

$$A = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 & 2 \\ 1 & 2 & 3 & 3 & 3 \\ 1 & 2 & 3 & 4 & 4 \\ 1 & 2 & 3 & 4 & 5 \end{bmatrix}$$

la matrice dell'esempio 6.11 e sia $B \in \mathbf{R}^{5\times 5}$ tale che $b_{ij} = 1$ per $i,\ j = 1, \ldots, 5$. Scegliendo per ϵ i valori $10^{-r},\ r = 1, \ldots, 4$, si ottengono per la matrice perturbata $C = A + \epsilon B$ gli autovalori riportati nella seguente tabella.

ϵ	ν_1	ν_2	ν_3	ν_4	ν_5
10^{-1}	12.78558	1.491278	0.5942757	0.3566023	0.2722294
10^{-2}	12.38752	1.453030	0.5841669	0.3536236	0.2716302
10^{-3}	12.34792	1.449116	0.5830847	0.3532901	0.2715611
10^{-4}	12.34396	1.448725	0.5829763	0.3532561	0.2715544

$\blacksquare$

4. Introduzione ai metodi

Nei prossimi paragrafi vengono presentati metodi numerici per calcolare gli autovalori e gli autovettori di una matrice. Fra i diversi metodi considerati alcuni hanno carattere generale e sono convenientemente applicabili a matrici dense e senza stuttura, altri utilizzano in modo specifico eventuali proprietà di struttura o sparsità della matrice, permettendo di trattare problemi anche con dimensioni molto elevate. Alcuni dei metodi esposti possono essere utilizzati per calcolare tutti gli autovalori e autovettori di una matrice, altri invece servono per calcolare solo alcuni autovalori, per esempio quelli che si trovano all'estremità dello spettro, ed i corrispondenti autovettori, come è richiesto in molte applicazioni.

I metodi per il calcolo degli autovalori possono essere divisi in due classi.

1) Metodi in cui il calcolo viene effettuato in due fasi: riduzione con metodi diretti della matrice A in una matrice simile B, di cui sia più agevole calcolare gli autovalori, e calcolo degli autovalori di B con un metodo iterativo. Questi metodi si applicano in generale a problemi di piccole dimensioni, per i quali tutti i dati su cui si opera possono essere contenuti nella memoria centrale del calcolatore.

2) Metodi completamente iterativi, che richiedono ad ogni passo la moltiplicazione di una matrice per un vettore, o la risoluzione di un sistema lineare. Questi metodi si applicano in generale a problemi di grandi dimensioni, anche nel caso in cui non sia possibile contenere tutti i dati nella memoria centrale del calcolatore.

Nei metodi della prima classe per la riduzione della matrice A nella matrice B si utilizzano metodi diretti analoghi a quelli descritti per la fattorizzazione delle matrici. Nel caso più generale la matrice B che si ottiene è tale che

$$b_{ij} = 0, \quad \text{per} \quad i > j+1, \ i, \ j = 1, \ldots, n.$$

Una matrice B con questa proprietà è detta essere *in forma di Hessenberg superiore*. Se la matrice A è hermitiana, e la trasformazione viene eseguita con matrici unitarie, la matrice B risulta hermitiana e tridiagonale.

Se $B = T^{-1}AT$ e A è diagonalizzabile, cioè $A = SDS^{-1}$, dove D è la matrice diagonale i cui elementi principali sono gli autovalori di A, allora anche B è diagonalizzabile e risulta

$$B = (T^{-1}S)D(T^{-1}S)^{-1}.$$

La matrice $T^{-1}S$ ha per colonne gli autovettori di B. Poiché per il teorema 6.4 il condizionamento del problema del calcolo degli autovalori di una matrice diagonalizzabile è legato al numero di condizionamento della matrice

degli autovettori, è opportuno determinare la matrice T in modo tale che il numero di condizionamento di $T^{-1}S$ sia minore o uguale al numero di condizionamento di S. Ciò è senz'altro vero se $\mu(T) = \|T\| \, \|T^{-1}\| = 1$, in tal caso infatti

$$\begin{aligned}\mu(T^{-1}S) &= \|T^{-1}S\| \, \|(T^{-1}S)^{-1}\| \\ &\leq \|T\| \, \|T^{-1}\| \, \|S\| \, \|S^{-1}\| = \mu(T) \, \mu(S) = \mu(S).\end{aligned}$$

In generale conviene utilizzare trasformazioni per similitudine in cui $\mu(T)$ sia piccolo.

La trasformazione per similitudine della matrice A nella matrice B è fatta per passi successivi

$$A^{(k+1)} = T_k^{-1} A^{(k)} T_k, \;\; k = 1, 2, \ldots, m-1, \tag{16}$$

dove

$$A^{(1)} = A \quad \text{e} \quad A^{(m)} = B,$$

per cui, posto $T = T_1 T_2 \ldots T_{m-1}$, risulta $B = T^{-1}AT$, e se $\mathbf{x}$ è autovettore di B, $T\mathbf{x}$ è autovettore di A.

Le matrici T_k sono di solito matrici elementari di Gauss o di Householder oppure matrici di Givens. Se T_k è una matrice di Householder o di Givens, risulta

$$\|T_k\|_2 \, \|T_k^{-1}\|_2 = 1,$$

se T_k è una matrice di Gauss con elementi non principali di modulo minore o uguale ad 1 (massimo pivot per colonne), risulta

$$\|T_k\|_\infty \, \|T_k^{-1}\|_\infty \leq 4.$$

I metodi iterativi per il calcolo degli autovalori di B potrebbero essere anche applicati direttamente alla matrice A. Trasformando però prima la matrice A nella matrice B, si abbassa notevolmente il numero delle operazioni richieste da ogni iterazione (ad esempio per il metodo QR, descritto nel paragrafo 8, si passa da un numero di operazioni dell'ordine n^3 ad uno dell'ordine di n^2).

Per il calcolo degli autovalori della matrice B, due sono le tecniche più usate:

a) se sono richiesti solo pochi autovalori rispetto alla dimensione della matrice (non più del 25%), conviene usare un metodo iterativo che calcoli un singolo autovalore per volta, come ad esempio un metodo di iterazione funzionale applicato all'equazione caratteristica o il metodo delle potenze inverse (paragrafo 11). È questo il modo migliore di

procedere per matrici hermitiane tridiagonali o per matrici in forma di Hessenberg superiore e sparse;

b) se sono richiesti tutti o molti degli autovalori, il metodo migliore è in generale il QR (paragrafo 8).

I metodi della seconda classe sono basati sul calcolo di successioni di vettori del tipo $\mathbf{x}_{k+1} = B\mathbf{x}_k$, dove la matrice B può essere, a seconda del metodo considerato, la A, la A^{-1} oppure una matrice $(A-\alpha I)^{-1}$. In questo modo ad ogni passo viene effettuata sempre la stessa trasformazione sul vettore corrente $\mathbf{x}_k$. Se la matrice B ha particolari proprietà di struttura o di sparsità questa trasformazione può essere fatta senza dover memorizzare tutti gli elementi della matrice nella memoria principale. Questi metodi sono particolarmente adatti a problemi di grosse dimensioni con matrici sparse, quando si richiede il calcolo di un numero limitato di autovalori e autovettori. Se la matrice A ha proprietà di struttura, ad esempio è una matrice a banda, è possibile ridurre il numero delle operazioni richieste ad ogni passo sfruttando queste proprietà. Un metodo classico, molto semplice, appartenente a questa classe è il metodo delle potenze (paragrafo 10) che approssima l'autovalore di modulo massimo e il corrispondente autovettore. Opportune varianti del metodo delle potenze consentono di calcolare anche altri autovalori e autovettori della matrice. In particolare la variante delle potenze inverse di Wielandt (paragrafo 11) è la più usata per calcolare l'autovettore corrispondente ad un autovalore di cui è nota un'approssimazione. Fra i metodi di questa seconda classe il metodo di Lanczos (paragrafo 13) è il più importante per calcolare gli autovalori di matrici reali, simmetriche e sparse, di grosse dimensioni.

Un metodo iterativo classico che non appartiene alle due classi sopra descritte è il metodo di Jacobi (paragrafo 9), che con successive trasformazioni unitarie costruisce una successione di matrici che converge a una matrice diagonale e consente quindi di calcolare tutti gli autovalori e gli autovettori contemporaneamente.

5. Riduzione di una matrice hermitiana in forma tridiagonale: i metodi di Householder, di Givens e di Lanczos

Una matrice hermitiana può essere trasformata in una matrice tridiagonale hermitiana mediante trasformazioni per similitudine unitarie utilizzando le matrici di Householder o quelle di Givens, oppure con il procedimento di Lanczos.

a) *Metodo di Householder.*

Sia $A \in \mathbf{C}^{n\times n}$ una matrice hermitiana; si considerino le trasformazioni (16), con $m = n-1$, in cui le matrici T_k siano matrici elementari di Householder (hermitiane e unitarie):

$$T_k = I - \beta_k \mathbf{u}_k \mathbf{u}_k^H,$$

costruite in modo che nella matrice $T_k A^{(k)}$ siano nulli tutti gli elementi della k-esima colonna, con l'indice di riga maggiore di $k+1$.

Al primo passo, posto

$$A^{(1)} = A = \begin{bmatrix} a_{11}^{(1)} & \mathbf{a}_1^H \\ \mathbf{a}_1 & B^{(1)} \end{bmatrix} \begin{matrix} \} & 1 \text{ riga} \\ \} & n-1 \text{ righe} \end{matrix}$$

si consideri la matrice elementare di Householder $P^{(1)} \in \mathbf{C}^{(n-1)\times(n-1)}$ tale che

$$P^{(1)}\mathbf{a}_1 = \alpha_1 \mathbf{e}_1,$$

dove $\mathbf{e}_1$ è il primo vettore della base canonica di $\mathbf{C}^{n-1}$. La matrice

$$T_1 = \begin{bmatrix} 1 & \mathbf{0}^H \\ \mathbf{0} & P^{(1)} \end{bmatrix},$$

è tale che nella matrice

$$A^{(2)} = T_1^{-1} A^{(1)} T_1 = T_1 A^{(1)} T_1$$

sono nulli tutti gli elementi della prima colonna con indice di riga maggiore di due e dei simmetrici elementi della prima riga.

Al k-esimo passo la sottomatrice principale di testa di ordine $k+1$ di $A^{(k)}$ risulta tridiagonale hermitiana e $A^{(k)}$ ha la forma

$$A^{(k)} = \begin{bmatrix} C^{(k)} & \mathbf{b}_k & O \\ \mathbf{b}_k^H & a_{kk}^{(k)} & \mathbf{a}_k^H \\ O & \mathbf{a}_k & B^{(k)} \end{bmatrix} \begin{matrix} \} & k-1 \text{ righe} \\ \} & 1 \text{ riga} \\ \} & n-k \text{ righe} \end{matrix}$$

dove $C^{(k)} \in \mathbf{C}^{(k-1)\times(k-1)}$ è tridiagonale hermitiana e $\mathbf{b}_k \in \mathbf{C}^{k-1}$ ha nulle le prime $k-2$ componenti. Sia $P^{(k)} \in \mathbf{C}^{(n-k)\times(n-k)}$ la matrice di Householder tale che

$$P^{(k)}\mathbf{a}_k = \alpha_k \mathbf{e}_1,$$

dove $\mathbf{e}_1$ è il primo vettore della base canonica di $\mathbf{C}^{n-k}$. Posto

$$T_k = \begin{bmatrix} I_k & \mathbf{0}^H \\ \mathbf{0} & P^{(k)} \end{bmatrix}$$

risulta

$$A^{(k+1)} = T_k^{-1} A^{(k)} T_k = T_k A^{(k)} T_k = \begin{bmatrix} C^{(k)} & \mathbf{b}_k & O \\ \mathbf{b}_k^H & a_{kk}^{(k)} & \mathbf{a}_k^H P^{(k)} \\ O & P^{(k)} \mathbf{a}_k & P^{(k)} B^{(k)} P^{(k)} \end{bmatrix}.$$

Poiché il vettore $P^{(k)}\mathbf{a}_k \in \mathbf{C}^{n-k}$ ha nulle le componenti di indice maggiore o uguale a due, la sottomatrice principale di testa di ordine $k+2$ della matrice $A^{(k+1)}$ è tridiagonale hermitiana. Applicando il procedimento $n-2$ volte si ottiene la matrice $B = A^{(n-1)}$ tridiagonale hermitiana.

Per calcolare $P^{(k)}B^{(k)}P^{(k)}$ non si utilizza esplicitamente la matrice $P^{(k)}$, ma si procede in modo analogo a quanto fatto nella risoluzione dei sistemi lineari, sfruttando inoltre il fatto che $B^{(k)}$ è una matrice hermitiana. Poiché

$$\begin{aligned} P^{(k)}B^{(k)}P^{(k)} &= (I - \beta_k \mathbf{u}_k \mathbf{u}_k^H) B^{(k)} (I - \beta_k \mathbf{u}_k \mathbf{u}_k^H) \\ &= B^{(k)} - \beta_k B^{(k)} \mathbf{u}_k \mathbf{u}_k^H - \beta_k \mathbf{u}_k \mathbf{u}_k^H B^{(k)} + \beta_k^2 \mathbf{u}_k (\mathbf{u}_k^H B^{(k)} \mathbf{u}_k) \mathbf{u}_k^H \\ &= B^{(k)} - [\beta_k B^{(k)} \mathbf{u}_k - \tfrac{1}{2}\beta_k (\mathbf{u}_k^H \beta_k B^{(k)} \mathbf{u}_k) \mathbf{u}_k] \mathbf{u}_k^H \\ &\quad - \mathbf{u}_k [\beta_k \mathbf{u}_k^H B^{(k)} - \tfrac{1}{2}\beta_k (\mathbf{u}_k^H \beta_k B^{(k)} \mathbf{u}_k) \mathbf{u}_k^H], \end{aligned}$$

ponendo

$$\mathbf{r}_k = \beta_k B^{(k)} \mathbf{u}_k$$

e

$$\mathbf{q}_k = \mathbf{r}_k - \tfrac{1}{2}\beta_k (\mathbf{r}_k^H \mathbf{u}_k) \mathbf{u}_k,$$

si ha:

$$P^{(k)}B^{(k)}P^{(k)} = B^{(k)} - \mathbf{q}_k \mathbf{u}_k^H - \mathbf{u}_k \mathbf{q}_k^H.$$

La trasformazione $A^{(k)} \to A^{(k+1)}$ richiede allora solo $2(n-k)^2$ operazioni moltiplicative. Il metodo di Householder per tridiagonalizzare una matrice hermitiana richiede dunque

$$\sum_{k=1}^{n-2} 2(n-k)^2 \simeq \frac{2}{3}\, n^3 \quad \text{operazioni moltiplicative.}$$

Il metodo di Householder per la riduzione di una matrice hermitiana in forma tridiagonale gode delle stesse proprietà di stabilità del corrispondente metodo per la risoluzione dei sistemi lineari. È infatti possibile dimostrare che la matrice B' effettivamente calcolata è simile ad una matrice "vicina" ad A, cioè che esiste una matrice ortogonale Q e una matrice E tali che

$$B' = Q^H(A + E)Q,$$

con $\|E\|_F \leq \gamma u \|A\|_F$, in cui $\gamma = \gamma(n)$ è una funzione polinomiale di n di grado basso e u è la precisione di macchina con cui sono stati effettuati i calcoli [28].

b) *Metodo di Givens.*

Sia $A \in \mathbf{C}^{n\times n}$ una matrice hermitiana; si considerino le trasformazioni (16), in cui le matrici T_k siano matrici di Givens:

$$T_k = G_{pq} = \begin{bmatrix} 1 & & & & & & & & & & \\ & \ddots & & & & & & & & & \\ & & 1 & & & & & & & & \\ & & & c & & & & -\overline{s} & & & \leftarrow p \\ & & & & 1 & & & & & & \\ & & & & & \ddots & & & & & \\ & & & & & & 1 & & & & \\ & & & s & & & & c & & & \leftarrow q \\ & & & & & & & & 1 & & \\ & & & & & & & & & \ddots & \\ & & & & & & & & & & 1 \end{bmatrix}$$
$$\begin{matrix} & & \uparrow & & \uparrow & & \\ & & p & & q & & \end{matrix}$$

in cui $c = \cos\phi$ e $s = \psi \sin\phi$, $\phi \in \mathbf{R}$ e $\psi \in \mathbf{C}$, con $|\psi| = 1$. È $G_{pq}^{-1} = G_{pq}^H$.

La matrice $A^{(k+1)} = G_{pq}^H A^{(k)} G_{pq}$ differisce da $A^{(k)}$ solo nella p-esima e q-esima riga e nella p-esima e q-esima colonna. Infatti, considerando per semplicità il caso in cui $A^{(k)}$ è reale e indicando con a_{rj} gli elementi di $A^{(k)}$ e con $\hat{a}_{rj}$ gli elementi di $A^{(k+1)}$, si ha

$$\begin{aligned}
&\hat{a}_{rp} = \hat{a}_{pr} = ca_{pr} + sa_{qr}, \qquad r \neq p,\, q; \quad r = 1, 2, \ldots, n,\\
&\hat{a}_{rq} = \hat{a}_{qr} = -sa_{pr} + ca_{qr}, \quad r \neq p,\, q; \quad r = 1, 2, \ldots, n,\\
&\hat{a}_{pp} = c^2 a_{pp} + s^2 a_{qq} + 2csa_{pq},\\
&\hat{a}_{qq} = c^2 a_{qq} + s^2 a_{pp} - 2csa_{pq},\\
&\hat{a}_{qp} = \hat{a}_{pq} = (c^2 - s^2)a_{pq} - cs(a_{pp} - a_{qq}),\\
&\hat{a}_{rj} = a_{rj}, \quad \text{altrimenti.}
\end{aligned} \tag{17}$$

La scelta di G_{pq} può essere effettuata in modo che per un assegnato valore di r risulti

$$\hat{a}_{qr} = \hat{a}_{rq} = 0.$$

Infatti se $a_{qr} = 0$, è sufficiente porre $G_{pq} = I$, se invece $a_{qr} \neq 0$, si possono ricavare c ed s dalla condizione

$$\hat{a}_{qr} = -sa_{pr} + ca_{qr} = 0,$$

ossia

$$c = \frac{a_{pr}}{\sqrt{a_{pr}^2 + a_{qr}^2}} \quad \text{ed} \quad s = \frac{a_{qr}}{\sqrt{a_{pr}^2 + a_{qr}^2}}\ ,$$

da cui, procedendo come nel paragrafo 16 del capitolo 4, si ottengono le formule più stabili:

$$\text{se } |a_{pr}| \geq |a_{qr}|, \text{ allora si pone } t = \frac{a_{qr}}{a_{pr}},\ c = \frac{1}{\sqrt{1+t^2}},\ s = tc,$$
$$\text{altrimenti si pone } t = \frac{a_{pr}}{a_{qr}},\ s = \frac{1}{\sqrt{1+t^2}},\ c = ts.$$

Il processo completo di riduzione in forma tridiagonale, richiede $m = \dfrac{(n-1)(n-2)}{2}$ trasformazioni, con la seguente scelta di indici (p, q) ed r:

$$\begin{array}{lllll} (2,3) & (2,4) & \dots & (2,n) & \text{con } r = 1 \\ & (3,4) & \dots & (3,n) & \text{con } r = 2 \\ & & \ddots & \vdots & \\ & & & (n-1,n) & \text{con } r = n-2. \end{array}$$

Ogni trasformazione $A^{(k)} \to A^{(k+1)}$ richiede $4(n-r)$ operazioni moltiplicative, e per ogni r il numero di trasformazioni richieste è $n - r$. In totale la riduzione di una matrice hermitiana in forma tridiagonale con il metodo di Givens richiede

$$\sum_{r=1}^{n-2} 4(n-r)^2 \simeq \frac{4}{3}\, n^3 \quad \text{operazioni moltiplicative.}$$

Dal punto di vista della stabilità numerica, il comportamento del metodo di Givens è analogo a quello del metodo di Householder. Il numero delle operazioni richieste dal metodo di Householder risulta inferiore a quello richiesto dal metodo di Givens. Però il metodo di Givens è più adatto a sfruttare l'eventuale presenza di elementi nulli nella matrice A e in quelle

generate durante il metodo, ed è quindi possibile in questi casi che il metodo di Givens richieda meno operazioni del metodo di Householder.

6.17 Esempio. Si consideri la matrice $A \in \mathbf{R}^{4\times 4}$

$$A = \begin{bmatrix} 4 & 3 & 2 & 1 \\ 3 & 4 & 3 & 2 \\ 2 & 3 & 4 & 3 \\ 1 & 2 & 3 & 4 \end{bmatrix}.$$

Applicando il metodo di Householder, al primo passo si ottiene

$$\beta_1 = 0.03964327$$

$$\mathbf{u}_1 = [0,\ 6.741657,\ 2,\ 1]^T,$$

e quindi

$$A^{(2)} = \begin{bmatrix} 4 & \text{-3.741657} & 0 & 0 \\ \text{-3.741657} & 8.285713 & \text{-1.301424} & \text{-2.254283} \\ 0 & \text{-1.301424} & 1.070671 & 0.9112844 \\ 0 & \text{-2.254283} & 0.9112844 & 2.643615 \end{bmatrix};$$

al secondo passo si ottiene

$$\beta_2 = 0.09839517$$

$$\mathbf{u}_2 = [0,\ 0,\ \text{-3.904409},\ \text{-2.254283}]^T,$$

e quindi

$$A^{(3)} = \begin{bmatrix} 4 & \text{-3.741657} & 0 & 0 \\ \text{-3.741657} & 8.285713 & 2.602978 & 0 \\ 0 & 2.602978 & 3.039586 & \text{-0.2253981} \\ 0 & 0 & \text{-0.2253981} & 0.6746988 \end{bmatrix}.$$

Applicando il metodo di Givens, al primo passo si pone $r = 1,\ p = 2,\ q = 3$ e si ottiene $c = 0.8320504,\ s = 0.5547003$ e quindi

$$A^{(2)} = \begin{bmatrix} 4 & 3.605551 & 0 & 1 \\ 3.605551 & 6.769231 & 1.153846 & 3.328201 \\ 0 & 1.153846 & 1.230768 & 1.386751 \\ 1 & 3.328201 & 1.386751 & 4 \end{bmatrix};$$

al secondo passo si pone $r = 1,\ p = 2,\ q = 4$ e si ottiene $c = 0.9636238,\ s = 0.2672612$ e quindi

$$A^{(3)} = \begin{bmatrix} 4 & 3.741654 & 0 & 0 \\ 3.741654 & 8.285707 & 1.482497 & 2.139555 \\ 0 & 1.482497 & 1.230768 & 1.027927 \\ 0 & 2.139555 & 1.027927 & 2.483513 \end{bmatrix};$$

al terzo passo si pone $r = 2,\ p = 3,\ q = 4$ e si ottiene $c = 0.5695390,\ s = 0.8219641$ e quindi

$$A^{(4)} = \begin{bmatrix} 4 & 3.741654 & 0 & 0 \\ 3.741654 & 8.285707 & 2.602977 & 0 \\ 0 & 2.602977 & 3.039581 & 0.2254009 \\ 0 & 0 & 0.2254009 & 0.6746972 \end{bmatrix}.$$

Si noti che la matrice $A^{(4)}$, non tenendo conto degli errori di arrotondamento, è uguale a quella ottenuta con il metodo di Householder, a meno di una matrice di fase reale, cioè, detta $H^{(3)}$ la matrice ottenuta con il metodo di Householder, è

$$H^{(3)} = D^{-1} A^{(4)} D,$$

dove D è una matrice diagonale con elementi principali uguali a 1 o a -1. ∎

c) *Metodo di Lanczos.*

Sia $A \in \mathbf{C}^{n\times n}$ una matrice hermitiana e $Q \in \mathbf{C}^{n\times n}$ una matrice unitaria le cui colonne sono $\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n$, tali che

$$Q^H A Q = T = \begin{bmatrix} \alpha_1 & \beta_1 & & \\ \beta_1 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \beta_{n-1} \\ & & \beta_{n-1} & \alpha_n \end{bmatrix}, \tag{18}$$

dove $\alpha_i \in \mathbf{R}$ per $i = 1, \ldots, n$, e $\beta_i \in \mathbf{R}$, $\beta_i \geq 0$ per $i = 1, \ldots, n-1$. Il metodo di Lanczos permette di generare, a partire dalla prima colonna $\mathbf{q}_1$ di Q, attraverso un processo di ortogonalizzazione, le rimanenti colonne di Q e gli elementi di T. Infatti scrivendo la (18) come

$$AQ = QT,$$

e confrontando le i-esime colonne, per $i = 1, 2, \ldots, n$, a primo e a secondo membro si ottengono le relazioni

$$\begin{aligned} A\mathbf{q}_1 &= \alpha_1 \mathbf{q}_1 + \beta_1 \mathbf{q}_2, \\ A\mathbf{q}_i &= \beta_{i-1}\mathbf{q}_{i-1} + \alpha_i \mathbf{q}_i + \beta_i \mathbf{q}_{i+1}, \quad i = 2, \ldots, n-1, \\ A\mathbf{q}_n &= \beta_{n-1}\mathbf{q}_{n-1} + \alpha_n \mathbf{q}_n, \end{aligned} \tag{19}$$

Sfruttando il fatto che i vettori $\mathbf{q}_i$ sono ortonormali, se $\beta_i \neq 0$, per $i = 1, \ldots, n-1$, si ottengono le relazioni

$$\begin{aligned}
&\alpha_1 = \mathbf{q}_1^H A \mathbf{q}_1, \quad \mathbf{q}_2 = \frac{(A - \alpha_1 I)\mathbf{q}_1}{\beta_1}, \quad \beta_1 = \|(A - \alpha_1 I)\mathbf{q}_1\|_2 \\
&\alpha_i = \mathbf{q}_i^H A \mathbf{q}_i, \quad \mathbf{q}_{i+1} = \frac{(A - \alpha_i I)\mathbf{q}_i - \beta_{i-1}\mathbf{q}_{i-1}}{\beta_i}, \\
&\qquad\qquad \beta_i = \|(A - \alpha_i I)\mathbf{q}_i - \beta_{i-1}\mathbf{q}_{i-1}\|_2, \quad i = 2, \ldots, n-1, \\
&\alpha_n = \mathbf{q}_n^H A \mathbf{q}_n,
\end{aligned} \tag{20}$$

che permettono di calcolare gli elementi di T e le colonne di Q se tutti i β_i sono non nulli. Se uno dei β_i fosse nullo, il procedimento può proseguire solo conoscendo il vettore $\mathbf{q}_{i+1}$.

Il procedimento di Lanczos comunque si può applicare a partire da un qualunque vettore $\mathbf{u} \in \mathbf{C}^n$, tale che $\|\mathbf{u}\|_2 = 1$, scegliendo, se uno dei β_i risultasse nullo, come $\mathbf{q}_{i+1}$ un qualsiasi vettore ortonormale ai vettori $\mathbf{q}_j$ già calcolati. Il procedimento può quindi essere portato a termine in ogni caso. I vettori $\mathbf{q}_i$ così calcolati sono ortonormali. Vale infatti il seguente teorema.

6.18 Teorema. *Sia $\mathbf{u} \in \mathbf{C}^n$, tale che $\|\mathbf{u}\|_2 = 1$. Scelto $\mathbf{q}_1 = \mathbf{u}$, i vettori $\mathbf{q}_1$ e $\mathbf{q}_i$, $i = 2, \ldots, n$, calcolati con la (20) (se $\beta_i = 0$ si sceglie come $\mathbf{q}_{i+1}$ un qualunque vettore ortogonale a $\mathbf{q}_1, \ldots, \mathbf{q}_i$, con $\|\mathbf{q}_{i+1}\|_2 = 1$), sono ortonormali. La matrice Q avente per colonne i vettori $\mathbf{q}_1, \ldots, \mathbf{q}_n$, e gli α_i, $i = 1, \ldots, n$, e i β_i, $i = 1, \ldots, n-1$, verificano la (18). Inoltre se i β_i sono tutti non nulli, la matrice Q così ottenuta è l'unica matrice per cui vale la (18) e tale che $Q\mathbf{e}_1 = \mathbf{u}$.*

Dim. I vettori $\mathbf{q}_1, \ldots, \mathbf{q}_n$ verificano la relazione $\|\mathbf{q}_i\|_2 = 1$ per costruzione. Per dimostrarne l'ortogonalità si procede per induzione. Si suppone che i vettori $\mathbf{q}_1, \ldots, \mathbf{q}_k$ siano ortogonali e si dimostra che $\mathbf{q}_{k+1}$ è ortogonale a $\mathbf{q}_1, \ldots, \mathbf{q}_k$. Per $k = 2$, i vettori $\mathbf{q}_1$ e $\mathbf{q}_2$ sono ortogonali per costruzione. Se $\beta_k = 0$ l'ortogonalità è verificata per costruzione. Altrimenti basta dimostrare che $\mathbf{q}_{k+1}$ è ortogonale a $\mathbf{q}_j$, $j = 1, \ldots, k-2$, perché $\mathbf{q}_{k+1}$ è ortogonale a $\mathbf{q}_k$ e $\mathbf{q}_{k-1}$ per le (20) e per l'ipotesi induttiva. Si ha dalla (19) per $j = 1, \ldots, k-2$,

$$\beta_k \mathbf{q}_{k+1}^H \mathbf{q}_j = \mathbf{q}_k^H A \mathbf{q}_j - \beta_{k-1} \mathbf{q}_{k-1}^H \mathbf{q}_j - \alpha_k \mathbf{q}_k^H \mathbf{q}_j,$$

e per l'ipotesi induttiva

$$\beta_k \mathbf{q}_{k+1}^H \mathbf{q}_j = \mathbf{q}_k^H A \mathbf{q}_j.$$

Poichè per la (19) è

$$A\mathbf{q}_j = \beta_{j-1}\mathbf{q}_{j-1} + \alpha_j\mathbf{q}_j + \beta_j\mathbf{q}_{j+1},$$

si ha

$$\mathbf{q}_k^H A\mathbf{q}_j = \beta_{j-1}\mathbf{q}_k^H\mathbf{q}_{j-1} + \alpha_j\mathbf{q}_k^H\mathbf{q}_j + \beta_j\mathbf{q}_k^H\mathbf{q}_{j+1},$$

e per l'ipotesi induttiva

$$\beta_k\mathbf{q}_{k+1}^H\mathbf{q}_j = 0,$$

da cui, poiché $\beta_k \neq 0$, segue che $\mathbf{q}_{k+1}^H\mathbf{q}_j = 0$. Per dimostrare che vale la (18) è sufficiente verificare che valgono le (19). Le prime $n-1$ relazioni in (19) sono verificate per costruzione dai vettori $\mathbf{q}_i$. L'ultima delle relazioni (19) è verificata perché il vettore

$$\mathbf{v} = A\mathbf{q}_n - \beta_{n-1}\mathbf{q}_{n-1} - \alpha_n\mathbf{q}_n$$

risulta nullo essendo ortogonale a $\mathbf{q}_1, \ldots, \mathbf{q}_n$. Infatti $\mathbf{q}_n^H\mathbf{v} = 0$ per la definizione di α_n e per $j = 1, \ldots, n-1$ è $\mathbf{q}_j^H\mathbf{v} = 0$, poiché

$$\begin{aligned}\mathbf{q}_j^H\mathbf{v} &= \mathbf{q}_j^H(A\mathbf{q}_n - \beta_{n-1}\mathbf{q}_{n-1} - \alpha_n\mathbf{q}_n) = \mathbf{q}_j^H A\mathbf{q}_n - \beta_{n-1}\mathbf{q}_j^H\mathbf{q}_{n-1}\\ &= (\beta_{j-1}\mathbf{q}_{j-1} + \alpha_j\mathbf{q}_j + \beta_j\mathbf{q}_{j+1})^H\mathbf{q}_n - \beta_{n-1}\mathbf{q}_j^H\mathbf{q}_{n-1}\\ &= \beta_j\mathbf{q}_{j+1}^H\mathbf{q}_n - \beta_{n-1}\mathbf{q}_j^H\mathbf{q}_{n-1}\end{aligned}$$

e quest'ultima relazione è nulla anche per $j = n-1$.

L'unicità della decomposizione (18) nel caso in cui $\beta_i \neq 0$ per $i = 1, \ldots, n-1$, segue dal fatto che la (18) e le (20) sono equivalenti se $\beta_i > 0$. ∎

In pratica se si genera solo la matrice T e non la matrice Q, il procedimento di Lanczos può essere implementato utilizzando solamente due vettori. Se il numero delle operazioni moltiplicative richieste dal prodotto della matrice A per un vettore è dato da hn (se A è una matrice piena è $h = n$, mentre se A è sparsa è $h << n$), il costo computazionale ad ogni passo è dato da $(5+h)n$ operazioni moltiplicative. Se A non è sparsa, il costo computazionale totale di questo metodo è dell'ordine di n^3 operazioni moltiplicative (quindi superiore a quello dei metodi di Householder e di Givens), se A è una matrice a banda, con $2p+1$ diagonali, allora il costo totale è di $(2p+6)n^2$ operazioni moltiplicative. Quindi tale metodo sembra essere molto indicato per matrici sparse e di dimensioni molto grandi.

Il metodo di tridiagonalizzazione di Lanczos presenta grossi problemi di stabilità numerica: infatti se uno dei β_i è piccolo, nel calcolo di $\mathbf{q}_{i+1}$ si possono presentare elevati errori di cancellazione, con una conseguente perdita

di ortogonalità dei vettori calcolati successivamente. Anche per questo motivo il metodo di Lanczos non è competitivo con i metodi di Givens e di Householder e non viene abitualmente usato per tridiagonalizzare matrici di dimensioni tali da poter essere contenute nella memoria del calcolatore. Il metodo di Lanczos risulta però particolarmente utile per il calcolo degli autovalori estremi dello spettro di matrici (si veda il paragrafo 13).

6.19 Esempio. Si applica il metodo di Lanczos alla matrice $A \in \mathbf{R}^{4\times 4}$

$$A = \begin{bmatrix} 4 & 3 & 2 & 1 \\ 3 & 4 & 3 & 2 \\ 2 & 3 & 4 & 3 \\ 1 & 2 & 3 & 4 \end{bmatrix},$$

già vista nell'esempio 6.17, assumendo come vettore iniziale il vettore $\mathbf{q}_1 = \mathbf{e}_1$. Si ha

$$\alpha_1 = 4, \quad \mathbf{q}_2 = \begin{bmatrix} 0 \\ 0.8017837 \\ 0.5345225 \\ 0.2672612 \end{bmatrix}, \quad \beta_1 = 3.741658,$$

$$\alpha_2 = 8.285710, \quad \mathbf{q}_3 = \begin{bmatrix} 0 \\ \text{-}0.4987068 \\ 0.3520293 \\ 0.7920648 \end{bmatrix}, \quad \beta_2 = 2.602981,$$

$$\alpha_3 = 3.039589, \quad \mathbf{q}_4 = \begin{bmatrix} 0.1041032\ 10^{-4} \\ 0.3293015 \\ \text{-}0.7683433 \\ 0.5488251 \end{bmatrix}, \quad \beta_3 = 0.2253997,$$

$$\alpha_4 = 0.6746987.$$

La matrice tridiagonale così ottenuta risulta quindi

$$T = \begin{bmatrix} 4 & 3.741658 & 0 & 0 \\ 3.741658 & 8.285710 & 2.602981 & 0 \\ 0 & 2.602981 & 3.039589 & 0.2253997 \\ 0 & 0 & 0.2253997 & 0.6746987 \end{bmatrix}.$$

Se si sceglie $\mathbf{q}_1 = \frac{1}{2}\,[1\ ,\ 1\ ,\ 1\ ,\ 1]^T$, si ottiene

$$\alpha_1 = 11, \quad \mathbf{q}_2 = \frac{1}{2}\,[-1\ ,\ 1\ ,\ 1\ ,\ -1]^T, \quad \beta_1 = 1, \quad \alpha_2 = 1, \quad \beta_2 = 0.$$

A questo punto, per poter proseguire occorre scegliere un vettore $\mathbf{q}_3$ ortonormale a $\mathbf{q}_1$ e $\mathbf{q}_2$. Scegliendo $\mathbf{q}_3 = \frac{1}{2}\,[1\ ,\ 1\ ,\ -1\ ,\ -1]^T$ si ha

$$\alpha_3 = 3, \quad \mathbf{q}_4 = \frac{1}{2}\,[1\ ,\ -1\ ,\ 1\ ,\ -1]^T, \quad \beta_3 = 1, \quad \alpha_4 = 1.$$

Si è quindi ottenuta la seguente tridiagonalizzazione

$$A = QTQ^H,$$

dove

$$Q = \frac{1}{2}\begin{bmatrix} 1 & \text{-1} & 1 & 1 \\ 1 & 1 & 1 & \text{-1} \\ 1 & 1 & \text{-1} & 1 \\ 1 & \text{-1} & \text{-1} & \text{-1} \end{bmatrix}, \quad T = \begin{bmatrix} 11 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 3 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} \qquad \blacksquare$$

6. Calcolo degli autovalori delle matrici tridiagonali hermitiane con la successione di Sturm

Per calcolare gli autovalori di una matrice tridiagonale hermitiana conviene utilizzare metodi iterativi che facciano ricorso al polinomio caratteristico solo se il numero degli autovalori che si vogliono determinare è piccolo rispetto alle dimensioni della matrice. Sia $B_n \in \mathbf{C}^{n\times n}$ la matrice tridiagonale hermitiana definita da

$$B_n = \begin{bmatrix} \alpha_1 & \overline{\beta}_2 & & & \\ \beta_2 & \alpha_2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \overline{\beta}_n \\ & & & \beta_n & \alpha_n \end{bmatrix}$$

e sia $P_n(\lambda) = \det(B_n - \lambda I)$ il suo polinomio caratteristico. Se la matrice B_n è riducibile, cioè se esiste almeno un indice $j, 2 \le j \le n$, tale che $\beta_j = 0$, allora il problema del calcolo degli autovalori di B_n è ricondotto al calcolo degli autovalori di due matrici di ordine inferiore. Infatti si ha

$$B_n = \begin{bmatrix} C_{j-1} & O \\ O & D_{n-j+1} \end{bmatrix},$$

in cui $C_{j-1} \in \mathbf{C}^{(j-1)\times(j-1)}, D_{n-j+1} \in \mathbf{C}^{(n-j+1)\times(n-j+1)}$ e quindi

$$\det(B_n - \lambda I) = \det(C_{j-1} - \lambda I_{j-1})\ \det(D_{n-j+1} - \lambda I_{n-j+1}).$$

Se le matrici C_{j-1} e D_{n-j+1} sono a loro volta riducibili, si procede in modo analogo.

Si consideri perciò il caso che B_n sia irriducibile cioè che $\beta_j \neq 0$ per $j = 2, 3, \ldots, n$. Calcolando $\det(B_n - \lambda I)$ con la regola di Laplace rispetto all'ultima riga, si ottengono le relazioni

$$\begin{aligned} &P_0(\lambda) = 1, \quad P_1(\lambda) = \alpha_1 - \lambda, \\ &P_i(\lambda) = (\alpha_i - \lambda)P_{i-1}(\lambda) - |\beta_i|^2 P_{i-2}(\lambda),\ i = 2, 3, \ldots, n, \end{aligned} \tag{21}$$

con cui è possibile calcolare il valore che il polinomio $P_n(\lambda)$ assume in un punto con $2(n-1)$ moltiplicazioni (supponendo di aver già calcolato $|\beta_i|^2,\ i = 2, 3, \ldots, n$).

6.20 Esempio. Si consideri la matrice $B_6 \in \mathbf{R}^{6\times 6}$ i cui elementi sono dati da:

$$b_{ij} = \begin{cases} 2 & \text{se } i = j, \\ 1 & \text{se } |i-j| = 1, \\ 0 & \text{altrimenti.} \end{cases}$$

Dalla (21) si ha

$$\begin{aligned} &P_0(\lambda) = 1, \quad P_1(\lambda) = 2 - \lambda, \\ &P_i(\lambda) = (2 - \lambda)P_{i-1}(\lambda) - P_{i-2}(\lambda),\ i = 2, 3, \ldots, 6, \end{aligned} \tag{22}$$

da cui

$$\begin{aligned} P_2(\lambda) &= \lambda^2 - 4\lambda + 3 \\ P_3(\lambda) &= -\lambda^3 + 6\lambda^2 - 10\lambda + 4 \\ P_4(\lambda) &= \lambda^4 - 8\lambda^3 + 21\lambda^2 - 20\lambda + 5 \\ P_5(\lambda) &= -\lambda^5 + 10\lambda^4 - 36\lambda^3 + 56\lambda^2 - 35\lambda + 6 \\ P_6(\lambda) &= \lambda^6 - 12\lambda^5 + 55\lambda^4 - 120\lambda^3 + 126\lambda^2 - 56\lambda + 7. \end{aligned}$$

Per calcolare il valore di $P_6(\lambda)$ in un punto sono richieste 5 moltiplicazioni e 6 addizioni sia con la relazione ricorrente (22) che con la regola di Ruffini-Horner che richiede però un lavoro preliminare per il calcolo dei coefficienti di $P_6(\lambda)$. Un aspetto particolarmente importante è che con la relazione ricorrente (22) si ottengono anche i valori $P_i(\lambda)$, $i = 1, \ldots, 5$, in un punto, che consentono di utilizzare un metodo semplice per calcolare gli zeri di $P_6(\lambda)$ (si veda il teorema 6.22).

La relazione ricorrente (22) e la regola di Ruffini-Horner possono generare errori algoritmici diversi per valori di λ vicini agli zeri di $P_6(\lambda)$. Ad esempio, per $\lambda = 3.8$ (uno zero di $P_6(\lambda)$ è 3.801973) si ottengono per $P_6(3.8)$ i valori

-0.3596973 con le (22) (sono esatte 4 cifre significative)

-0.3388882 con la regola di Ruffini (è esatta una sola cifra significativa). ▪

Gli autovalori di B_n vengono quindi calcolati risolvendo l'equazione caratteristica

$$P_n(\lambda) = 0, \tag{23}$$

con un metodo iterativo. Se si utilizza il metodo di Newton, il calcolo di $P'_n(\lambda)$ può essere fatto con le seguenti relazioni ricorrenti, ottenute derivando rispetto a λ entrambi i membri delle (21):

$$P'_0(\lambda) = 0, \quad P'_1(\lambda) = -1,$$
$$P'_i(\lambda) = (\alpha_i - \lambda)P'_{i-1}(\lambda) - P_{i-1}(\lambda) - |\beta_i|^2 P'_{i-2}(\lambda), \ i = 2, 3, \ldots, n.$$

Quindi il rapporto $P_n(\lambda)/P'_n(\lambda)$, che interviene ad ogni passo del metodo di Newton, può essere calcolato con $4(n-1)$ moltiplicazioni e una divisione. Nel caso in cui si debba calcolare più di un autovalore, possono essere anche utilizzate delle tecniche di deflazione, quale la variante di *Maehly* della *deflazione implicita* (si veda [26]).

Per separare le radici di (23) conviene sfruttare le proprietà delle successioni di Sturm. Infatti nel seguente teorema si dimostra che i polinomi $P_i(\lambda)$ formano una successione di Sturm.

6.21 Teorema. *Se $\beta_i \neq 0$, per $i = 2, 3, \ldots, n$, la successione dei polinomi $P_i(\lambda)$, $i = 0, 1, \ldots, n$, verifica le seguenti proprietà:*

1) *$P_0(\lambda)$ non cambia segno;*
2) *se $P_i(\lambda) = 0$, allora $P_{i-1}(\lambda)P_{i+1}(\lambda) < 0$, per $i = 1, 2, \ldots, n-1$;*
3) *se $P_n(\lambda) = 0$, allora $P'_n(\lambda)P_{n-1}(\lambda) < 0$ (e quindi $P_n(\lambda)$ ha tutti zeri di molteplicità 1).*

Una successione di polinomi che verifica le proprietà 1), 2) e 3) è detta successione di Sturm.

Dim. La 1) è ovvia. Per la 2), si osservi che da (21) si ha $P_{i-1}(\lambda)P_{i+1}(\lambda) \leq 0$. Ma se fosse $P_{i-1}(\lambda)P_{i+1}(\lambda) = 0$ e $P_i(\lambda) = 0$, allora sarebbe $P_{i-1}(\lambda) = P_{i+1}(\lambda) = 0$, da cui, per ricorrenza, seguirebbe $P_0(\lambda) = 0$, che è assurdo. Da ciò segue anche che gli zeri $\lambda_j^{(n)}$, $j = 1, 2, \ldots, n$, di $P_n(\lambda)$ sono distinti dagli zeri $\lambda_j^{(n-1)}$, $j = 1, 2, \ldots, n-1$, di $P_{n-1}(\lambda)$ e quindi, per il teorema 6.10, gli zeri $\lambda_j^{(n-1)}$ separano strettamente gli zeri $\lambda_j^{(n)}$, cioè

$$\lambda_{j+1}^{(n)} < \lambda_j^{(n-1)} < \lambda_j^{(n)}, \ j = 1, 2, \ldots, n-1.$$

Da questo fatto, tenendo presente che il coefficiente di λ^i in $P_i(\lambda)$ è $(-1)^i$, e quindi $\lim_{\lambda\to-\infty} P_i(\lambda) = +\infty$, segue la 3) (si veda la figura 6.1 per il caso $n = 4$). ∎

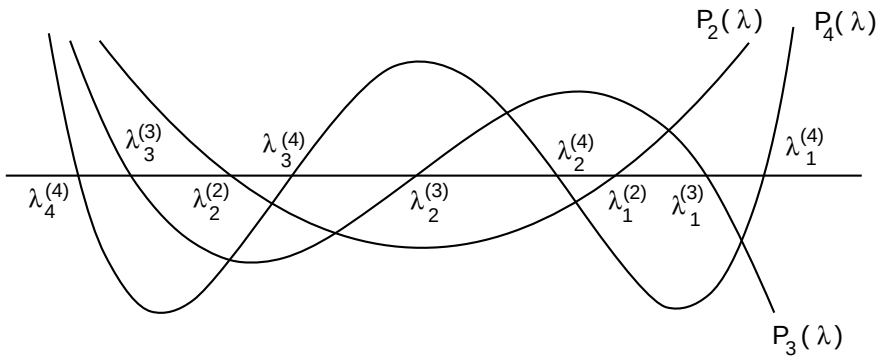


Fig. 6.1 - Grafico dei polinomi $P_2(\lambda)$, $P_3(\lambda)$ e $P_4(\lambda)$.

Si consideri, in un punto λ^*, la successione $P_0(\lambda^*), P_1(\lambda^*), \ldots, P_n(\lambda^*)$ (se fosse $P_i(\lambda^*) = 0$ per un indice $i \geq 1$, si attribuisca a tale valore il segno di $P_{i-1}(\lambda^*)$) e si indichi con $w(\lambda^*)$ il numero di cambiamenti di segno di tale successione. Vale il seguente teorema.

6.22 Teorema. *Se $\{P_i(\lambda)\}, i = 0, 1, \ldots, n$, è una successione di Sturm, il numero $w(b) - w(a)$ è uguale al numero di zeri di $P_n(\lambda)$ appartenenti all'intervallo $[a, b)$.*

Dim. Si faccia variare λ con continuità da a verso b. Si può avere una variazione nel numero $w(\lambda)$ solo quando λ incontra uno zero di uno dei polinomi $P_i(\lambda)$. Si consideri perciò un λ^* tale che $P_i(\lambda^*) = 0$ per un indice i. Per la proprietà 1) del teorema 6.21 deve essere $i \neq 0$. Si distinguono allora i due casi:

a) $i \neq n$. In questo caso, per la proprietà 2) del teorema 6.21 si ha

$$P_{i-1}(\lambda^*)P_{i+1}(\lambda^*) < 0.$$

Esiste perciò un numero h tale che nell'intervallo $[\lambda^* - h, \lambda^* + h]$ è ancora

$$P_{i-1}(\lambda)P_{i+1}(\lambda) < 0$$

e

$$P_i(\lambda) \neq 0,$$

eccetto che nel punto λ^*. Poiché per ogni $\lambda \in [\lambda^* - h, \lambda^* + h]$ i due polinomi $P_{i-1}(\lambda)$ e $P_{i+1}(\lambda)$ hanno segno discorde, $P_i(\lambda)$ deve avere in

questo intervallo segno concorde con uno dei due e discorde con l'altro. Quindi nella sequenza $P_{i-1}(\lambda), P_i(\lambda), P_{i+1}(\lambda)$ vi è una sola variazione di segno in tutto l'intervallo $[\lambda^* - h, \lambda^* + h]$, cioè il fatto che $P_i(\lambda)$ si annulli in λ^* non comporta variazioni del numero $w(\lambda)$.

b) $i = n$. In questo caso, poiché per la proprietà 3) del teorema 6.21 il polinomio $P_n(\lambda)$ ha radici semplici, la sua derivata $P'_n(\lambda)$ non si annulla in λ^* ed esiste un numero h tale che nell'intervallo $[\lambda^* - h, \lambda^* + h]$ $P'_n(\lambda)$ ha lo stesso segno che $P_n(\lambda)$ ha in $\lambda^* + h$ e segno opposto a quello che $P_n(\lambda)$ ha in $\lambda^* - h$. Se h è tale che nell'intervallo $[\lambda^* - h, \lambda^* + h]$ anche $P_{n-1}(\lambda)$ non si annulla, poiché per la proprietà 3) del teorema 6.21 $P_{n-1}(\lambda)$ ha segno opposto a quello di $P'_n(\lambda)$ per $\lambda \in [\lambda^* - h, \lambda^* + h]$, la sequenza $P_{n-1}(\lambda^* + h),\ P_n(\lambda^* + h)$ presenta una variazione di segno, mentre la sequenza $P_{n-1}(\lambda^* - h),\ P_n(\lambda^* - h)$ non presenta alcuna variazione di segno.

Se ne conclude che il numero di variazioni di segno in tutta la sequenza $P_0(\lambda), P_1(\lambda), \ldots, P_n(\lambda)$ può cambiare solo nei punti in cui si annulla $P_n(\lambda)$, ed esattamente aumenta di 1 ogni volta che si annulla $P_n(\lambda)$.

Nella tesi del teorema l'intervallo $[a, b)$ è aperto a destra perché se fosse $P_n(b) = 0$, poiché a $P_n(b)$ viene assegnato lo stesso segno assunto in b da $P_{n-1}(\lambda)$, che è diverso da zero in un intorno sinistro di b, $w(\lambda)$ non cambia in tale intorno. Perciò la radice b non altera il numero di variazioni di segno.

∎

Poiché

$$\lim_{\lambda \to -\infty} P_i(\lambda) = +\infty, \quad \text{per } i = 1, 2, \ldots, n,$$

esiste $\mu \in \mathbf{R}$ tale che per ogni $\lambda \leq \mu$ è $w(\lambda) = 0$. Quindi per ogni λ^* il numero di cambiamenti di segno $w(\lambda^*)$, per il teorema 6.22, fornisce il numero di autovalori di B_n minori di λ^*.

Sul teorema 6.22 è basato il seguente procedimento per calcolare il k-esimo autovalore λ_k di una matrice B_n tridiagonale, hermitiana e irriducibile, i cui autovalori sono $\lambda_1 > \lambda_2 > \ldots > \lambda_k > \ldots > \lambda_n$:

1) sia $[a_0, b_0)$ tale che $\lambda_k \in [a_0, b_0)$,

2) per $j = 0, 1, \ldots$, sia $\xi = \frac{1}{2}(a_j + b_j)$,
   se $w(\xi) \geq n - k + 1$, allora $a_{j+1} = a_j$ e $b_{j+1} = \xi$,
   se $w(\xi) < n - k + 1$, allora $a_{j+1} = \xi$ e $b_{j+1} = b_j$.

Questo procedimento, basato sul principio della bisezione, fornisce una successione di intervalli $[a_j, b_j)$ di ampiezza $2^{-j}(b_0 - a_0)$ che contengono λ_k ed è utile per separare gli autovalori di B_n, cioè per determinare intervalli che contengono un solo autovalore della matrice. Per l'effettiva approssimazione di un autovalore conviene in generale usare il metodo di Newton.

6.23 Esempio. Nel caso della successione di Sturm ottenuta nell'esempio 6.20 si ha

λ	$P_0(\lambda)$	$P_1(\lambda)$	$P_2(\lambda)$	$P_3(\lambda)$	$P_4(\lambda)$	$P_5(\lambda)$	$P_6(\lambda)$	$w(\lambda)$
0	1	2	3	4	5	6	7	0
1	1	1	0	-1	-1	0	1	2
2	1	0	-1	0	1	0	-1	3
3	1	-1	0	1	-1	0	1	4
4	1	-2	3	-4	5	-6	7	6

Dall'ultima colonna risulta che tutti gli autovalori sono positivi, che ve ne sono due nell'intervallo (0,1), uno nell'intervallo (1,2), uno nell'intervallo (2,3), due nell'intervallo (3,4). Poiché inoltre $w(0.5) = 1$ e $w(3.5) = 5$, risulta

$$0 < \lambda_6 < 0.5 < \lambda_5 < 1 < \lambda_4 < 2 < \lambda_3 < 3 < \lambda_2 < 3.5 < \lambda_1 < 4.$$

Se si vuole ridurre l'intervallo di separazione di λ_4, si può applicare l'algoritmo di bisezione all'intervallo [1,2], ottenendo per ξ e $w(\xi)$ successivamente i valori

ξ	$w(\xi)$
1.5	2
1.75	3
1.625	3
1.5625	3
1.53125	2
1.546875	2
1.554688	2
1.558594	3

da cui si ha che

$$1.554688 < \lambda_4 < 1.558594.$$

Per approssimare λ_1 si può applicare il metodo di Newton. Scegliendo come approssimazione iniziale il punto $x_0 = 4$, si ottiene la successione

i	x_i
1	3.875000
2	3.816162
3	3.802620
4	3.801939
5	3.801973
6	3.801973

■

7. Riduzione di una matrice in forma di Hessenberg superiore

Se applicati ad una matrice A non hermitiana, i metodi di Householder e di Givens, forniscono una matrice $B = T^{-1}AT$ in forma di Hessenberg superiore. Il costo computazionale è in questo caso di $5n^3/3$ operazioni moltiplicative per il metodo di Householder e di $10n^3/3$ operazioni moltiplicative per il metodo di Givens.

6.24 Esempio. Si consideri la matrice $A \in \mathbf{R}^{4\times 4}$

$$A = \begin{bmatrix} 4 & 3 & 2 & 1 \\ 1 & 4 & 3 & 2 \\ 1 & 1 & 4 & 3 \\ 1 & 1 & 1 & 4 \end{bmatrix}.$$

Applicando il metodo di Householder, al primo passo si ottiene

$$\beta_1 = 0.2113248$$

$$\mathbf{u}_1 = [0,\ 2.732051,\ 1,\ 1]^T,$$

e quindi

$$A^{(2)} = \begin{bmatrix} 4 & -3.464098 & -0.3660240 & -1.366024 \\ -1.732050 & 7.666641 & -0.5446615 & -1.122009 \\ 0 & 0.08931351 & 1.877991 & 1.032692 \\ 0 & 1.244013 & -0.6993599 & 2.455341 \end{bmatrix};$$

al secondo passo si ottiene

$$\beta_2 = 0.5999027$$

$$\mathbf{u}_2 = [0,\ 0,\ 1.336528,\ 1.244013]^T,$$

e quindi

$$A^{(3)} = \begin{bmatrix} 4 & -3.464098 & 1.388726 & 0.2672625 \\ -1.732050 & 7.666641 & 1.158131 & 0.4629154 \\ 0 & -1.247213 & 2.476189 & -0.7423077 \\ 0 & 0 & 0.9897442 & 1.857142 \end{bmatrix},$$

che è in forma di Hessenberg superiore. Applicando invece il metodo di Givens, al primo passo si pone $r = 1,\ j = 2,\ p = 3$ e si ottiene

$$c = s = 0.7071069$$

e quindi

$$A^{(2)} = \begin{bmatrix} 4 & 3.535534 & -0.7071069 & 1 \\ 1.414213 & 6 & 1 & 3.535534 \\ 0 & -1 & 1 & 0.7071069 \\ 1 & 1.414213 & 0 & 4 \end{bmatrix};$$

al secondo passo si pone $r = 1,\ j = 2,\ p = 4$ e si ottiene

$$c = 0.8164967, \quad s = 0.5773506$$

e quindi

$$A^{(3)} = \begin{bmatrix} 4 & 3.464101 & -0.7071069 & -1.224745 \\ 1.732049 & 7.666669 & 0.8164975 & 0.9428082 \\ 0 & -0.4082471 & 2 & 1.154700 \\ 0 & -1.178512 & -0.5773511 & 2.333335 \end{bmatrix};$$

al terzo passo si pone $r = 2,\ j = 3,\ p = 4$ e si ottiene

$$c = 0.3273256, \quad s = 0.9449114$$

e quindi

$$A^{(4)} = \begin{bmatrix} 4 & 3.464101 & -1.388729 & 0.2672636 \\ 1.732049 & 7.666669 & 1.158131 & -0.4629126 \\ 0 & -1.247218 & 2.476188 & 0.7423077 \\ 0 & 0 & 0.9897417 & 1.857139 \end{bmatrix},$$

che risulta, non tenendo conto degli errori di arrotondamento, uguale, a meno di una matrice di fase reale, a quella ottenuta con il metodo di Householder. ∎

Per ridurre una matrice in forma di Hessenberg superiore attraverso trasformazioni per similitudine, si possono anche utilizzare le matrici elementari di Gauss. Per questioni di stabilità, analoghe a quelle già viste per il caso dei sistemi lineari, è necessario applicare il metodo con la tecnica del massimo pivot.

Al primo passo, posto

$$A^{(1)} = A = \begin{bmatrix} a_{11}^{(1)} & \mathbf{b}_1^H \\ \mathbf{a}_1 & B^{(1)} \end{bmatrix} \begin{matrix} \} & 1 \text{ riga} \\ \} & n-1 \text{ righe,} \end{matrix}$$

se $\mathbf{a}_1 = \mathbf{0}$, si pone $A^{(2)} = A^{(1)}$ e $T_1 = I$, altrimenti sia $\Pi_1 \in \mathbf{R}^{(n-1)\times(n-1)}$ una matrice di permutazione tale che il vettore $\mathbf{a}'_1 = \Pi_1 \mathbf{a}_1$ abbia come prima componente una componente di $\mathbf{a}_1$ di modulo massimo, e si consideri la matrice elementare di Gauss $M_1 \in \mathbf{C}^{(n-1)\times(n-1)}$ tale che il vettore

$$M_1 \mathbf{a}'_1 = \mathbf{a}''_1$$

abbia nulle tutte le componenti, esclusa la prima. Allora nella matrice

$$A^{(2)} = T_1 A^{(1)} T_1^{-1}, \quad T_1 = \begin{bmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & M_1 \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & \Pi_1 \end{bmatrix},$$

sono nulli tutti gli elementi della prima colonna con indice di riga maggiore di due.

Al k-esimo passo, si supponga che $A^{(k)}$ abbia la struttura seguente

$$A^{(k)} = \begin{bmatrix} C^{(k)} & \mathbf{b}_k & D^{(k)} \\ \mathbf{c}_k^H & a_{kk}^{(k)} & \mathbf{d}_k^H \\ O & \mathbf{a}_k & B^{(k)} \end{bmatrix} \begin{matrix} \} & k-1 \text{ righe} \\ \} & 1 \text{ riga} \\ \} & n-k \text{ righe,} \end{matrix}$$

dove $C^{(k)} \in \mathbf{C}^{(k-1)\times(k-1)}$ è in forma di Hessenberg superiore e $\mathbf{c}_k \in \mathbf{C}^{k-1}$ ha tutte le componenti nulle eccetto al più l'ultima. Se $\mathbf{a}_k = \mathbf{0}$, si pone $A^{(k+1)} = A^{(k)}$ e $T_k = I$, altrimenti sia $\Pi_k \in \mathbf{R}^{(n-k)\times(n-k)}$ una matrice di permutazione tale che il vettore $\mathbf{a}'_k = \Pi_k \mathbf{a}_k$ abbia come prima componente una componente di $\mathbf{a}_k$ di modulo massimo, e si consideri la matrice elementare di Gauss $M_k \in \mathbf{C}^{(n-k)\times(n-k)}$ tale che il vettore

$$M_k \mathbf{a}'_k = \mathbf{a}''_k$$

abbia nulle tutte le componenti, esclusa la prima. Allora la matrice

$$A^{(k+1)} = T_k A^{(k)} T_k^{-1}, \quad T_k = \begin{bmatrix} I_k & O \\ O & M_k \end{bmatrix} \begin{bmatrix} I_k & O \\ O & \Pi_k \end{bmatrix},$$

ha la struttura

$$A^{(k+1)} = \begin{bmatrix} C^{(k+1)} & \mathbf{b}_{k+1} & D^{(k+1)} \\ \mathbf{c}_{k+1}^H & a_{k+1,k+1}^{(k+1)} & \mathbf{d}_{k+1}^H \\ O & \mathbf{a}_{k+1} & B^{(k+1)} \end{bmatrix} \begin{matrix} \} & k \text{ righe} \\ \} & 1 \text{ riga} \\ \} & n-k-1 \text{ righe.} \end{matrix}$$

Al termine del procedimento $A^{(n-1)}$ è in forma di Hessenberg superiore.

Per moltiplicare la matrice M_k per $\Pi_k B^{(k)}$ sono richieste $(n-k)^2$ operazioni moltiplicative, per moltiplicare la matrice $M_k \Pi_k B^{(k)} \Pi_k^T$ per M_k^{-1} sono richieste ancora $(n-k)^2$ operazioni moltiplicative e per moltiplicare la matrice $D^{(k)} \Pi_k^T$ per M_k^{-1} sono richieste $k(n-k)$ operazioni moltiplicative. Per trasformare la matrice $A^{(k)}$ nella matrice $A^{(k+1)}$ sono quindi richieste $(n-k)(2n-k)$ operazioni moltiplicative. Perciò per trasformare una matrice in forma di Hessenberg superiore, sono richieste $5n^3/6$ operazioni moltiplicative. Quindi il costo computazionale di questo metodo è inferiore a quello dei metodi di Householder e di Givens. Però con questo metodo possono presentarsi problemi di instabilità numerica, in particolare quando gli elementi delle matrici $A^{(k)}$ hanno modulo molto elevato rispetto agli elementi di A, come accade nel caso dei sistemi lineari. Può accadere infatti che il massimo dei moduli degli elementi di $A^{(k)}$ sia una funzione esponenziale di k. Se ciò accade, conviene utilizzare metodi di riduzione che fanno uso di matrici ortogonali (Householder e Givens) e che risultano più stabili.

6.25 Esempio. Facendo uso delle matrici elementari di Gauss, la matrice

$$A = \begin{bmatrix} 4 & 3 & 2 & 1 \\ 1 & 4 & 3 & 2 \\ 1 & 1 & 4 & 3 \\ 1 & 1 & 1 & 4 \end{bmatrix}$$

dell'esempio 6.24 è trasformata successivamente nelle matrici

$$A^{(2)} = \begin{bmatrix} 4 & 6 & 2 & 1 \\ 1 & 9 & 3 & 2 \\ 0 & -1 & 1 & 1 \\ 0 & -3 & -2 & 2 \end{bmatrix}$$

$$A^{(3)} = \begin{bmatrix} 4 & 6 & \frac{5}{3} & 2 \\ 1 & 9 & 3 & 3 \\ 0 & -3 & \frac{4}{3} & -2 \\ 0 & 0 & \frac{8}{9} & \frac{5}{3} \end{bmatrix}.$$

La matrice $A^{(3)}$ è in forma di Hessenberg superiore e differisce naturalmente dalle due matrici ottenute con i metodi di Householder e di Givens. ∎

Anche per le matrici in forma di Hessenberg superiore è possibile calcolare il valore assunto in un punto dal polinomio caratteristico senza determinarne effettivamente i coefficienti, con il seguente metodo di *Hyman*.

Sia $A \in \mathbf{C}^{n\times n}$ in forma di Hessenberg superiore e irriducibile. Fissato un punto λ, si determinano un vettore $\mathbf{x}$ con l'ultima componente $x_n = 1$ e uno scalare γ, dipendenti da λ, tali che

$$(A - \lambda I)\mathbf{x} = \gamma \mathbf{e}_1, \tag{24}$$

nel modo seguente: si ricava x_{n-1} dall'ultima equazione e procedendo mediante sostituzioni all'indietro, si ricava infine x_1 dalla seconda equazione e γ dalla prima equazione. Il costo computazionale di questo procedimento è di $n^2/2$ operazioni moltiplicative. Poiché per la regola di Cramer risulta

$$x_n = \frac{(-1)^{n+1}}{\det(A - \lambda I)}\ \gamma a_{21}a_{32}\dots a_{n,n-1},$$

essendo $x_n = 1$, si ha

$$P(\lambda) = \det(A - \lambda I) = (-1)^{n+1}\gamma a_{21}a_{32}\dots a_{n,n-1}. \tag{25}$$

È possibile, in modo analogo, calcolare anche la derivata prima

$$P'(\lambda) = \frac{\mathrm{d}}{\mathrm{d}\lambda}\ \det(A - \lambda I).$$

Derivando entrambi i membri della (24) si ha

$$(A - \lambda I)\mathbf{x}' - \mathbf{x} = \gamma' \mathbf{e}_1,$$

da cui, attraverso il processo di sostituzione all'indietro, è possibile ricavare $\mathbf{x}'$ e γ' dopo avere calcolato $\mathbf{x}$ dal sistema (24). Dalla (25) si ha poi:

$$P'(\lambda) = (-1)^{n+1}\gamma' a_{21}a_{32}\dots a_{n,n-1}.$$

8. Metodo QR per il calcolo degli autovalori

Il metodo QR è il metodo più usato per calcolare tutti gli autovalori di una matrice, in quanto è il più efficiente e può essere applicato anche a matrici non hermitiane. Il metodo è assai complicato, sia come descrizione che come implementazione, anche se il principio su cui si basa è semplice. Il metodo richiede tutta una serie di accorgimenti, senza i quali non potrebbe essere efficiente: riduzione preliminare della matrice in forma tridiagonale o di Hessenberg superiore, per ridurre il costo computazionale ad ogni iterazione; utilizzazione di una tecnica di traslazione per aumentare la velocità di convergenza; riduzione dell'ordine della matrice quando un autovalore è

stato approssimato con sufficiente precisione, per calcolare un altro autovalore.

Il metodo QR, che è stato descritto da Francis nel 1961, utilizza la fattorizzazione QR di una matrice; esso deriva da un precedente metodo, detto *metodo LR*, proposto da Rutishauser nel 1958, che utilizza la fattorizzazione LU di una matrice.

La descrizione del metodo si articola nei seguenti punti:

a) algoritmo di base,

b) teorema di convergenza,

c) costo computazionale e stabilità,

d) convergenza in ipotesi più deboli,

e) condizioni di arresto e riduzione dell'ordine della matrice,

f) tecnica di traslazione,

g) calcolo degli autovettori.

a) Algoritmo di base

Nel metodo QR viene generata una successione $\{A_k\}$ di matrici nel modo seguente: posto

$$A_1 = A,$$

per $k = 1, 2, \ldots$, si calcola una fattorizzazione QR di A_k

$$A_k = Q_k R_k, \tag{26}$$

dove Q_k è unitaria e R_k è triangolare superiore, e si definisce la matrice A_{k+1} per mezzo della relazione

$$A_{k+1} = R_k Q_k. \tag{27}$$

Da (26) e (27) risulta che

$$A_{k+1} = Q_k^H A_k Q_k, \tag{28}$$

e quindi le matrici della successione $\{A_k\}$ sono tutte simili fra di loro. Sotto opportune ipotesi la successione converge ad una matrice triangolare superiore (diagonale se A è hermitiana) che ha come elementi diagonali gli autovalori di A.

6.26 Esempio. Il metodo QR viene applicato alla matrice

$$A = \begin{bmatrix} 4 & 3 & 2 & 1 \\ 3 & 4 & 3 & 2 \\ 2 & 3 & 4 & 3 \\ 1 & 2 & 3 & 4 \end{bmatrix}.$$

dell'esempio 6.17. Si ottiene

$$A_2 = \begin{bmatrix} 9.733320 & 2.834947 & 0.8783645 & -0.2318690 \\ 2.834947 & 3.783903 & 1.539931 & -0.6027983 \\ 0.8783645 & 1.539931 & 1.515015 & -0.5091640 \\ -0.2318690 & -0.6027983 & -0.5091640 & 0.9677416 \end{bmatrix},$$

$$A_3 = \begin{bmatrix} 10.95491 & 1.039794 & 0.8115560\ 10^{-1} & 0.1726981\ 10^{-1} \\ 1.039794 & 3.494645 & 0.3880183 & 0.1244645 \\ 0.8115560\ 10^{-1} & 0.3880183 & 0.8236489 & 0.1754468 \\ 0.1726981\ 10^{-1} & 0.1244645 & 0.1754468 & 0.7267331 \end{bmatrix},$$

$$\vdots$$

Gli elementi non principali formano successioni decrescenti in modulo, e dopo 9 iterazioni la matrice A_{10} è data da

$$\begin{bmatrix} 11.09831 & 0.2762247\ 10^{-3} & 0.1729076\ 10^{-8} & -0.2617231\ 10^{-10} \\ 0.2762247\ 10^{-3} & 3.414135 & 0.3305371\ 10^{-4} & -0.7052673\ 10^{-6} \\ 0.1729076\ 10^{-8} & 0.3305372\ 10^{-4} & 0.9003896 & -0.1320110\ 10^{-1} \\ -0.2617231\ 10^{-10} & -0.7052673\ 10^{-6} & -0.1320110\ 10^{-1} & 0.5863345 \end{bmatrix}$$

L'elemento non principale di massimo modulo è dell'ordine di 10^{-2}. Ripetendo il procedimento fino a quando il massimo modulo degli elementi non principali è minore di 10^{-4}, gli elementi sulla diagonale principale alla 21-esima iterazione sono

$$11.09720 \quad 3.414135 \quad 0.9008932 \quad 0.5857800,$$

che si assumono come approssimazioni degli autovalori di A. ∎

b) Teorema di convergenza

Il seguente teorema di convergenza viene dato con ipotesi piuttosto restrittive, allo scopo di renderne più semplice la dimostrazione. La convergenza del metodo si può dimostrare anche con ipotesi assai più deboli, che verranno esaminate in seguito.

6.27 Teorema. *Sia $A \in \mathbf{C}^{n\times n}$ tale che i suoi autovalori λ_i, $i = 1, 2, \ldots, n$, abbiano moduli tutti distinti, cioè*

$$|\lambda_1| > |\lambda_2| > \ldots > |\lambda_n| > 0. \tag{29}$$

Indicata con X la matrice degli autovettori di A, tale che

$$A = XDX^{-1}, \tag{30}$$

in cui D è la matrice diagonale il cui i-esimo elemento principale è λ_i, si supponga che la matrice X^{-1} ammetta la fattorizzazione LU. Allora esistono delle matrici di fase S_k tali che

$$\lim_{k\to\infty} S_k^H R_k S_{k-1} = \lim_{k\to\infty} S_{k-1}^H A_k S_{k-1} = T, \tag{31}$$

e

$$\lim_{k\to\infty} S_{k-1}^H Q_k S_k = I,$$

dove T è triangolare superiore con gli elementi principali uguali a $\lambda_1, \lambda_2, \ldots, \lambda_n$. Quindi gli elementi principali di A_k tendono agli autovalori di A. Se A è una matrice hermitiana, allora T è diagonale.

Dim. Il teorema viene dimostrato confrontando due fattorizzazioni QR della matrice A^k ottenute in due modi diversi. Una prima fattorizzazione è data dalla seguente relazione

$$A^k = H_k U_k, \tag{32}$$

dove

$$H_k = Q_1 Q_2 \ldots Q_k$$

è una matrice unitaria e

$$U_k = R_k R_{k-1} \ldots R_1$$

è una matrice triangolare superiore. Per dimostrare la (32) si procede per induzione: per $k = 1$ risulta $A = A_1 = H_1 U_1$. Per $k > 1$, supposta valida la (32), da (26) e (27) si ottiene

$$Q_k A_{k+1} = A_k Q_k,$$

da cui

$$Q_1 \ldots Q_{k-1} Q_k A_{k+1} = Q_1 \ldots Q_{k-1} A_k Q_k = \ldots = A Q_1 \ldots Q_{k-1} Q_k \tag{33}$$

e quindi

$$\begin{aligned} H_{k+1} U_{k+1} &= Q_1 \ldots Q_k Q_{k+1} R_{k+1} R_k \ldots R_1 \\ &= Q_1 \ldots Q_{k-1} Q_k A_{k+1} R_k R_{k-1} \ldots R_1 \\ &= A Q_1 \ldots Q_{k-1} Q_k R_k R_{k-1} \ldots R_1 = A H_k U_k = A^{k+1}, \end{aligned}$$

cioè $A^{k+1} = H_{k+1} U_{k+1}$.

Una seconda fattorizzazione QR della matrice A^k viene ottenuta dalla relazione (30). Sia $X^{-1} = LU$ la fattorizzazione LU di X^{-1}. Allora

$$A^k = XD^kX^{-1} = XD^kLU = XD^kLD^{-k}D^kU.$$

Poiché gli elementi della matrice D^kLD^{-k} sono dati da

$$\begin{cases} l_{ij}\left(\dfrac{\lambda_i}{\lambda_j}\right)^k & \text{per } i > j, \\ 1 & \text{per } i = j, \\ 0 & \text{per } i < j, \end{cases} \tag{34}$$

e $|\lambda_i| < |\lambda_j|$ per $i > j$, si può porre

$$D^kLD^{-k} = I + E_k,$$

dove

$$\lim_{k\to\infty} E_k = 0,$$

e quindi è

$$A^k = X(I + E_k)D^kU.$$

Indicata con

$$X = QR$$

una fattorizzazione QR della matrice X, si ha

$$A^k = QR(I + E_k)D^kU = Q(I + RE_kR^{-1})RD^kU,$$

e indicata con

$$I + RE_kR^{-1} = P_kT_k \tag{35}$$

una fattorizzazione QR della matrice $I + RE_kR^{-1}$, si ha

$$A^k = (QP_k)\ (T_kRD^kU). \tag{36}$$

La (36) dà una seconda fattorizzazione QR di A^k: infatti QP_k è unitaria e T_kRD^kU è triangolare superiore.

Poiché la fattorizzazione QR di una matrice è unica a meno di una matrice di fase, confrontando le due fattorizzazioni di A^k ottenute, cioè la (32) e la (36) segue che esiste una matrice di fase $\hat{S}_k$ tale che

$$H_k = QP_k\hat{S}_k^H \quad \text{e} \quad U_k = \hat{S}_kT_kRD^kU.$$

Risulta

$$Q_k = (H_{k-1})^{-1}H_k = \hat{S}_{k-1}P_{k-1}^HQ^HQP_k\hat{S}_k^H = \hat{S}_{k-1}P_{k-1}^HP_k\hat{S}_k^H,$$

da cui

$$\hat{S}_{k-1}^H Q_k \hat{S}_k = P_{k-1}^H P_k,$$

e

$$\begin{aligned} R_k = U_k (U_{k-1})^{-1} &= \hat{S}_k T_k R D^k U U^{-1} D^{-k+1} R^{-1} T_{k-1}^{-1} \hat{S}_{k-1}^H \\ &= \hat{S}_k T_k R D R^{-1} T_{k-1}^{-1} \hat{S}_{k-1}^H, \end{aligned}$$

e quindi

$$\hat{S}_k^H R_k \hat{S}_{k-1} = T_k R D R^{-1} T_{k-1}^{-1}.$$

Poiché $\lim_{k\to\infty} E_k = 0$, per la (35) risulta

$$\lim_{k\to\infty} (I + R E_k R^{-1}) = \lim_{k\to\infty} P_k T_k = I,$$

e quindi (si veda l'esercizio 6.30) esiste una matrice di fase $\check{S}_k$ tale che

$$\lim_{k\to\infty} P_k \check{S}_k = \lim_{k\to\infty} \check{S}_k^H T_k = I.$$

Allora posto $S_k = \hat{S}_k \check{S}_k$, è

$$\lim_{k\to\infty} S_{k-1}^H Q_k S_k = \lim_{k\to\infty} P_{k-1}^H P_k = I,$$

$$\lim_{k\to\infty} S_k^H R_k S_{k-1} = \lim_{k\to\infty} T_k R D R^{-1} T_{k-1}^{-1} = R D R^{-1},$$

e

$$\begin{aligned} \lim_{k\to\infty} S_{k-1}^H A_k S_{k-1} &= \lim_{k\to\infty} S_{k-1}^H Q_k R_k S_{k-1} = \lim_{k\to\infty} S_{k-1}^H Q_k S_k S_k^H R_k S_{k-1} \\ &= \lim_{k\to\infty} S_k^H R_k S_{k-1} = R D R^{-1}. \end{aligned}$$

La matrice $T = RDR^{-1}$ è triangolare superiore e quindi per gli elementi diagonali di A_k vale

$$\lim_{k\to\infty} a_{jj}^{(k)} = \lambda_j.$$

Se A è hermitiana, dalla (28) segue che le matrici A_k, e quindi le matrici $S_{k-1}^H A_k S_{k-1}$, sono hermitiane. Dalla (31) segue allora che T è hermitiana e quindi diagonale. ∎

c) Costo computazionale e stabilità

Il metodo QR applicato a una matrice di ordine n ha ad ogni passo un costo computazionale dell'ordine di n^3 operazioni moltiplicative (per calcolare la fattorizzazione $A_k = Q_k R_k$ e per moltiplicare la matrice triangolare R_k per le matrici elementari della fattorizzazione). Per abbassare il costo

computazionale globale conviene prima trasformare la matrice A in forma di Hessenberg superiore. Questa trasformazione viene eseguita una sola volta perché il metodo QR, applicato a matrici in forma di Hessenberg superiore produce matrici A_k in forma di Hessenberg superiore. Infatti se A_k è in forma di Hessenberg superiore, la matrice Q_k è data dal prodotto di $n-1$ matrici elementari di Householder (o di Givens) che sono in forma di Hessenberg superiore e quindi la matrice A_{k+1}, prodotto di una matrice triangolare superiore R_k per una matrice Q_k in forma di Hessenberg superiore, risulta ancora in forma di Hessenberg superiore. Se la matrice A è hermitiana, la matrice in forma di Hessenberg superiore, ottenuta applicando ad A i metodi di Householder o di Givens, è ancora hermitiana, e quindi risulta tridiagonale. Inoltre anche tutte le matrici A_k generate dal metodo QR sono hermitiane e quindi tridiagonali.

Il metodo QR applicato a una matrice A in forma di Hessenberg superiore ha ad ogni passo un costo computazionale di $2n^2$ operazioni moltiplicative (che è il costo computazionale per calcolare la fattorizzazione $A_k = Q_k R_k$, infatti il numero delle operazioni richieste per moltiplicare la matrice triangolare R_k per le matrici elementari della fattorizzazione è di ordine inferiore al secondo). Se A è una matrice tridiagonale, il costo computazionale di ogni passo del metodo è lineare in n.

In [28] viene dimostrato che il metodo QR gode delle stesse proprietà di stabilità di cui gode la fattorizzazione QR di una matrice.

6.28 Esempio. Il metodo QR viene applicato alla matrice tridiagonale

$$A_1 = \begin{bmatrix} 4 & 3.741654 & 0 & 0 \\ 3.741654 & 8.285707 & 2.602977 & 0 \\ 0 & 2.602977 & 3.039581 & 0.2254009 \\ 0 & 0 & 0.2254009 & 0.6746972 \end{bmatrix},$$

ottenuta con il metodo di Givens nell'esempio 6.17 dalla matrice A di cui sono stati approssimati gli autovalori nell'esempio 6.26. Si ottiene

$$A_2 = \begin{bmatrix} 9.733309 & 2.976943 & 0 & 0 \\ 2.976943 & 4.547497 & 0.7894507 & 0 \\ 0 & 0.7894507 & 1.094460 & \text{-}0.1081211 \\ 0 & 0 & \text{-}0.1081211 & 0.6246958 \end{bmatrix},$$

$$A_3 = \begin{bmatrix} 10.95485 & 1.043100 & 0 & 0 \\ 1.043100 & 3.542464 & 0.2006968 & 0 \\ 0 & 0.2006968 & 0.8992355 & 0.7246423\ 10^{-1} \\ 0 & 0 & 0.7246423\ 10^{-1} & 0.6033282 \end{bmatrix},$$

$$\vdots$$

Ripetendo il procedimento fino a quando il massimo modulo degli elementi non principali è minore di 10^{-4}, alla 18-esima iterazione gli elementi principali sono

$$11.09809 \quad 3.414161 \quad 0.9008972 \quad 0.5857840,$$

che si assumono come approssimazioni degli autovalori di A. ∎

d) Convergenza in ipotesi più deboli

La dimostrazione della convergenza del metodo QR è stata fatta nell'ipotesi che la matrice X^{-1} fosse fattorizzabile nella forma LU. In questo caso gli elementi principali di T coincidono, nell'ordine, con $\lambda_1, \lambda_2, \ldots, \lambda_n$. Se X^{-1} non ammette fattorizzazione LU, si può dimostrare [28] che il metodo QR è ancora convergente. In questo caso gli elementi principali di T coincidono ancora con i λ_i, ma non sono più in ordine di modulo decrescente.

Se l'ipotesi (29) del teorema 6.27, che tutti gli autovalori abbiano modulo distinto, non è verificata, la successione formata dagli elementi diagonali di A_k non converge. Questa ipotesi è troppo restrittiva, e non consente di utilizzare il metodo QR in casi particolarmente importanti nelle applicazioni, come quelli in cui la matrice A ha elementi reali e autovalori non reali. Però anche in questo caso il metodo QR può essere applicato con opportune varianti. Sia ad esempio

$$|\lambda_1| > \ldots > |\lambda_r| = |\lambda_{r+1}| > \ldots > |\lambda_n| > 0,$$

dove λ_r e λ_{r+1} sono due numeri complessi coniugati, oppure due numeri reali. Allora nella (34) la successione degli elementi

$$l_{r+1,r}\Big(\frac{\lambda_{r+1}}{\lambda_r}\Big)^k$$

non converge a zero per $k \to \infty$. Ne segue che le matrici P_k, e quindi le matrici $S_{k-1}^H Q_k S_k$, non convergono alla matrice I per la presenza nella posizione $(r+1, r)$ di elementi che non tendono a zero. Sia

$$A_r^{(k)} = \begin{bmatrix} a_{rr}^{(k)} & a_{r,r+1}^{(k)} \\ a_{r+1,r}^{(k)} & a_{r+1,r+1}^{(k)} \end{bmatrix}$$

la sottomatrice principale di ordine 2 di A_k formata dalle righe e colonne di indici r e $r+1$. La successione $\{A_r^{(k)}\}$ non converge, ma gli autovalori delle sottomatrici $A_r^{(k)}$ convergono a λ_r e λ_{r+1} [28]. Gli elementi principali di A_k di indice diverso da r e $r+1$ convergono agli altri autovalori. Situazioni analoghe si presentano quando la matrice A ha più autovalori di modulo

uguale e in questo caso il metodo QR genera matrici R_k con struttura triangolare a blocchi, in cui gli autovalori dei blocchi diagonali convergono ad autovalori di A.

6.29 Esempio. Si applica il metodo QR alla matrice

$$A_1 = \begin{bmatrix} 4 & 3.464101 & -1.388729 & 0.2672636 \\ 1.732049 & 7.666669 & 1.158131 & -0.4629126 \\ 0 & -1.247218 & 2.476188 & 0.7423077 \\ 0 & 0 & 0.9897417 & 1.857139 \end{bmatrix},$$

in forma di Hessenberg superiore ottenuta nell'esempio 6.24. Si ottiene

$$A_2 = \begin{bmatrix} 6.473673 & 4.062625 & 0.3739953\ 10^{-1} & -0.7850719\ 10^{-1} \\ 2.302595 & 4.968325 & 2.265884 & -0.9043741\ 10^{-1} \\ 0 & -0.6322317 & 2.717972 & -0.9186863 \\ 0 & 0 & 0.6584795 & 1.840011 \end{bmatrix},$$

$$A_3 = \begin{bmatrix} 8.314364 & 2.673093 & 1.266714 & 0.3786074 \\ 1.132446 & 2.832693 & 2.120341 & 0.3386071 \\ 0 & -0.5868980 & 2.906489 & 1.142451 \\ 0 & 0 & -0.4178842 & 1.946413 \end{bmatrix},$$

$$\vdots$$

$$A_{10} = \begin{bmatrix} 8.783114 & -1.744292 & -1.348724 & -0.9030869 \\ 0.3378619\ 10^{-3} & 2.480618 & 1.925506 & 0.7184988 \\ 0 & -0.7168258 & 2.609033 & 0.9730009 \\ 0 & 0 & 0.6583786\ 10^{-1} & 2.126765 \end{bmatrix},$$

$$A_{11} = \begin{bmatrix} 8.783008 & -1.301571 & -1.799058 & 0.8643191 \\ 0.9932932\ 10^{-4} & 2.168211 & 1.806849 & -0.3811607 \\ 0 & -0.8444027 & 2.947040 & -1.116467 \\ 0 & 0 & -0.4549125\ 10^{-1} & 2.101224 \end{bmatrix}.$$

Come si può notare, le successioni degli elementi di indici (2,1) e (4,3) sono decrescenti in modulo, più rapidamente la prima, più lentamente la seconda, mentre questo non accade nella successione delle sottomatrici principali formate dagli elementi delle righe e colonne di indici 2 e 3. Ripetendo il procedimento fino a quando l'elemento di indici (4,3) risulta inferiore in modulo a 10^{-4}, alla 30-esima iterazione gli elementi principali di indici 1 e 4 sono

$$a_{11}^{(31)} = 8.782016, \quad a_{44}^{(31)} = 2.089449,$$

che sono delle buone approssimazioni degli autovalori λ_1 e λ_4 di massimo e minimo modulo. La sottomatrice principale $A_2^{(31)}$ di ordine 2 risulta

$$A_2^{(31)} = \begin{bmatrix} 2.576344 & 1.940763 \\ -0.6846419 & 2.550408 \end{bmatrix},$$

da cui si ricavano per λ_2 e λ_3 le approssimazioni

$$\lambda_2 = 2.563376 + \mathbf{i}\, 1.152632\ , \quad \lambda_3 = 2.563376 - \mathbf{i}\, 1.152632\ . \qquad \blacksquare$$

e) Condizioni di arresto e riduzione dell'ordine della matrice

Fissato un valore ϵ di tolleranza, si procede applicando il metodo QR alla matrice A in forma di Hessenberg superiore fino a quando per un indice $p,\ 1 \le p < n$, l'elemento $a_{p+1,p}^{(k)}$ diventa sufficientemente piccolo. Un criterio utilizzato è il seguente

$$|a_{p+1,p}^{(k)}| < \epsilon(|a_{pp}^{(k)}| + |a_{p+1,p+1}^{(k)}|). \tag{37}$$

Quando la condizione (37) è verificata, nella matrice A_k

$$A_k = \begin{bmatrix} B_k & D_k \\ \\ E_k & C_k \end{bmatrix} \begin{matrix} \} & p \text{ righe} \\ \\ \} & n-p \text{ righe} \end{matrix}$$

dove $B_k \in \mathbf{C}^{p\times p}, C_k \in \mathbf{C}^{(n-p)\times(n-p)}$, la sottomatrice E_k ha un elemento di modulo piccolo e gli altri tutti nulli. Si procede quindi operando separatamente con le matrici B_k e C_k. Se la matrice A è hermitiana, gli autovalori di B_k e C_k sono delle buone approssimazioni degli autovalori di A_k (si veda l'esercizio 7.3).

f) Tecnica di traslazione

La velocità di convergenza del metodo QR dipende per la (34) dai rapporti $|\lambda_i/\lambda_j|$ per $i > j$, e quindi per l'ipotesi (29) dal numero

$$\max_{1\le i\le n-1} \left|\frac{\lambda_{i+1}}{\lambda_i}\right|. \tag{38}$$

Se tale rapporto è vicino ad 1, la convergenza può essere lenta. In questo caso per accelerare la convergenza si utilizza una tecnica di traslazione dello spettro degli autovalori di A, detta *di shift*.

Sia μ un numero che approssima un autovalore λ meglio degli altri autovalori. Le matrici Q_k e R_k, generate dal metodo QR a partire dalla

matrice $A - \mu I$ possono essere costruite anche per mezzo delle seguenti relazioni (*metodo QR con shift*)

$$\left.\begin{aligned} A_k - \mu I &= Q_k R_k, \\ A_{k+1} &= R_k Q_k + \mu I, \end{aligned}\right\} \quad \text{per } k = 1, 2, \ldots$$

e risulta

$$Q_k A_{k+1} = A_k Q_k - \mu Q_k + \mu Q_k = A_k Q_k.$$

Tenendo presente che gli autovalori di $A - \mu I$ sono $\lambda_i - \mu$ e che la velocità di convergenza è regolata dalla (38), è possibile scegliere un parametro μ in modo da accelerare la convergenza del metodo QR con shift. È conveniente scegliere per μ un valore che approssima λ_n. Ciò può essere ottenuto applicando il metodo QR inizialmente senza shift per un certo numero p di iterazioni, e scegliendo $\mu = a_{nn}^{(p)}$ per le successive iterazioni con shift.

Poiché μ può essere modificato ad ogni iterazione è più conveniente scegliere

$$\mu_k = a_{nn}^{(k)}, \quad k = 1, 2, \ldots. \tag{39}$$

Nel caso delle matrici hermitiane è possibile dimostrare [29] che con questa strategia la convergenza a zero dell'elemento $a_{n,n-1}^{(k)}$ è del terzo ordine (si veda anche l'esercizio 6.31).

Quando la (37) è verificata per $p = n-1$, si passa a operare sulla matrice B_k di ordine $n - 1$ ottenuta dalla matrice A_k eliminando l'ultima riga e l'ultima colonna. Per l'approssimazione degli altri autovalori si procede in modo analogo.

6.30 Esempio. Si applica il metodo QR con lo shift (39) alla matrice tridiagonale

$$A_1 = \begin{bmatrix} 4 & 3.741654 & 0 & 0 \\ 3.741654 & 8.285707 & 2.602977 & 0 \\ 0 & 2.602977 & 3.039581 & 0.2254009 \\ 0 & 0 & 0.2254009 & 0.6746972 \end{bmatrix},$$

ottenuta con il metodo di Givens nell'esempio 6.17, di cui sono stati approssimati gli autovalori negli esempi 6.26 e 6.28. Si ha

$$A_2 = \begin{bmatrix} 10.11023 & 2.576269 & 0 & 0 \\ 2.576269 & 4.380260 & 0.2509962 & 0 \\ 0 & 0.2509962 & 0.9084192 & 0.6795645\ 10^{-1} \\ 0 & 0 & 0.6795645\ 10^{-1} & 0.6010619 \end{bmatrix},$$

$$A_3 = \begin{bmatrix} 11.01885 & 0.7804608 & 0 & 0 \\ 0.7804608 & 3.494054 & 0.2471317\ 10^{-1} & 0 \\ 0 & 0.2471317\ 10^{-1} & 0.9011788 & 0.3621320\ 10^{-2} \\ 0 & 0 & 0.3621320\ 10^{-2} & 0.5858268 \end{bmatrix},$$

$$A_4 = \begin{bmatrix} 11.09302 & 0.2120095 & 0 & 0 \\ 0.2120095 & 3.420040 & 0.2740411\ 10^{-2} & 0 \\ 0 & 0.2740411\ 10^{-2} & 0.9009814 & 0.4774449\ 10^{-6} \\ 0 & 0 & 0.4774449\ 10^{-6} & 0.5857852 \end{bmatrix}.$$

Poiché l'elemento $a_{43}^{(4)}$ soddisfa alla condizione (37) con $\epsilon = 10^{-6}$, si passa a operare sulla sottomatrice di ordine 3 ottenuta eliminando l'ultima riga e colonna. Dopo altre 3 iterazioni, si ottiene la matrice

$$A_7 = \begin{bmatrix} 11.09874 & 0.4131589\ 10^{-2} & 0 \\ 0.4131589\ 10^{-2} & 3.414158 & \text{-}0.3791176\ 10^{-5} \\ 0 & \text{-}0.3791176\ 10^{-5} & 0.9009783 \end{bmatrix},$$

che può essere a sua volta ridotta. Gli altri due autovalori possono essere calcolati direttamente dalla sottomatrice principale di testa di ordine 2. Si ottengono così le approssimazioni degli autovalori di A

$$\lambda_1 = 11.09835, \quad \lambda_2 = 3.414142, \quad \lambda_3 = 0.9009783, \quad \lambda_4 = 0.5857852. \quad \blacksquare$$

Il metodo QR con shift può essere applicato anche ai casi in cui esistono più autovalori con lo stesso modulo. In particolare, se $|\lambda_{n-1}| = |\lambda_n|$, allora conviene scegliere come $\mu^{(k)}$, $k = 1, 2, \ldots$, l'autovalore della sottomatrice

$$A_{n-1}^{(k)} = \begin{bmatrix} a_{n-1,n-1}^{(k)} & a_{n-1,n}^{(k)} \\ a_{n,n-1}^{(k)} & a_{nn}^{(k)} \end{bmatrix}$$

che è più vicino ad $a_{nn}^{(k)}$.

In questo caso, anche se la matrice A ha elementi reali, l'utilizzazione dello shift può portare ad una matrice A_k ad elementi complessi, con conseguente aumento del costo computazionale. Questo può essere evitato eseguendo due iterazioni successive

$$A_k \to A_{k+1} \to A_{k+2},$$

e usando come costanti di traslazione $\mu^{(k)} = \alpha$ e $\mu^{(k+1)} = \beta$, dove α e β sono i due autovalori della sottomatrice $A_{n-1}^{(k)}$. Si ha infatti

$$A_k - \alpha I = Q_k R_k,$$

$$\begin{aligned} A_{k+1} &= R_k Q_k + \alpha I, \\ A_{k+1} - \beta I &= Q_{k+1} R_{k+1}, \\ A_{k+2} &= R_{k+1} Q_{k+1} + \beta I, \end{aligned} \tag{40}$$

e quindi

$$\begin{aligned} Q_k Q_{k+1} R_{k+1} R_k &= Q_k (A_{k+1} - \beta I) R_k = Q_k (R_k Q_k + \alpha I - \beta I) R_k \\ &= Q_k R_k (Q_k R_k + \alpha I - \beta I) = (A_k - \alpha I)\ (A_k - \beta I). \end{aligned} \tag{41}$$

La matrice $M = (A_k - \alpha I)(A_k - \beta I)$ ha elementi reali se A_k è reale, perché α e β sono radici di un'equazione di secondo grado a coefficienti reali. Ne segue che ponendo

$$Z = Q_k Q_{k+1} \quad \text{e} \quad S = R_{k+1} R_k,$$

dalla (41) si ricava che ZS è una fattorizzazione QR della matrice reale M e quindi Z e S sono, a meno di moltiplicazione per una matrice di fase, matrici reali rispettivamente ortogonale e triangolare superiore. D'altra parte dalle (40) risulta che

$$\begin{aligned} Z A_{k+2} &= Q_k Q_{k+1} A_{k+2} = Q_k Q_{k+1} R_{k+1} Q_{k+1} + \beta Q_k Q_{k+1} = Q_k A_{k+1} Q_{k+1} \\ &= Q_k R_k Q_k Q_{k+1} + \alpha Q_k Q_{k+1} = A_k Q_k Q_{k+1} = A_k Z, \end{aligned}$$

da cui

$$A_{k+2} = Z^H A_k Z.$$

È quindi possibile ricavare A_{k+2} direttamente da A_k utilizzando la fattorizzazione QR della matrice reale M. Questo modo di procedere però ha un consistente costo computazionale, in quanto la sola costruzione della matrice M, che non è in forma di Hessenberg superiore anche se lo è la A_k, richiede $n^3/6$ operazioni moltiplicative.

Per superare questo inconveniente si utilizza il seguente procedimento suggerito da Francis, che richiede $6n^2$ operazioni moltiplicative:

1. si costruisce la prima colonna $\mathbf{m}_1$ della matrice M e la matrice elementare di Householder P_0 tale che

$$P_0 \mathbf{m}_1 = \gamma \mathbf{e}_1, \quad \text{dove} \quad |\gamma| = \|\mathbf{m}_1\|_2.$$

2. si costruiscono le matrici di Householder $P_1, P_2, \ldots, P_{n-2}$ tali che, posto $Z' = P_0 P_1 \ldots P_{n-2}$, la matrice $(Z')^H A_k Z'$ sia in forma di Hessenberg superiore. È possibile dimostrare che le matrici

$$A_{k+2} = Z^H A_k Z \quad \text{e} \quad A'_{k+2} = (Z')^H A_k Z'$$

sono "essenzialmente" uguali, cioè uguali nel senso che esiste una matrice di fase reale D, tale che

$$A_{k+2} = D^{-1} A'_{k+2} D$$

(si vedano come esempio di matrici essenzialmente uguali le due matrici $A^{(3)}$ e $H^{(3)}$ ottenute nell'esempio 6.17).

6.31 Esempio. Applicando il metodo QR con lo shift (39) alla matrice

$$A_1 = \begin{bmatrix} 4 & 3.464101 & -1.388729 & 0.2672636 \\ 1.732049 & 7.666669 & 1.158131 & -0.4629126 \\ 0 & -1.247218 & 2.476188 & 0.7423077 \\ 0 & 0 & 0.9897417 & 1.857139 \end{bmatrix},$$

in forma di Hessenberg superiore ottenuta nell'esempio 6.24, si ottiene (il metodo senza shift è stato applicato alla matrice A_1 nell'esempio 6.29)

$$A_2 = \begin{bmatrix} 7.989221 & 2.992575 & 1.020255 & -0.6292015 \\ 1.667174 & 3.078217 & 2.177103 & -0.5580172 \\ 0 & -0.7987912 & 2.882763 & -1.142143 \\ 0 & 0 & 0.1381161 & 2.049773 \end{bmatrix},$$

$$A_3 = \begin{bmatrix} 8.842974 & -1.227221 & 1.630806 & 0.8734073 \\ 0.2213716 & 2.623399 & 0.7493563 & -0.9821870 \\ 0 & -1.915713 & 2.440961 & 0.6386327 \\ 0 & 0 & -0.3020395\ 10^{-2} & 2.092627 \end{bmatrix}.$$

Dopo altre due iterazioni l'elemento $a_{43}^{(5)}$ soddisfa alla condizione (37) con $\epsilon = 10^{-10}$. Il valore

$$a_{44}^{(5)} = 2.089537$$

viene assunto come approssimazione di λ_4 e si passa a operare sulla sottomatrice di ordine 3 ottenuta eliminando l'ultima riga e l'ultima colonna. In questa matrice gli autovalori di minimo modulo sono due e sono complessi, per cui si applica il procedimento suggerito da Francis, ottenendo dopo altre 2 iterazioni la matrice

$$A_9 = \begin{bmatrix} 8.783265 & 1.081047 & -1.955143 \\ 0.2728484\ 10^{-11} & 2.064380 & -1.694178 \\ 0 & 0.9313574 & 3.062660 \end{bmatrix},$$

in cui l'elemento $a_{21}^{(9)}$ è in modulo minore di 10^{-11}. L'elemento $a_{11}^{(9)}$ viene assunto come approssimazione dell'autovalore λ_1, mentre gli autovalori λ_2 e

λ_3 vengono approssimati calcolando gli autovalori della sottomatrice principale $A_2^{(9)}$. ∎

g) Calcolo degli autovettori

Con il metodo QR si ottiene la forma normale di Schur della matrice A. Infatti dalla (33) si ha

$$H_k A_{k+1} = A H_k$$

e per la (31) è

$$\lim_{k\to\infty} S_k^H H_k^H A H_k S_k = T,$$

in cui T è una matrice triangolare superiore e $H_k S_k$ è una matrice unitaria. Se la matrice A è normale, è facile dimostrare (si veda l'esercizio 6.29) che esiste una sottosuccessione $\{ H_{k_i} S_{k_i} \}$ della successione $\{ H_k S_k \}$ che converge alla matrice unitaria le cui colonne sono gli autovettori di A. Il costo computazionale del calcolo degli autovettori è elevato perché ad ogni passo è richiesta la costruzione e la memorizzazione della matrice $H_k = H_{k-1} Q_k$. Per il calcolo degli autovettori conviene ricorrere al metodo delle potenze con la variante di Wielandt, descritto più avanti.

9. Metodo di Jacobi

Il *metodo di Jacobi* è un metodo classico per calcolare gli autovalori e gli autovettori di matrici hermitiane. Per la semplicità con cui può essere implementato viene ancora oggi usato nel caso di matrici di piccole dimensioni, quando sono richiesti tutti gli autovalori. Inoltre questo metodo si presta bene per una utilizzazione in ambiente di calcolo parallelo, dove si assume che ad ogni passo possano essere effettuate simultaneamente p operazioni aritmetiche, con $p > 1$.

Il metodo di Jacobi è un metodo iterativo che utilizza trasformazioni della forma (16)

$$A^{(1)} = A, \quad A^{(k+1)} = T_k^{-1} A^{(k)} T_k, \quad k = 1, 2, \ldots,$$

in cui le matrici T_k sono matrici di Givens. T_k viene scelta in modo da rendere nullo un opportuno elemento non principale di $A^{(k+1)}$. La successione $\{A^{(k)}\}$, se è convergente, converge ad una matrice diagonale D.

Considerando per semplicità il caso in cui $A^{(k)}$ è reale e seguendo la notazione del paragrafo 5, in cui si indicano con a_{rj} gli elementi di $A^{(k)}$ e con $\hat{a}_{rj}$ gli elementi di $A^{(k+1)}$, la matrice $T_k = G_{pq}$, viene determinata in modo che risulti $\hat{a}_{pq} = 0$ (se $a_{pq} = 0$, basta porre $T_k = I$). Dalla (17), posto $t = \text{tg}\ \phi$, risulta

$$(1 - t^2) a_{pq} - t(a_{pp} - a_{qq}) = 0,$$

da cui si ottiene l'equazione

$$t^2 + 2mt - 1 = 0,$$

dove

$$m = \frac{a_{pp} - a_{qq}}{2a_{pq}}.$$

Fra le due soluzioni dell'equazione si sceglie quella di minimo modulo, per cui $|\phi| \leq \frac{\pi}{4}$, calcolata nella forma

$$t = \frac{\text{sgn}(m)}{|m| + \sqrt{1 + m^2}}$$

per evitare possibili fenomeni di cancellazione. Si calcola poi

$$c = \frac{1}{\sqrt{1 + t^2}} \quad \text{e} \quad s = tc.$$

Con questa trasformazione si annulla quindi un elemento non principale della matrice, che in generale può essere modificato al passo successivo. Lo scopo del metodo è quello di ridurre ad ogni passo la quantità

$$S(A^{(k)}) = \sum_{\substack{r,j=1 \\ r \neq j}}^{n} |a_{rj}^{(k)}|^2.$$

Tenendo conto delle relazioni, riportate nel paragrafo 5, che legano gli elementi di $A^{(k)}$ e di $A^{(k+1)}$ e del fatto che $c^2 + s^2 = 1$, si ricava che

$$|\hat{a}_{pp}|^2 + |\hat{a}_{qq}|^2 + 2|\hat{a}_{pq}|^2 = |a_{pp}|^2 + |a_{qq}|^2 + 2|a_{pq}|^2$$

e

$$|\hat{a}_{rp}|^2 + |\hat{a}_{rq}|^2 = |a_{rp}|^2 + |a_{rq}|^2, \text{ per } r \neq p,\ q.$$

Poiché gli altri elementi delle due matrici $A^{(k)}$ e $A^{(k+1)}$ non cambiano, si ha

$$\sum_{r,j=1}^{n} |\hat{a}_{rj}|^2 = \sum_{r,j=1}^{n} |a_{rj}|^2$$

e quindi, avendo imposto la condizione che $\hat{a}_{pq} = 0$, se $a_{pq} \neq 0$ si ha

$$\begin{aligned} S(A^{(k+1)}) &= \sum_{r,j=1}^{n} |\hat{a}_{rj}|^2 - \sum_{r=1}^{n} |\hat{a}_{rr}|^2 \\ &= \sum_{r,j=1}^{n} |a_{rj}|^2 - \sum_{r=1}^{n} |a_{rr}|^2 - 2|a_{pq}|^2 = S(A^{(k)}) - 2|a_{pq}|^2 \\ &< S(A^{(k)}). \end{aligned} \tag{42}$$

La successione dei numeri positivi $\{S(A^{(k)})\}$ risulta allora decrescente. Si può dimostrare che questa successione tende a zero solo individuando ad ogni passo un'opportuna strategia per la scelta degli elementi da azzerare. Nella *strategia classica* al k-esimo passo si sceglie un elemento non principale di massimo modulo di $A^{(k)}$. Il procedimento si arresta quando tale modulo risulta inferiore ad una quantità prefissata che dipende dalla precisione che si vuole ottenere.

6.32 Teorema. *Sia* $\{A^{(k)}\}$ *la successione ottenuta applicando il metodo di Jacobi alla matrice hermitiana* $A \in \mathbf{C}^{n\times n}$ *secondo la strategia classica. Allora* $\lim_{k\to\infty} S(A^{(k)}) = 0$ *e quindi il* $\lim_{k\to\infty} A^{(k)}$ *è una matrice diagonale.*

Dim. Poiché a_{pq} è un elemento non principale di massimo modulo di $A^{(k)}$, risulta

$$a_{pq}^2 \geq \frac{S(A^{(k)})}{n(n-1)} \ .$$

Dalla (42) si ha allora

$$S(A^{(k+1)}) \leq S(A^{(k)}) - \frac{2S(A^{(k)})}{n(n-1)} = \gamma S(A^{(k)}) \tag{43}$$

dove $\gamma = 1 - \dfrac{2}{n(n-1)} < 1$ per $n \geq 2$. Applicando in modo ricorrente la (43) si ha

$$S(A^{(k+1)}) \leq \gamma^k S(A^{(1)}),$$

da cui la tesi. ∎

Per individuare nella matrice $A^{(k)}$ un elemento non principale di massimo modulo si devono confrontare fra di loro $\dfrac{n(n-1)}{2}$ elementi. Perciò la strategia classica ha un costo computazionale elevato, e per questo motivo è preferibile adottare una *strategia ciclica* in cui la scelta della successione degli indici (p, q) avviene nel modo seguente

$$\begin{array}{ccccc} (1,2) & (1,3) & \dots & (1,n) \\ & (2,3) & \dots & (2,n) \\ & & \ddots & \vdots \\ & & & (n-1,n) \end{array}$$

e tale successione viene ripetuta ciclicamente saltando gli indici (p, q) corrispondenti a elementi che in modulo sono minori di una quantità prefissata.

Anche per il metodo di Jacobi applicato con la strategia ciclica si può dimostrare un teorema di convergenza analogo al 6.32. Inoltre sia per la strategia classica che per quella ciclica è possibile dimostrare [27] che se A è hermitiana con autovalori distinti $\lambda_i,\ i = 1, 2, \ldots, n$, allora da un certo passo k in poi risulta

$$S(A^{(k+N)}) \leq \frac{2S(A^{(k)})^2}{\delta^2} \ ,$$

dove $N = \dfrac{n(n-1)}{2}$ e $\delta = \min\limits_{i \neq j} |\lambda_i - \lambda_j|$, cioè il metodo di Jacobi ha convergenza localmente quadratica.

6.33 Esempio. Applicando il metodo di Jacobi con la strategia classica alla matrice

$$A = \begin{bmatrix} 4 & 3 & 2 & 1 \\ 3 & 4 & 3 & 2 \\ 2 & 3 & 4 & 3 \\ 1 & 2 & 3 & 4 \end{bmatrix} .$$

dell'esempio 6.17, si ha

k	p	q	$S(A^{(k)})$
1	1	2	72.00000
2	2	3	53.99997
3	2	4	28.99997
4	3	4	5.540541
5	1	4	2.681776
6	1	2	1.627234
7	2	4	1.058018
8	2	3	0.5056292
9	1	3	$0.5161846\ 10^{-1}$
10	3	4	$0.2429459\ 10^{-2}$
11	1	4	$0.1281016\ 10^{-3}$
12	1	2	$0.5892318\ 10^{-4}$
13	2	3	$0.2793697\ 10^{-4}$
14	2	4	$0.1687663\ 10^{-5}$
15	1	3	$0.1030698\ 10^{-7}$

Gli elementi principali della matrice $A^{(15)}$ sono

$$0.9009814 \quad 11.09902 \quad 0.5857867 \quad 3.414210,$$

che si assumono come approssimazioni degli autovalori della matrice A. Con la strategia ciclica, per ottenere le stesse approssimazioni degli autovalori, sono richiesti due passi in più. ∎

Posto $Q_k = T_k T_{k-1} \dots T_1$, la matrice

$$Q = \lim_{k\to\infty} Q_k$$

ha per colonne gli autovettori di A.

10. Metodo delle potenze

Il *metodo delle potenze* è un classico metodo iterativo per approssimare l'autovalore di modulo massimo di una matrice e il corrispondente autovettore. Sulla base di questo metodo sono stati sviluppati altri metodi che sono particolarmente adatti per approssimare gli autovalori di matrici sparse di grosse dimensioni. È facile dimostrare la convergenza del metodo nel caso che la matrice sia diagonalizzabile e abbia un solo autovalore di modulo massimo.

Sia $A \in \mathbf{C}^{n\times n}$, con n autovettori $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ linearmente indipendenti e autovalori $\lambda_1, \lambda_2, \dots, \lambda_n$ tali che

$$|\lambda_1| > |\lambda_2| \geq \dots \geq |\lambda_n|,$$

cioè l'autovalore di modulo massimo ha molteplicità algebrica 1 e non esistono altri autovalori con lo stesso modulo.

Fissato un vettore $\mathbf{t}_0 \in \mathbf{C}^n$, si genera la successione $\{\mathbf{y}_k\},\ k = 1, 2, \dots$, così definita

$$\begin{aligned} &\mathbf{y}_0 = \mathbf{t}_0, \\ &\mathbf{y}_k = A\mathbf{y}_{k-1}, \quad k = 1, 2, \dots \end{aligned} \tag{44}$$

Poiché i vettori $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ sono linearmente indipendenti, il vettore $\mathbf{t}_0$ può essere espresso per mezzo della combinazione lineare

$$\mathbf{t}_0 = \sum_{i=1}^{n} \alpha_i \mathbf{x}_i,$$

e si supponga scelto in modo tale che $\alpha_1 \neq 0$; risulta quindi

$$\mathbf{y}_k = A^k \mathbf{t}_0 = \sum_{i=1}^{n} \alpha_i A^k \mathbf{x}_i = \sum_{i=1}^{n} \alpha_i \lambda_i^k \mathbf{x}_i = \lambda_1^k \left[\alpha_1 \mathbf{x}_1 + \sum_{i=2}^{n} \alpha_i \Big(\frac{\lambda_i}{\lambda_1}\Big)^k \mathbf{x}_i \right]. \tag{45}$$

Indicate con $y_r^{(k)}$ e con $x_r^{(i)}$ le r-esime componenti dei vettori $\mathbf{y}_k$ e $\mathbf{x}_i$, per gli indici j per cui $y_j^{(k)} \neq 0$ e $x_j^{(1)} \neq 0$, si ha

$$\frac{y_j^{(k+1)}}{y_j^{(k)}} = \lambda_1 \frac{\alpha_1 x_j^{(1)} + \sum_{i=2}^{n} \alpha_i \Big(\frac{\lambda_i}{\lambda_1}\Big)^{k+1} x_j^{(i)}}{\alpha_1 x_j^{(1)} + \sum_{i=2}^{n} \alpha_i \Big(\frac{\lambda_i}{\lambda_1}\Big)^{k} x_j^{(i)}}\ , \tag{46}$$

e poiché $|\lambda_i/\lambda_1| < 1$ per $i \geq 2$ si ha

$$\lim_{k\to\infty} \frac{y_j^{(k+1)}}{y_j^{(k)}} = \lambda_1.$$

Quindi da un certo indice k in poi l'autovalore λ_1 può essere approssimato mediante uno dei rapporti $y_j^{(k+1)}/y_j^{(k)}$.

Con questo metodo si può approssimare anche l'autovettore $\mathbf{x}_1$. Dalla (45) risulta infatti

$$\lim_{k\to\infty} \frac{\mathbf{y}_k}{\lambda_1^k} = \alpha_1 \mathbf{x}_1,$$

e quindi per $j = 1, \ldots, n$, è

$$\lim_{k\to\infty} \frac{y_j^{(k)}}{\lambda_1^k} = \alpha_1 \mathbf{x}_j^{(1)},$$

e

$$\lim_{k\to\infty} \frac{\mathbf{y}_k}{y_j^{(k)}} = \frac{\mathbf{x}_1}{x_j^{(1)}}\ , \tag{47}$$

per tutti gli indici j per cui $x_j^{(1)} \neq 0$. Poiché per k sufficientemente elevato l'indice m di una componente di massimo modulo di $\mathbf{y}_k$ rimane costante, la successione $\mathbf{y}_k/y_m^{(k)}$ converge all'autovettore $\mathbf{x}_1$ normalizzato in norma ∞.

Questo metodo richiede ad ogni passo il calcolo del prodotto di una matrice A per un vettore: se A non è sparsa ogni passo richiede n^2 operazioni moltiplicative, mentre se A è sparsa ogni passo richiede θ operazioni moltiplicative, dove $\theta << n^2$ è il numero di elementi non nulli di A (ad esempio se A è tridiagonale, il numero degli elementi non nulli di A è $3n-2$).

Però operando in aritmetica finita, con la (44) dopo pochi passi si possono presentare condizioni di overflow o di underflow. Per evitare che ciò accada è necessario eseguire ad ogni passo una normalizzazione del vettore ottenuto, costruendo una successione $\mathbf{t}_k,\ k = 1, 2, \ldots$ così definita

$$\left.\begin{aligned} \mathbf{u}_k &= A\mathbf{t}_{k-1}, \\ \mathbf{t}_k &= \frac{1}{\beta_k}\ \mathbf{u}_k, \end{aligned}\right\}, \quad k = 1, 2, \ldots, \tag{48}$$

dove β_k è uno scalare tale che $\|\mathbf{t}_k\| = 1$, per qualche norma vettoriale $\| \, . \, \|$.

Si ha allora

$$\mathbf{t}_k = \frac{1}{\gamma_k} \, \mathbf{y}_k = \frac{1}{\gamma_k} A^k \mathbf{t}_0, \quad \text{dove } \gamma_k = \prod_{i=1}^{k} \beta_i,$$

e poiché

$$\mathbf{u}_{k+1} = \frac{1}{\gamma_k} \, A^{k+1} \mathbf{t}_0,$$

operando come nella (46) si ha che il rapporto fra le j-esime componenti di $\mathbf{u}_{k+1}$ e $\mathbf{t}_k$, per gli indici j per cui $t_j^{(k)} \neq 0$ e $x_j^{(1)} \neq 0$, è dato da

$$\frac{u_j^{(k+1)}}{t_j^{(k)}} = \lambda_1 \frac{\alpha_1 x_j^{(1)} + \sum_{i=2}^{n} \alpha_i \Big(\frac{\lambda_i}{\lambda_1}\Big)^{k+1} x_j^{(i)}}{\alpha_1 x_j^{(1)} + \sum_{i=2}^{n} \alpha_i \Big(\frac{\lambda_i}{\lambda_1}\Big)^{k} x_j^{(i)}}, \tag{49}$$

e quindi

$$\lim_{k \to \infty} \frac{u_j^{(k+1)}}{t_j^{(k)}} = \lambda_1.$$

Si esaminano ora in dettaglio i casi particolari in cui la normalizzazione sia fatta con la norma ∞ o con la norma 2.

Utilizzando la norma ∞, sia $\|\mathbf{t}_0\|_\infty = 1$ e sia β_k una componente di massimo modulo di $\mathbf{u}_k$, cioè tale che

$$\beta_k = u_m^{(k)}, \quad \text{con} \quad |u_m^{(k)}| = \max_{j=1,\ldots,n} |u_j^{(k)}| = \|\mathbf{u}_k\|_\infty.$$

I vettori $\mathbf{t}_k$ ottenuti con la (48) sono quindi tali che $t_m^{(k)} = 1$. Dalla (49) risulta

$$u_m^{(k+1)} = \lambda_1 \left(1 + O\Big(\frac{\lambda_2}{\lambda_1}\Big)^k\right).$$

Poiché si può assumere che da una certa iterazione in poi l'indice m, corrispondente a una componente di massimo modulo di $\mathbf{u}_k$, resti sempre lo stesso, ne segue che la successione dei β_k converge a λ_1 e che l'errore che si commette approssimando λ_1 con β_k tende a zero come $|\lambda_2/\lambda_1|^k$. Inoltre, poiché $\|\mathbf{t}_k\|_\infty = 1$, dalla (47) risulta

$$\lim_{k \to \infty} \mathbf{t}_k = \frac{\mathbf{x}_1}{x_m^{(1)}},$$

e quindi la successione $\mathbf{t}_k$ converge all'autovettore $\mathbf{x}_1$ normalizzato in norma ∞.

Fissata una tolleranza ϵ, come condizione di arresto del metodo iterativo si può utilizzare una delle condizioni seguenti:

$$|\beta_{k+1} - \beta_k| < \epsilon, \tag{50}$$

o

$$\left|\frac{\beta_{k+1} - \beta_k}{\beta_{k+1}}\right| < \epsilon.$$

6.34 Esempio. Si consideri la matrice

$$A = \begin{bmatrix} 15 & -2 & 2 \\ 1 & 10 & -3 \\ -2 & 1 & 0 \end{bmatrix}$$

che, come si è visto nell'esempio 2.36, ha due autovalori λ_1 e λ_2 in

$$\{\ z \in \mathbf{C} : |z - 15| \leq 3\ \} \cup \{\ z \in \mathbf{C} : |z - 10| \leq 3\ \},$$

e un autovalore λ_3 in

$$\{\ z \in \mathbf{C} : |z| \leq 3\ \}.$$

Il metodo delle potenze, applicato ad A normalizzando rispetto alla norma ∞, a partire dal vettore $\mathbf{t}_0 = [1, 1, 1]^T$, fornisce le seguenti successioni di valori che approssimano l'autovalore λ_1 e l'autovettore corrispondente:

k	β_k	$\mathbf{t}_k^T$		
1	15.00000	1.000000	0.5333333	-0.06666667
2	13.80000	1.000000	0.4734299	-0.1062801
3	13.84058	1.000000	0.4373472	-0.1102966
4	13.90471	1.000000	0.4102466	-0.1123829
⋮	⋮	⋮	⋮	⋮
41	14.10255	1.000000	0.3303270	-0.1183949

Con il criterio di arresto (50) e la tolleranza $\epsilon = 10^{-6}$, il metodo si arresta al 41-esimo passo fornendo i valori approssimati

$$\lambda_1 = 14.10255$$

e

$$\mathbf{x}_1 = [1.000000, 0.3303270, -0.1183949]^T. \qquad \blacksquare$$

Utilizzando la norma 2, sia $\|\mathbf{t}_0\|_2 = 1$ e sia $\beta_k = \|\mathbf{u}_k\|_2$. Questa scelta di β_k è particolarmente conveniente nel caso che la matrice A sia normale, perché si ottiene una successione che converge a λ_1 più velocemente che nel caso precedente. Infatti, tenendo conto che gli autovettori $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ di una matrice normale A possono essere scelti ortonormali, risulta che

$$\begin{aligned}\sigma_k = \mathbf{t}_k^H \mathbf{u}_{k+1} = \frac{\mathbf{t}_k^H A \mathbf{t}_k}{\mathbf{t}_k^H \mathbf{t}_k} &= \frac{(A^k\mathbf{t}_0)^H (A^{k+1}\mathbf{t}_0)}{(A^k\mathbf{t}_0)^H (A^k\mathbf{t}_0)} \\ &= \lambda_1 \, \frac{|\alpha_1|^2 + \sum_{i=2}^{n} |\alpha_i|^2 \left|\frac{\lambda_i}{\lambda_1}\right|^{2k} \left(\frac{\lambda_i}{\lambda_1}\right)}{|\alpha_1|^2 + \sum_{i=2}^{n} |\alpha_i|^2 \left|\frac{\lambda_i}{\lambda_1}\right|^{2k}} \\ &= \lambda_1 \left[1 + O\Big(\Big|\frac{\lambda_2}{\lambda_1}\Big|^{2k}\Big)\right].\end{aligned}$$

La successione dei σ_k converge a λ_1 e l'errore che si commette approssimando λ_1 con σ_k tende a zero con $|\lambda_2/\lambda_1|^{2k}$. Quindi la successione dei σ_k converge più rapidamente della successione dei β_k.

In questo caso, invece della (50), poiché la matrice A è normale, si può utilizzare come criterio di arresto la condizione

$$\|\mathbf{u}_{k+1} - \sigma_k \mathbf{t}_k\|_2 < \epsilon, \tag{51}$$

che oltre ad essere facilmente applicabile, fornisce una maggiorazione dell'errore assoluto: infatti per la (2) risulta che esiste un autovalore λ di A tale che

$$|\lambda - \sigma_k| \leq \frac{\|(A - \sigma_k I)\mathbf{t}_k\|_2}{\|\mathbf{t}_k\|_2} = \|\mathbf{u}_{k+1} - \sigma_k \mathbf{t}_k\|_2 < \epsilon.$$

In modo analogo, se la matrice A non è singolare, una condizione di arresto per l'errore relativo è data da

$$\frac{\|\mathbf{u}_{k+1} - \sigma_k \mathbf{t}_k\|_2}{\|\mathbf{u}_{k+1}\|_2} < \epsilon,$$

infatti per la (3) risulta che esiste un autovalore λ di A tale che

$$\Big|\frac{\lambda_1 - \sigma_k}{\lambda_1}\Big| \leq \frac{\|(A - \sigma_k I)A^{-1}\mathbf{u}_{k+1}\|_2}{\|\mathbf{u}_{k+1}\|_2} = \frac{\|\mathbf{u}_{k+1} - \sigma_k \mathbf{t}_k\|_2}{\|\mathbf{u}_{k+1}\|_2} < \epsilon.$$

Si noti che con la normalizzazione in norma 2 la successione $\{\mathbf{t}_k\}$ può non avere limite, ma per la (47) ogni successione $\{\dfrac{\mathbf{t}_k}{t_j^{(k)}},\ t_j^{(k)} \neq 0\}$ ha per limite l'autovettore $\mathbf{x}_1$ opportunamente normalizzato.

6.35 Esempio. Alla matrice

$$A = \begin{bmatrix} 4 & 3 & 2 & 1 \\ 3 & 4 & 3 & 2 \\ 2 & 3 & 4 & 3 \\ 1 & 2 & 3 & 4 \end{bmatrix}$$

dell'esempio 6.17 si applica il metodo delle potenze con la normalizzazione di $\mathbf{t}_k$ rispetto alla norma ∞ a partire dal vettore $\mathbf{t}_0 = [1, 1, 1, 1]^T$ e rispetto alla norma 2, a partire dal vettore $\mathbf{t}_0 = [0.5, 0.5, 0.5, 0.5]^T$. Nella figura 6.2 sono riportati gli errori $|\beta_k - \lambda_1|$ (indicati con i quadratini vuoti) e $|\sigma_k - \lambda_1|$ (indicati con i quadratini pieni).

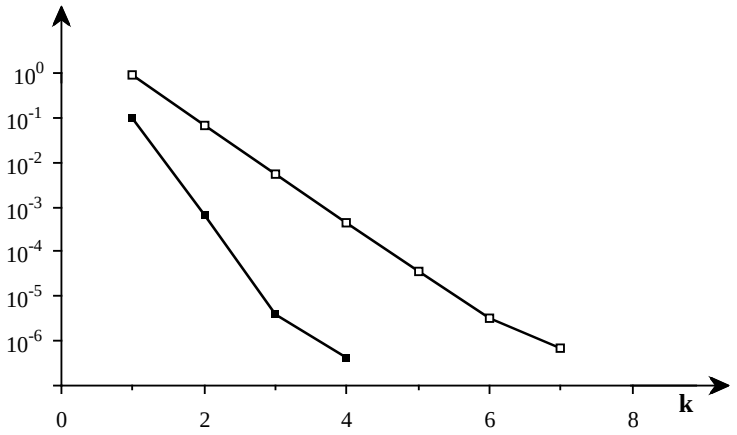


Fig. 6.2 - Errori del metodo delle potenze con la normalizzazione rispetto alla norma ∞ e alla norma 2.

Fissata la tolleranza $\epsilon = 10^{-6}$, il metodo si arresta alla settima iterazione quando la normalizzazione viene fatta rispetto alla norma ∞ e si usa il criterio (50) e alla quarta iterazione quando la normalizzazione viene fatta rispetto alla norma 2 e si usa il criterio (51). Si noti la maggiore velocità di convergenza della successione dei σ_k. ∎

Il metodo delle potenze è convergente anche nel caso in cui l'autovalore di modulo massimo abbia molteplicità algebrica maggiore di 1, cioè $\lambda_1 = \lambda_2 = \ldots = \lambda_r$, con

$$|\lambda_1| = |\lambda_2| = \ldots = |\lambda_r| > |\lambda_{r+1}| \geq \ldots \geq |\lambda_n|.$$

Infatti al posto della (45) si ha

$$\mathbf{y}_k = \lambda_1^k \Big[\sum_{i=1}^{r} \alpha_i \mathbf{x}_i + \sum_{i=r+1}^{n} \alpha_i \Big(\frac{\lambda_i}{\lambda_1} \Big)^k \mathbf{x}_i \Big];$$

l'autovalore λ_1 si approssima con la successione dei β_k o dei σ_k, e l'errore dell'approssimazione tende a zero come $(\lambda_{r+1}/\lambda_1)^k$ o come $|\lambda_{r+1}/\lambda_1|^{2k}$. Inoltre

$$\lim_{k\to\infty} \frac{\mathbf{y}_k}{y_j^{(k)}} = \frac{1}{\theta_j} \sum_{i=1}^{r} \alpha_i \mathbf{x}_i, \quad \text{dove} \quad \theta_j = \sum_{i=1}^{r} \alpha_i x_j^{(i)},$$

e quindi la successione $\{\mathbf{y}_k/y_m^{(k)}\}$, dove m è l'indice di una componente di massimo modulo di $\mathbf{y}_k$, converge ad un autovettore normalizzato in norma ∞ appartenente allo spazio vettoriale generato da $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_r$.

Se invece esistono più autovalori di modulo massimo diversi fra loro, il metodo delle potenze non è convergente (si veda l'esercizio 6.35).

6.36 Esempio. La matrice

$$A = \begin{bmatrix} 8 & -1 & -5 \\ -4 & 4 & -2 \\ 18 & -5 & -7 \end{bmatrix}$$

ha gli autovalori $2 \pm 4\,\mathbf{i}$ e 1. Gli autovalori di modulo massimo sono quelli complessi e quindi sono distinti. Con il metodo delle potenze, applicato a partire dal vettore $\mathbf{t}_0 = [1, 1, 1]^T$, normalizzando $\mathbf{t}_k$ rispetto alla norma ∞, si ottiene la successione:

k	β_k
1	6.000000
2	-4.666667
3	3.714286
4	4.769224
5	-5.096744
⋮	⋮
44	-3.377194
45	5.844138
⋮	⋮

che risulta non convergente. ∎

Il metodo delle potenze può essere modificato in modo da approssimare anche autovalori distinti con lo stesso modulo, come nel caso di autovalori complessi coniugati [28] (si veda anche l'esercizio 6.34).

Come risulta dalle considerazioni precedenti, la condizione $\alpha_1 \neq 0$ è, in teoria, necessaria per la convergenza a λ_1 della successione (46). Se infatti

fosse $\alpha_1 = 0$, $\alpha_2 \neq 0$ e $|\lambda_2| > |\lambda_3|$, allora è possibile dimostrare con argomentazioni analoghe che la successione $\{y_j^{(k+1)}/y_j^{(k)}\}$ tende all'autovalore λ_2. In pratica però, anche se $\mathbf{t}_0$ fosse tale che $\alpha_1 = 0$, per la presenza degli errori di arrotondamento, i vettori $\mathbf{t}_k$ effettivamente calcolati sarebbero comunque rappresentabili come combinazioni lineari degli autovettori con una componente non nulla rispetto a $\mathbf{x}_1$. Perciò la successione effettivamente calcolata convergerebbe ugualmente a λ_1. Inoltre nel caso che le componenti del vettore $\mathbf{t}_0$ vengano scelte casualmente nell'insieme dei numeri complessi, la probabilità di ottenere un vettore per cui $\alpha_1 = 0$ è nulla.

11. Varianti del metodo delle potenze

Varianti del metodo delle potenze consentono di calcolare anche gli altri autovalori e i corrispondenti autovettori.

a) Variante di Wielandt (metodo delle potenze inverse)

Se A è una matrice non singolare, diagonalizzabile, con autovalori $\lambda_i,\ i = 1, \ldots, n$, tali che

$$|\lambda_1| \geq \ldots \geq |\lambda_{n-1}| > |\lambda_n| > 0,$$

la matrice A^{-1} ha autovalori $\dfrac{1}{\lambda_i},\ i = 1, \ldots, n$, tali che

$$\frac{1}{|\lambda_n|} > \frac{1}{|\lambda_{n-1}|} \geq \ldots \geq \frac{1}{|\lambda_1|}.$$

Per calcolare l'autovalore di modulo minimo di A si applica il metodo delle potenze alla matrice A^{-1}, nel modo seguente

$$\left.\begin{aligned} A\mathbf{u}_k &= \mathbf{t}_{k-1}, \\ \mathbf{t}_k &= \frac{1}{\beta_k}\ \mathbf{u}_k, \end{aligned}\right\}, \quad k = 1, 2, \ldots, \tag{52}$$

dove β_k è uno scalare tale che $\|\mathbf{t}_k\| = 1$, per la norma scelta. Ogni passo del metodo richiede la risoluzione del sistema lineare $A\mathbf{u}_k = \mathbf{t}_{k-1}$. Per $k \to \infty$ la successione dei β_k, se si usa la $\| \ . \ \|_\infty$, o dei σ_k, se si usa la $\| \ . \ \|_2$ e la matrice A è normale, tende a $\dfrac{1}{\lambda_n}$ e $\mathbf{t}_k$ tende al corrispondente autovettore della matrice A^{-1} (e quindi di A).

Se di un autovalore λ_j è nota una stima μ, tale che

$$0 < |\mu - \lambda_j| < |\mu - \lambda_i|, \quad j \neq i,$$

questo autovalore può essere calcolato, applicando il metodo delle potenze alla matrice $(A-\mu I)^{-1}$, nel modo seguente

$$\left.\begin{array}{l}(A-\mu I)\mathbf{u}_k = \mathbf{t}_{k-1},\\ \mathbf{t}_k = \dfrac{1}{\beta_k}\ \mathbf{u}_k,\end{array}\right\},\quad k=1,2,\ldots, \tag{53}$$

dove β_k è uno scalare tale che $\|\mathbf{t}_k\| = 1$, per la norma scelta. Per $k \to \infty$ la successione dei β_k o dei σ_k tende a $\dfrac{1}{\lambda_j - \mu}$ e $\mathbf{t}_k$ tende al corrispondente autovettore della matrice $(A-\mu I)^{-1}$ (e quindi di A).

L'autovalore λ_j viene calcolato a meno di un errore che tende a zero con

$$\Big(\frac{|\lambda_j - \mu|}{\min\{\ |\lambda_{j-1}-\mu|\ ,\ |\lambda_{j+1}-\mu|\ \}}\Big)^k .$$

Questo metodo è spesso usato per migliorare l'approssimazione di un autovalore ottenuta con altri metodi. Va però rilevato che più μ è vicino a λ_j più rapida è la convergenza del metodo, ma aumentano le difficoltà numeriche nel calcolo di $\mathbf{u}_k$ perché la matrice $A-\mu I$ tende a diventare mal condizionata.

Per il calcolo effettivo della (52) o (53), conviene prima fattorizzare, con un costo computazionale di $n^3/3$ operazioni moltiplicative, la matrice A o la matrice $A-\mu I$ nella forma LU. Poi ad ogni passo si risolvono due sistemi con matrice dei coefficienti triangolare e il costo computazionale di questo metodo è quindi confrontabile ad ogni passo con quello del metodo delle potenze.

6.37 Esempio. Fissata la tolleranza $\epsilon = 10^{-6}$, si applica il metodo di Wielandt, normalizzando $\mathbf{t}_k$ rispetto alla norma ∞, alla matrice

$$A = \begin{bmatrix} 15 & -2 & 2 \\ 1 & 10 & -3 \\ -2 & 1 & 0 \end{bmatrix}$$

dell'esempio 6.34, ponendo $\mu = 14$ e $\mathbf{t}_0 = [1,1,1]^T$. Si ottengono le successioni:

k	β_k	$\mathbf{t}_k^T$		
1	9.399977	1.000000	0.3191492	-0.1276596
2	9.782953	1.000000	0.3305786	-0.1183123
3	9.749693	1.000000	0.3303205	-0.1183960
4	9.750801	1.000000	0.3303275	-0.1183950
5	9.750768	1.000000	0.3303272	-0.1183950
6	9.750769	1.000000	0.3303273	-0.1183950

cioè $\dfrac{1}{\lambda_1 - \mu} = 9.750769$, da cui si ricava $\lambda_1 = 14.10256$.

Con il metodo di Wielandt il risultato viene raggiunto con 6 passi, mentre con il metodo delle potenze (si veda l'esempio 6.34) occorrono 41 passi.

Ponendo $\mu = 13$ oppure $\mu = 15$ e partendo dallo stesso vettore $\mathbf{t}_0$, si ottengono ancora successioni convergenti, ma il numero di iterazioni richieste per ottenere la stessa precisione è maggiore (rispettivamente 17 e 9 iterazioni). Ponendo invece $\mu = 12$ e partendo dallo stesso vettore $\mathbf{t}_0$, la successione dei β_k converge in 50 iterazioni a -0.6193352, da cui si ricava l'autovalore $\lambda_2 = 10.38537$.

Per approssimare l'autovalore λ_3 della matrice A, si pone $\mu = 0$ (in questo caso il metodo di Wielandt coincide con il metodo delle potenze applicato alla matrice A^{-1}) e si ottiene per β_k la successione:

k	β_k
1	2.160000
2	1.959506
3	1.953039
4	1.952810
5	1.952802
6	1.952801

cioè $\dfrac{1}{\lambda_3} = 1.952801$, da cui si ricava l'approssimazione $\lambda_3 = 0.5120849$. ∎

Anche per il metodo di Wielandt valgono le stesse considerazioni fatte per il metodo delle potenze. In particolare, se μ si trova alla stessa distanza da due autovalori distinti, allora il metodo non è convergente, come risulta anche dall'esempio seguente.

6.38 Esempio. La matrice

$$A = \begin{bmatrix} 33 & 16 & 72 \\ -24 & -10 & -57 \\ -8 & -4 & -17 \end{bmatrix}$$

ha gli autovalori $\lambda_1 = 3$, $\lambda_2 = 2$, $\lambda_3 = 1$. Ponendo $\mu = 2.5$, a partire da $\mathbf{t}_0 = [1, 1, 1]^T$, si ottiene la successione:

k	β_k
1	73.99426
2	3.657217
3	0.4363987
4	10.16883
⋮	⋮
97	0.2937573
98	13.61612
99	0.2933033

La successione β_k non è convergente perché il valore scelto per μ è equidistante dai due autovalori λ_1 e λ_2. ∎

b) Metodo delle iterazioni del quoziente di Rayleigh

Questo metodo è una variante del metodo di Wielandt applicato a una matrice hermitiana con la normalizzazione in norma 2. La (53) viene così modificata

$$\left.\begin{aligned} \mu_{k-1} &= \mathbf{t}_{k-1}^H A \mathbf{t}_{k-1}, \\ (A - \mu_{k-1} I)\mathbf{u}_k &= \mathbf{t}_{k-1}, \\ \mathbf{t}_k &= \frac{1}{\|\mathbf{u}_k\|_2}\, \mathbf{u}_k, \end{aligned}\right\}, \quad k = 1, 2, \ldots \tag{54}$$

Si può dimostrare [17], in modo analogo a quanto fatto nel caso del metodo delle potenze, che la successione dei μ_k converge ad un autovalore λ della matrice A e che localmente la convergenza è del terzo ordine (per il caso che la matrice abbia autovalori distinti, si veda l'esercizio 6.30). Però ogni passo del metodo richiede in generale un numero di operazioni moltiplicative dell'ordine di $n^3/6$, perché la matrice del sistema (54) cambia ogni volta ed è hermitiana. Inoltre all'aumentare di k aumenta il numero di condizionamento della matrice $A - \mu_{k-1} I$ e quindi aumentano le difficoltà numeriche del calcolo di $\mathbf{u}_k$.

6.39 Esempio. Si calcola l'autovalore λ_1 della matrice

$$A = \begin{bmatrix} 4 & 3 & 2 & 1 \\ 3 & 4 & 3 & 2 \\ 2 & 3 & 4 & 3 \\ 1 & 2 & 3 & 4 \end{bmatrix}$$

dell'esempio 6.17 con il metodo di Wielandt con la normalizzazione in norma ∞, $\mu = 10$ e $\mathbf{t}_0 = [1, 1, 1, 1]^T$, il metodo di Wielandt con la normalizzazione

in norma 2, $\mu = 10$ e $\mathbf{t}_0 = [0.5, 0.5, 0.5, 0.5]^T$, e il metodo del quoziente di Rayleigh con $\mu_0 = 10$ e $\mathbf{t}_0 = [0.5, 0.5, 0.5, 0.5]^T$. Nella figura 6.3 sono riportati gli errori assoluti della successione β_k (indicati con quadratini vuoti), ottenuta con il metodo di Wielandt con la normalizzazione in norma ∞, gli errori assoluti della successione σ_k (indicati con quadratini pieni), ottenuta con il metodo di Wielandt con la normalizzazione in norma 2 e gli errori assoluti della successione μ_k (indicati con triangolini), ottenuta con il metodo del quoziente di Rayleigh. Si confrontino questi risultati con quelli ottenuti con metodo delle potenze e riportati nella figura 6.2. ∎

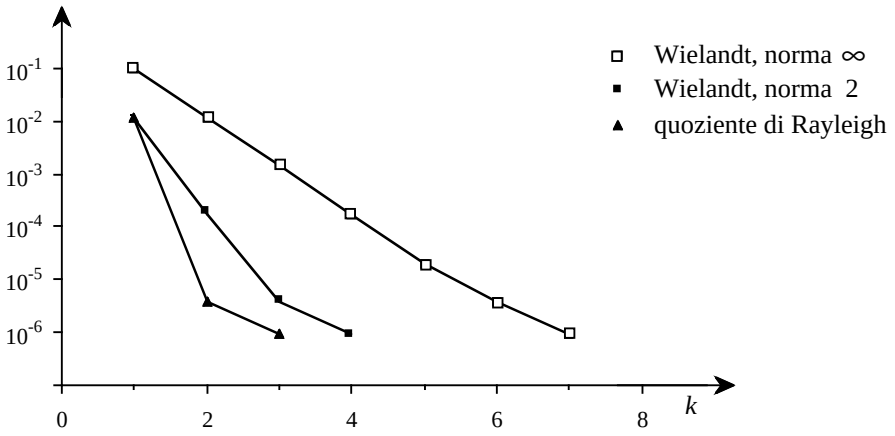


Fig. 6.3 - Errori delle soluzioni ottenute con il metodo di Wielandt con la normalizzazione rispetto alla norma ∞ e alla norma 2 e con il metodo del quoziente di Rayleigh.

c) Variante dell'ortogonalizzazione - 1

Sia A normale e tale che $|\lambda_1| > |\lambda_2| > \ldots > |\lambda_n|$. Dopo aver calcolato λ_1 e $\mathbf{x}_1$, con $\|\mathbf{x}_1\|_2 = 1$, si considera un qualunque vettore $\mathbf{y} \in \mathbf{C}^n$, $\mathbf{y} \neq \mathbf{0}$ e si applica il metodo delle potenze con la normalizzazione rispetto alla norma 2 partendo dal vettore

$$\mathbf{t}_0 = \frac{\mathbf{z}}{\|\mathbf{z}\|_2}\ , \quad \mathbf{z} = \mathbf{y} - (\mathbf{x}_1^H \mathbf{y})\mathbf{x}_1, \tag{55}$$

ortogonale a $\mathbf{x}_1$. Poiché i vettori $\mathbf{t}_k$ generati con il metodo delle potenze sono (in teoria) ortogonali a $\mathbf{x}_1$, il metodo calcola λ_2. In pratica però, per effetto degli errori di arrotondamento, i vettori $\mathbf{t}_k$ effettivamente calcolati hanno una componente diversa da zero lungo la direzione $\mathbf{x}_1$ che si accentua al crescere di k. Quindi per ottenere una successione dei σ_k che non converga nuovamente a λ_1, occorre *riortogonalizzare*, dopo un certo numero di passi, $\mathbf{t}_k$ rispetto a $\mathbf{x}_1$. Cioè ogni m passi, dove m è un intero opportuno, si

sostituisce il vettore $\mathbf{t}_k$ con il vettore

$$\mathbf{t}'_k = \frac{\mathbf{z}}{\|\mathbf{z}\|_2}\ , \quad \mathbf{z} = \mathbf{t}_k - (\mathbf{x}_1^H \mathbf{t}_k)\mathbf{x}_1.$$

In modo analogo, calcolati gli autovalori $\lambda_1, \lambda_2, \ldots, \lambda_j$ e i corrispondenti autovettori $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_j$, tutti di norma 2 unitaria, è possibile calcolare λ_{j+1} scegliendo

$$\mathbf{t}_0 = \frac{\mathbf{z}}{\|\mathbf{z}\|_2}\ , \quad \mathbf{z} = \mathbf{y} - \sum_{i=1}^{j} (\mathbf{x}_i^H \mathbf{y})\mathbf{x}_i,$$

in modo che $\mathbf{t}_0$ risulti ortogonale a $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_j$. Anche in questo caso occorre effettuare ogni m passi il processo di riortogonalizzazione, che richiede $2jn$ operazioni moltiplicative.

d) Variante dell'ortogonalizzazione - 2

Sia A normale e tale che $|\lambda_1| > |\lambda_2| > \ldots > |\lambda_n|$. Si ha

$$A = \sum_{i=1}^{n} \lambda_i \mathbf{x}_i \mathbf{x}_i^H, .$$

dove $\mathbf{x}_1, \ldots, \mathbf{x}_n$ sono autovettori ortonormali. La matrice

$$A_1 = A - \lambda_1 \mathbf{x}_1 \mathbf{x}_1^H$$

ha autovalori $\lambda_2, \lambda_3, \ldots, \lambda_n$, e 0. Quindi calcolati λ_1 e $\mathbf{x}_1$, il metodo delle potenze, applicato ad A_1 approssima λ_2. In generale, calcolati λ_1, $\lambda_2, \ldots, \lambda_j$ e $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_j$, per calcolare λ_{j+1} si applica il metodo delle potenze alla matrice

$$A - \sum_{i=1}^{j} \lambda_i \mathbf{x}_i \mathbf{x}_i^H.$$

Se la matrice A è sparsa, per utilizzare questa proprietà ad ogni passo il metodo delle potenze viene applicato nel modo seguente

$$\mathbf{u}_k = A\mathbf{t}_{k-1} - \sum_{i=1}^{j} \lambda_i (\mathbf{x}_i^H \mathbf{t}_{k-1})\mathbf{x}_i, \quad k = 1, 2, \ldots,$$

con un aumento ad ogni passo di $2jn$ operazioni moltiplicative.

6.40 Esempio. Fissata una tolleranza $\epsilon = 10^{-6}$ si applica il metodo delle potenze alla matrice dell'esempio 6.17

$$A = \begin{bmatrix} 4 & 3 & 2 & 1 \\ 3 & 4 & 3 & 2 \\ 2 & 3 & 4 & 3 \\ 1 & 2 & 3 & 4 \end{bmatrix}$$

per approssimare tutti gli autovalori (si veda per confronto il calcolo con il metodo QR nell'esempio 6.30). Calcolato il primo autovalore $\lambda_1 = 11.09901$ e il corrispondente autovettore $\mathbf{x}_1$, e posto $\mathbf{y} = [0.5, 0.5, 0.5, 0.5]^T$, si calcola il vettore $\mathbf{t}_0$, ortogonale a $\mathbf{x}_1$, con la (55). La successione dei σ_k che si ottiene è la seguente

k	σ_k
1	0.7694202
2	2.590006
3	3.351716
4	3.410331
⋮	⋮
10	3.414937
11	3.421915
12	3.494808
13	4.188173
14	7.579575
15	10.53008
16	11.04131

Si noti come dopo la decima iterazione per effetto di una progressiva perdita di ortogonalità di $\mathbf{t}_k$ rispetto a $\mathbf{x}_1$, la successione dei σ_k tenda nuovamente a λ_1. Se invece si riortogonalizza ogni 5 passi $\mathbf{t}_k$ rispetto a $\mathbf{x}_1$, si ottiene la successione

k	σ_k
⋮	⋮
4	3.410331
5	3.413966
6	3.414192
7	3.414209
8	3.414209

che converge a $\lambda_2 = 3.414209$. Il calcolo dei successivi autovalori diventa sempre più complicato: riortogonalizzando ogni 5 passi si determina λ_3 solo dopo 18 iterazioni, mentre non si riesce a determinare λ_4. Solamente riortogonalizzando ogni 2 passi si riesce a calcolare $\lambda_4 = 0.5857863$ in 7 iterazioni.

Con la seconda variante, applicando il metodo delle potenze alle matrici

$$A_1 = A - \lambda_1 \mathbf{x}_1 \mathbf{x}_1^T, \quad A_2 = A_1 - \lambda_2 \mathbf{x}_2 \mathbf{x}_2^T, \quad A_3 = A_2 - \lambda_3 \mathbf{x}_3 \mathbf{x}_3^T,$$

si ottengono risultati migliori: il numero di passi richiesti risulta infatti di 9 per λ_2, 4 per λ_3 e 7 per λ_4. ∎

e) Variante della deflazione

Sia $|\lambda_1| > |\lambda_2|$. Calcolati λ_1 e $\mathbf{x}_1$, di norma 2 unitaria, si considera la matrice di Householder P tale che $P\mathbf{x}_1 = \mathbf{e}_1$; risulta

$$PAP^H = \begin{bmatrix} \lambda_1 & \mathbf{0}^H \\ \mathbf{0} & A_1 \end{bmatrix},$$

se A è hermitiana o

$$PAP^H = \begin{bmatrix} \lambda_1 & \mathbf{a}^H \\ \mathbf{0} & A_1 \end{bmatrix},$$

se A non lo è.

Si applica il metodo delle potenze alla matrice A_1 di ordine $n-1$ e si calcolano λ_2 e il corrispondente autovettore $\mathbf{y}_2$ di A_1. L'autovettore $\mathbf{x}_2$ di A corrispondente a λ_2 è dato da

$$\mathbf{x}_2 = P^H \begin{bmatrix} \theta \\ \mathbf{y}_2 \end{bmatrix}, \quad \text{con} \quad \theta = \begin{cases} 0 & \text{se } A \text{ è hermitiana,} \\ \dfrac{\mathbf{a}^H \mathbf{y}_2}{\lambda_2 - \lambda_1} & \text{se } A \text{ non lo è.} \end{cases}$$

Procedendo in questo modo si costruisce la forma di Schur della matrice A. Poiché la trasformazione $A \to PAP^H$ può distruggere la eventuale struttura e sparsità di A, questo procedimento può non essere indicato per matrici sparse.

6.41 Esempio. Fissata una tolleranza $\epsilon = 10^{-6}$, si applica il metodo delle potenze con la variante della deflazione alla matrice

$$A = \begin{bmatrix} 4 & 3 & 2 & 1 \\ 3 & 4 & 3 & 2 \\ 2 & 3 & 4 & 3 \\ 1 & 2 & 3 & 4 \end{bmatrix}$$

dell'esempio 6.17 (si veda l'esempio 6.40 per la variante dell'ortogonalizzazione). Calcolato il primo autovalore $\lambda_1 = 11.09901$ e il corrispondente autovettore $\mathbf{x}_1$, si ottiene la matrice

$$A_1 = \begin{bmatrix} 0.7225373 & 0.1001084 & \text{-}0.1358454 \\ 0.1001084 & 1.477678 & 1.173688 \\ \text{-}0.1358454 & 1.173688 & 2.700763 \end{bmatrix}.$$

Applicando nuovamente il metodo delle potenze ad A_1, a partire dal vettore

$$\mathbf{t}_0 = \frac{1}{\sqrt{3}} [1, 1, 1]^T,$$

si calcola il secondo autovalore $\lambda_2 = 3.414209$ e il corrispondente autovettore $\mathbf{y}_2$ di A_1 in 9 passi. Si ottiene poi la matrice

$$A_2 = \begin{bmatrix} 0.8919271 & 0.05264682 \\ 0.05264682 & 0.5948396 \end{bmatrix},$$

a cui si riapplica il metodo delle potenze, a partire dal vettore

$$\mathbf{t}_0 = \frac{1}{\sqrt{2}} [1, 1]^T,$$

e occorrono 18 iterazioni per calcolare λ_3. ∎

12. Metodo delle iterazioni ortogonali

Questo metodo, noto anche con il nome di *metodo delle iterazioni di sottospazi*, è un'estensione a blocchi del metodo delle potenze ed è particolarmente conveniente quando la matrice A è sparsa e di grandi dimensioni e sono richiesti solo pochi dei suoi autovalori di maggior modulo. Come il metodo delle potenze, anche questo si basa sul fatto che se $\lambda_1, \lambda_2, \ldots, \lambda_n$ sono autovalori della matrice A, allora $\lambda_1^k, \lambda_2^k, \ldots, \lambda_n^k$ sono autovalori di A^k e che se alcuni di essi sono *dominanti* sugli altri, cioè di modulo maggiore degli altri, questa dominanza diventa sempre più grande per gli autovalori di A^k, al crescere di k.

Il metodo può essere applicato a matrici qualsiasi, anche se qui viene presentato solo il caso delle matrici hermitiane con autovalori di modulo distinto, per le quali è possibile applicare una tecnica di accelerazione che rende il metodo molto efficiente.

Sia $A \in \mathbf{C}^{n \times n}$, hermitiana, siano $\lambda_1, \lambda_2, \ldots, \lambda_n$ i suoi autovalori, tali che

$$|\lambda_1| > |\lambda_2| > \ldots > |\lambda_n|,$$

e siano $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ i corrispondenti autovettori, che si suppongono ortonormali. Sia p un intero tale che $1 \leq p < n$. Gli autovalori $\lambda_1, \lambda_2, \ldots, \lambda_p$ sono detti *autovalori dominanti* e i corrispondenti autovettori sono detti *autovettori dominanti*, mentre il sottospazio S_p generato da $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$ è detto *sottospazio invariante dominante*. Il seguente metodo delle *iterazioni ortogonali* consente di approssimare un tale sottospazio.

Sia $Q_0 \in \mathbf{C}^{n\times p}$, una matrice le cui colonne sono ortonormali, cioè tale che $Q_0^H Q_0 = I_p$. Si considerino le successioni di matrici $Q_k,\ R_k,\ Y_k \in \mathbf{C}^{n\times p}$ per $k = 1, 2, \ldots$, definite nel modo seguente:

a) si calcoli $Y_k = AQ_{k-1}$,

b) si calcolino le prime p colonne della matrice unitaria H_k e la matrice R_k di una fattorizzazione QR della matrice Y_k, cioè

$$Y_k = H_k R_k,$$

e sia Q_k la matrice formata dalle p colonne calcolate di H_k.

La successione dei sottospazi $S_p^{(k)}$ generati dalle colonne delle matrici Q_k tende al sottospazio invariante dominante S_p. Vale infatti il seguente teorema (per la dimostrazione si veda l'esercizio 6.41).

6.42 Teorema. *Sia $U \in \mathbf{C}^{n\times p}$, $p < n$, la matrice le cui colonne sono $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$. Posto*

$$d_k = \|UU^H - Q_k Q_k^H\|_2, \quad k = 1, 2, \ldots,$$

se $d_0 < 1$ risulta

$$d_k \leq \frac{d_0}{\sqrt{1 - d_0^2}} \left|\frac{\lambda_{p+1}}{\lambda_p}\right|^k.$$

Inoltre per gli elementi diagonali della matrice R_k si ha

$$|r_{ii}^{(k)} - \lambda_i| = O\Big(\left|\frac{\lambda_{i+1}}{\lambda_i}\right|^k + \left|\frac{\lambda_i}{\lambda_{i-1}}\right|^k\Big)\ ,\ i = 1, 2, \ldots, p, \tag{56}$$

dove si assume $\dfrac{\lambda_1}{\lambda_0} = 0$. La quantità d_k misura la distanza fra i due sottospazi $S_p^{(k)}$ e S_p. ∎

A differenza del metodo delle potenze, in cui il sottospazio invariante che viene determinato è di dimensione 1, generato cioè da un solo autovettore di A, nel metodo delle iterazioni ortogonali è possibile scegliere la dimensione del sottospazio invariante che si vuole determinare, e quindi il numero degli autovettori dominanti.

6.43 Esempio. Si applica il metodo delle iterazioni ortogonali con $p = 2$ alla matrice simmetrica

$$A = \begin{bmatrix} 4 & 3 & 2 & 1 \\ 3 & 4 & 3 & 2 \\ 2 & 3 & 4 & 3 \\ 1 & 2 & 3 & 4 \end{bmatrix}$$

dell'esempio 6.17 (i cui autovalori $\lambda_1 = 11.09902$, $\lambda_2 = 3.414210$, $\lambda_3 = 0.9009814$, $\lambda_4 = 0.5857867$ sono stati determinati in vari esempi di questo capitolo, fra cui il 6.26). Posto

$$Q_0 = [\mathbf{e}_1 \mid \mathbf{e}_2],$$

si ottiene la seguente successione di matrici:

$$R_1 = \begin{bmatrix} \text{-5.477224} & \text{-5.842374} \\ 0 & \text{-1.966382} \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \qquad R_2 = \begin{bmatrix} 10.17839 & 3.911512 \\ 0 & 3.128832 \\ 0 & 0 \\ 0 & 0 \end{bmatrix},$$

$$R_3 = \begin{bmatrix} 11.00446 & 1.368338 \\ 0 & 3.404015 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \qquad R_4 = \begin{bmatrix} 11.08995 & 0.4257860 \\ 0 & 3.414426 \\ 0 & 0 \\ 0 & 0 \end{bmatrix},$$

$$R_5 = \begin{bmatrix} 11.09813 & 0.1311131 \\ 0 & 3.414291 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \qquad R_6 = \begin{bmatrix} 11.09890 & 0.04034042 \\ 0 & 3.414214 \\ 0 & 0 \\ 0 & 0 \end{bmatrix},$$

$$\dots$$

Le successioni $\{r_{11}^{(k)}\}$ e $\{r_{22}^{(k)}\}$ degli elementi principali delle matrici R_k convergono rispettivamente ai primi due autovalori di A, con velocità di convergenza determinate dai due rapporti

$$\frac{\lambda_2}{\lambda_1} \approx 0.3, \qquad \frac{\lambda_3}{\lambda_2} \approx 0.25.$$

Arrestando il metodo all'ottava iterazione, quando

$$\max_{i=1,2} |r_{ii}^{(k)} - r_{ii}^{(k-1)}| < 10^{-5},$$

si ottengono i valori

$$r_{11}^{(8)} = 11.09898 \quad \text{e} \quad r_{22}^{(8)} = 3.414203,$$

che si assumono come approssimazioni di λ_1 e λ_2. Inoltre si ha

$$Q_8 = \begin{bmatrix} \text{-0.4483709} & 0.6532544 \\ \text{-0.5468667} & 0.2705123 \\ \text{-0.5468036} & \text{-0.2706826} \\ \text{-0.4482216} & \text{-0.6533069} \end{bmatrix},$$

le cui colonne si assumono come approssimazione di $\mathbf{x}_1$ e $\mathbf{x}_2$.

Applicando lo stesso metodo con $p = 2$ alla matrice non simmetrica (non è necessario che la matrice sia hermitiana per poter applicare il metodo) dell'esempio 6.24

$$A = \begin{bmatrix} 4 & 3 & 2 & 1 \\ 1 & 4 & 3 & 2 \\ 1 & 1 & 4 & 3 \\ 1 & 1 & 1 & 4 \end{bmatrix}$$

(i cui autovalori $\lambda_1 = 8.782016$, $\lambda_2 = 2.563376 + 1.152632\ \mathbf{i}$, $\lambda_3 = \overline{\lambda}_2$, $\lambda_4 = 2.089449$ sono stati determinati nell'esempio 6.29) a partire dalla stessa matrice Q_0, si ottengono per gli elementi principali delle matrici R_k le successioni

k	$r_{11}^{(k)}$	$r_{22}^{(k)}$
1	-4.358897	-3.153938
2	6.870981	2.365515
3	8.391135	2.291078
4	8.730016	2.895477
⋮	⋮	⋮
49	8.783383	3.458923
50	8.783383	3.145504

Mentre la successione delle prime componenti converge, la seconda non converge perché λ_2 e λ_3 hanno lo stesso modulo. ∎

La velocità di convergenza, che per la (56) dipende dai rapporti λ_{i+1}/λ_i, è bassa quando due autovalori successivi λ_i e λ_{i+1} hanno moduli che differiscono di poco. È possibile ottenere una convergenza migliore se si calcolano gli autovalori della matrice B_k, restrizione di A al sottospazio $S_p^{(k)}$

$$B_k = V_k^H A V_k \qquad \textit{(quoziente di Rayleigh generalizzato)},$$

dove V_k è una matrice le cui colonne formano una base ortonormale per $S_p^{(k)}$. Si ottiene così il seguente metodo delle iterazioni ortogonali con *accelerazione di Ritz*. Posto $Q_0 \in \mathbf{C}^{n\times p}$, tale che $Q_0^H Q_0 = I_p$, per $k = 1, 2, \ldots$

a) si calcoli $Y_k = AQ_{k-1}$,

b) si calcoli la matrice V_k formata dalle prime p colonne della matrice unitaria H_k della fattorizzazione QR della matrice Y_k,

c) si calcoli $B_k = V_k^H A V_k$,

d) si determini la decomposizione di Schur di B_k:

$$B_k = U_k D_k U_k^H,$$

in cui D_k e $U_k \in \mathbf{C}^{p\times p}$ sono matrici rispettivamente diagonale e unitaria,

e) si calcoli $Q_k = V_k U_k$.

Il passo d) richiede ovviamente la determinazione degli autovalori e degli autovettori normalizzati della matrice hermitiana B_k, di dimensione p.

Indicati con $d_i^{(k)}$ gli elementi principali di D_k, si può dimostrare [22] che

$$|d_i^{(k)} - \lambda_i| = O\Big(\Big|\frac{\lambda_{p+1}}{\lambda_i}\Big|^k\Big),\ i = 1, 2, \ldots, p.$$

e quindi i valori $d_i^{(k)}$ convergono più velocemente dei valori $r_{ii}^{(k)}$.

6.44 Esempio. La matrice $A \in \mathbf{R}^{4\times 4}$:

$$A = \frac{1}{45}\begin{bmatrix} 1427 & -280 & -64 & 1974 \\ -280 & 407 & -1024 & -492 \\ -64 & -1024 & 4241 & -114 \\ 1974 & -492 & -114 & 3060 \end{bmatrix}$$

ha gli autovalori $\lambda_1 = 100,\ \lambda_2 = 99,\ \lambda_3 = 3$ e $\lambda_4 = 1$. Applicando il metodo delle iterazioni ortogonali con $p = 2$ a partire dalla matrice

$$Q_0 = [\mathbf{e}_1 \mid \mathbf{e}_2],$$

si ottengono per gli elementi principali delle matrici R_k le successioni

k	$r_{11}^{(k)}$	$r_{22}^{(k)}$	$r_{33}^{(k)}$
1	-54.50334	-24.34186	-9.440650
2	98.97675	99.51717	2.877560
3	99.06108	99.93756	2.983708
4	99.06239	99.93680	2.998173
⋮	⋮	⋮	⋮
49	99.14136	99.85713	3.000000
50	99.14386	99.85464	3.000000

La convergenza al primo e al secondo autovalore è molto lenta perché è governata dal rapporto $|\lambda_2/\lambda_1|$, la convergenza al terzo autovalore, che viene approssimato in 8 iterazioni, dipende dai rapporti $|\lambda_3/\lambda_2|$ e $|\lambda_4/\lambda_3|$ e quindi è più rapida. Invece applicando l'accelerazione di Ritz al calcolo dei primi due autovalori si ottengono le successioni

k	$r_{11}^{(k)}$	$r_{22}^{(k)}$
1	99.58009	98.44128
2	100.0032	98.99542
3	100.0027	98.99692

e la velocità di convergenza è governata dal rapporto $|\lambda_3/\lambda_1| = 0.03$. ∎

13. Metodo di Lanczos

Come il metodo precedente, anche il *metodo di Lanczos* è particolarmente conveniente per matrici hermitiane sparse e di grandi dimensioni. Il metodo viene qui descritto per il caso di una matrice A ad elementi reali e simmetrica.

In questo metodo viene generata una successione di sottospazi S_k, $k = 1, 2, \ldots, n$, di dimensione k, tali che gli autovalori all'estremità dello spettro della restrizione di A ad S_k meglio approssimano gli autovalori estremi di A. Sia $\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_k$ una base ortonormale di S_k e $Q_k \in \mathbf{R}^{n\times k}$ la matrice che ha per colonne i vettori $\mathbf{q}_j$, $j = 1, 2, \ldots, k$. Dal teorema 6.9 segue che, se λ_1 e λ_n sono il massimo e minimo autovalore di A e $\mu_1^{(k)}$ e $\mu_k^{(k)}$ sono il massimo e minimo autovalore di $Q_k^T A Q_k$, allora

$$\lambda_n \leq \mu_k^{(k)} \quad \text{e} \quad \mu_1^{(k)} \leq \lambda_1.$$

All'aumentare di k, la successione dei $\mu_1^{(k)}$ è crescente e per $k = n$ si ha $\mu_1^{(k)} = \lambda_1$, e la successione dei $\mu_k^{(k)}$ è decrescente e per $k = n$ si ha $\mu_k^{(k)} = \lambda_n$. Nel metodo di Lanczos è fondamentale determinare i vettori $\mathbf{q}_j$, detti *vettori di Lanczos*, in modo tale che la successione dei $\mu_1^{(k)}$ cresca il più rapidamente possibile, e la successione dei $\mu_k^{(k)}$ decresca il più rapidamente possibile.

La matrice $Q_n^T A Q_n$ ottenuta dopo n passi dovrebbe in teoria avere gli stessi autovalori di A. Però a causa degli errori di arrotondamento ciò non accade. Per questo il procedimento, come si vedrà in seguito, viene utilizzato come metodo iterativo, arrestando il calcolo al k-esimo passo e approssimando gli autovalori all'estremità dello spettro.

Il metodo si basa sul calcolo della direzione di massima crescita del quoziente di Rayleigh (4) di una matrice A relativo ad un vettore $\mathbf{x} \neq \mathbf{0}$

$$r_A(\mathbf{x}) = \frac{\mathbf{x}^T A\mathbf{x}}{\mathbf{x}^T\mathbf{x}},$$

direzione che è quella individuata dal gradiente di $r_A(\mathbf{x})$, cioè dal vettore $\mathbf{g}(\mathbf{x})$, la cui componente i-esima $g_i(\mathbf{x})$, $i = 1, 2, \ldots, n$, è la derivata parziale di $r_A(\mathbf{x})$ rispetto alla componente x_i del vettore $\mathbf{x}$. Poiché

$$\begin{aligned} g_i(\mathbf{x}) &= \frac{\partial}{\partial x_i}\,\frac{\mathbf{x}^T A\mathbf{x}}{\mathbf{x}^T\mathbf{x}} = \frac{1}{\mathbf{x}^T\mathbf{x}}\,\left[\,\frac{\partial(\mathbf{x}^T A\mathbf{x})}{\partial x_i} - r_A(\mathbf{x})\,\frac{\partial(\mathbf{x}^T\mathbf{x})}{\partial x_i}\,\right] \\ &= \frac{2}{\mathbf{x}^T\mathbf{x}}\Big[\sum_{j=1}^{n} a_{ij}x_j - r_A(\mathbf{x})x_i\Big], \quad i = 1, \ldots, n, \end{aligned}$$

risulta

$$\mathbf{g}(\mathbf{x}) = \frac{2}{\mathbf{x}^T\mathbf{x}}\,[A\mathbf{x} - r_A(\mathbf{x})\mathbf{x}].$$

Cioè il vettore $\mathbf{g}(\mathbf{x})$ appartiene al sottospazio generato da $\mathbf{x}$ e $A\mathbf{x}$. Si noti che allo stesso sottospazio appartiene anche il vettore $-\mathbf{g}(\mathbf{x})$, che individua la direzione di massima decrescita.

Sia allora $k = 1$ e $\mathbf{q}_1$ un vettore di $\mathbf{R}^n$, $\|\mathbf{q}_1\|_2 = 1$; per l'autovalore $\mu_1^{(1)}$ di $Q_1^T A Q_1$ risulta

$$\mu_1^{(1)} = \mathbf{q}_1^T A\mathbf{q}_1 = r_A(\mathbf{q}_1).$$

Per quanto visto sopra, le direzioni di massima crescita e di massima decrescita di $r_A(\mathbf{x})$ nel punto $\mathbf{x} = \mathbf{q}_1$ appartengono al sottospazio generato da $\mathbf{q}_1$ e $A\mathbf{q}_1$. Se il vettore $A\mathbf{q}_1$ è linearmente indipendente da $\mathbf{q}_1$, come vettore $\mathbf{q}_2$ si sceglie un vettore, ortogonale a $\mathbf{q}_1$ tale che i due vettori $\mathbf{q}_1$ e $\mathbf{q}_2$ costituiscano una base ortonormale per il sottospazio S_2 generato da $\mathbf{q}_1$ e $A\mathbf{q}_1$. In questo modo il vettore $\mathbf{g}(\mathbf{q}_1)$ appartiene a S_2. Se invece $A\mathbf{q}_1$ è linearmente dipendente da $\mathbf{q}_1$, allora $\mathbf{q}_1$ è un autovettore di A e $\mu_1^{(1)}$ è un autovalore di A. Se in questa eventualità si vuole continuare ad applicare il metodo di Lanczos alla ricerca di un altro autovalore, si sceglie come vettore $\mathbf{q}_2$ un vettore ortonormale a $\mathbf{q}_1$.

In generale al k-esimo passo, $k > 1$, sia S_k il sottospazio, che si suppone di dimensione k, generato da $\mathbf{q}_1, A\mathbf{q}_1, \ldots, A^{k-1}\mathbf{q}_1$ e siano $\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_k$ i vettori di una base ortonormale di S_k. Poiché per il teorema 6.7 risulta

$$\mu_1^{(k)} = \max_{\mathbf{x}\neq\mathbf{0}} \frac{\mathbf{x}^T Q_k^T A Q_k \mathbf{x}}{\mathbf{x}^T\mathbf{x}} = \max_{\substack{\mathbf{y}\neq\mathbf{0} \\ \mathbf{y}\in S_k}} r_A(\mathbf{y}),$$

esiste un vettore $\mathbf{v} \in S_k,\ \mathbf{v} \neq \mathbf{0}$, tale che

$$\mu_1^{(k)} = r_A(\mathbf{v}).$$

Analogamente, poiché risulta

$$\mu_k^{(k)} = \min_{\mathbf{x}\neq\mathbf{0}} \frac{\mathbf{x}^T Q_k^T A Q_k \mathbf{x}}{\mathbf{x}^T \mathbf{x}} = \min_{\substack{\mathbf{y}\neq\mathbf{0} \\ \mathbf{y}\in S_k}} r_A(\mathbf{y}),$$

esiste un vettore $\mathbf{u} \in S_k,\ \mathbf{u} \neq \mathbf{0}$, tale che

$$\mu_k^{(k)} = r_A(\mathbf{u}).$$

La direzione di massima crescita di $r_A(\mathbf{x})$ nel punto $\mathbf{x} = \mathbf{v}$ è quella individuata dal vettore $\mathbf{g}(\mathbf{v})$, che appartiene al sottospazio generato da $\mathbf{v}$ e $A\mathbf{v}$. Poiché il vettore $\mathbf{v}$ appartiene al sottospazio generato da $\mathbf{q}_1, A\mathbf{q}_1, \ldots, A^{k-1}\mathbf{q}_1$, il vettore $\mathbf{g}(\mathbf{v})$ appartiene al sottospazio generato da $\mathbf{q}_1, A\mathbf{q}_1, \ldots, A^k\mathbf{q}_1$. Se il vettore $A^k\mathbf{q}_1$ è linearmente indipendente dai vettori $\mathbf{q}_1, A\mathbf{q}_1, \ldots, A^{k-1}\mathbf{q}_1$, come vettore $\mathbf{q}_{k+1}$ si sceglie allora un vettore, ortogonale a $\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_k$, tale che i vettori $\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_{k+1}$ costituiscano una base ortonormale per lo spazio S_{k+1} generato da $\mathbf{q}_1, A\mathbf{q}_1, \ldots, A^k\mathbf{q}_1$. Lo stesso ragionamento si può ripetere per la direzione di massima decrescita $-\mathbf{g}(\mathbf{u})$. Il sottospazio S_{k+1} risulta quindi includere le direzioni $\mathbf{g}(\mathbf{v})$ e $-\mathbf{g}(\mathbf{u})$.

Se invece il vettore $A^k\mathbf{q}_1$ è linearmente dipendente dai vettori $\mathbf{q}_1, A\mathbf{q}_1, \ldots, A^{k-1}\mathbf{q}_1$, allora il sottospazio S_k è *invariante*, cioè è generato da k autovettori $\mathbf{u}_{\sigma_1}, \mathbf{u}_{\sigma_2}, \ldots, \mathbf{u}_{\sigma_k}$ di A. Perciò ulteriori iterazioni del metodo di Lanczos produrrebbero sempre vettori appartenenti ad S_k. Questo fatto può essere sfruttato per calcolare autovalori corrispondenti agli autovettori $\mathbf{u}_{\sigma_i},\ i = 1, 2, \ldots, k$, ma non è possibile, partendo dal vettore $\mathbf{q}_1$, approssimare i successivi autovalori. Volendo comunque continuare ad applicare il metodo di Lanczos per calcolare i successivi autovalori, occorre generare un vettore $\mathbf{q}_{k+1}$ ortonormale a $\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_k$.

Il problema della determinazione dei vettori di Lanczos è così ricondotto al problema della determinazione di una base ortonormale del sottospazio generato da $\mathbf{q}_1, A\mathbf{q}_1, \ldots, A^k\mathbf{q}_1$, dove $\mathbf{q}_1$ è un vettore fissato inizialmente, tale che $\|\mathbf{q}_1\|_2 = 1$. Questa base viene determinata utilizzando il procedimento di tridiagonalizzazione di Lanczos, descritto nel paragrafo 5, che viene qui riportato:

siano $\alpha_1 = \mathbf{q}_1^T A\mathbf{q}_1, \quad \beta_0 = 0,$

per $i = 1, 2, \ldots, n-1$, si calcoli

$$\mathbf{r}_i = (A - \alpha_i I)\mathbf{q}_i - \beta_{i-1}\mathbf{q}_{i-1},$$

$$\begin{aligned}
&\beta_i = \|\mathbf{r}_i\|_2, \\
&\mathbf{q}_{i+1} = \begin{cases} \dfrac{\mathbf{r}_i}{\beta_i}, & \text{se } \beta_i \neq 0, \\ \text{un vettore ortonormale a } \mathbf{q}_1, \ldots, \mathbf{q}_i, & \text{se } \beta_i = 0, \end{cases} \\
&\alpha_{i+1} = \mathbf{q}_{i+1}^T A \mathbf{q}_{i+1}.
\end{aligned} \tag{57}$$

I vettori $\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n$ così generati sono, per il teorema 6.18, ortonormali e se i β_i sono tutti diversi da zero, coicidono con i vettori di Lanczos cercati. Infatti, indicata con Q la matrice che ha per colonne i vettori $\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n$ e con T la matrice tridiagonale

$$T = \begin{bmatrix} \alpha_1 & \beta_1 & & \\ \beta_1 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \beta_{n-1} \\ & & \beta_{n-1} & \alpha_n \end{bmatrix},$$

risulta che

$$A^i \mathbf{q}_1 = Q T^i \mathbf{e}_1, \ i = 0, 1, \ldots, n, \tag{58}$$

come si può dimostrare per induzione su i. Per $i = 0$ la (58) segue dal teorema 6.18, mentre per $i > 0$ è

$$A^i \mathbf{q}_1 = A(A^{i-1}\mathbf{q}_1) = A(QT^{i-1}\mathbf{e}_1) = QTQ^H QT^{i-1}\mathbf{e}_1 = QT^i\mathbf{e}_1.$$

Per la (58) è allora

$$[\mathbf{q}_1 \mid A\mathbf{q}_1 \mid \ \ldots \ \mid A^{n-1}\mathbf{q}_1] = Q[\mathbf{e}_1 \mid T\mathbf{e}_1 \mid \ \ldots \ \mid T^{n-1}\mathbf{e}_1], \tag{59}$$

e poiché la matrice T^i è a banda, di ampiezza i, il vettore $T^i\mathbf{e}_1$ ha nulle tutte le componenti di indice superiore a $i+1$, e quindi la (59) è la fattorizzazione QR della matrice

$$[\mathbf{q}_1 \mid A\mathbf{q}_1 \mid \ \ldots \ \mid A^{n-1}\mathbf{q}_1].$$

I vettori di Lanczos sono quindi le colonne della matrice ortogonale Q la cui prima colonna è $\mathbf{q}_1$ e tale che la matrice $Q^T AQ$ sia tridiagonale.

Se la matrice $[\mathbf{q}_1 \mid A\mathbf{q}_1 \mid \ \ldots \ \mid A^{n-1}\mathbf{q}_1]$ ha rango $k < n$, nella matrice T è nullo l'elemento β_k, e quindi il metodo di Lanczos si arresta alla k-esima iterazione, e viceversa. Se un tale elemento si annulla durante la costruzione della matrice T, ciò indica che è stato individuato un sottospazio S_k invariante.

Il procedimento di Lanczos consiste nell'applicare l'algoritmo precedente con arresto al k-esimo passo: si ottengono così i vettori $\mathbf{q}_i$, $i = 1, 2, \ldots, k$ e la sottomatrice principale di testa T_k della matrice tridiagonale T. Dalle (19) si ha che

$$AQ_k = Q_k T_k + \beta_k \mathbf{q}_{k+1}\mathbf{e}_k^T, \tag{60}$$

dove $\beta_k \mathbf{q}_{k+1} \mathbf{e}_k^T$ è la matrice i cui elementi non nulli compaiono solo nella k-esima colonna, che è uguale a $\beta_k \mathbf{q}_{k+1}$, e quindi si ha

$$Q_k^T A Q_k = T_k + \beta_k Q_k^T \mathbf{q}_{k+1} \mathbf{e}_k^T.$$

Poiché $Q_k^T \mathbf{q}_{k+1} = \mathbf{0}$, ne segue che

$$T_k = Q_k^T A Q_k.$$

Perciò la matrice T_k rappresenta la restrizione della matrice A al sottospazio S_k. I suoi autovalori estremi $\mu_1^{(k)}$ e $\mu_k^{(k)}$ possono essere poi determinati con uno dei metodi descritti per calcolare gli autovalori di matrici tridiagonali simmetriche. Vale la seguente stima degli autovalori, per la cui dimostrazione si rimanda a [7].

6.45 Teorema. *Sia $A \in \mathbf{R}^{n \times n}$ una matrice simmetrica i cui autovalori sono*

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$$

e i corrispondenti autovettori sono $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$. Sia $T_k \in \mathbf{R}^{k \times k}$ la matrice tridiagonale ottenuta al k-esimo passo del metodo di Lanczos, a partire dal primo vettore di Lanczos $\mathbf{q}_1$. Allora per gli autovalori estremi $\mu_1^{(k)}$ e $\mu_k^{(k)}$ di T_k valgono le seguenti limitazioni

$$\lambda_1 \geq \mu_1^{(k)} \geq \lambda_1 - \frac{(\lambda_1 - \lambda_n) tg^2(\phi_1)}{[c_{k-1}(\xi_1)]^2},$$
$$\lambda_n \leq \mu_k^{(k)} \leq \lambda_n + \frac{(\lambda_1 - \lambda_n) tg^2(\phi_n)}{[c_{k-1}(\xi_n)]^2},$$

dove

$$\cos(\phi_j) = |\mathbf{q}_1^T \mathbf{u}_j|, \quad \textit{per } j = 1, n,$$
$$\xi_1 = 1 + 2\,\frac{\lambda_1 - \lambda_2}{\lambda_2 - \lambda_n}, \qquad \xi_n = 1 + 2\,\frac{\lambda_{n-1} - \lambda_n}{\lambda_1 - \lambda_{n-1}},$$

e $c_{k-1}(x)$ è il polinomio di Chebyshev di 1^a specie di grado $k-1$, definito ricorsivamente per mezzo delle relazioni

$$\begin{cases} c_0(x) = 1 \\ c_1(x) = x \\ c_j(x) = 2x c_{j-1}(x) - c_{j-2}(x), \ j = 2, \ldots, k-1. \end{cases}$$

∎

Anche per gli altri autovalori di T_k si ha convergenza ad autovalori di A, ma l'approssimazione è migliore per gli autovalori estremi dello spettro di A e comunque il metodo non consente di valutare la molteplicità di un autovalore.

Come condizione di arresto per il metodo di Lanczos si può usare una qualunque delle condizioni usate per i metodi iterativi, applicate alla successione dei $\mu_1^{(k)}$ o $\mu_k^{(k)}$ calcolati. Un criterio semplice, che consente di valutare la precisione raggiunta dagli autovalori di T_k quando si disponga anche dei corrispondenti autovettori, si basa sul teorema 6.2 e si ottiene nel modo seguente.

Sia

$$T_k = U_k D_k U_k^T,$$

la forma normale di Schur di T_k, dove U_k è ortogonale e D_k è diagonale. Ponendo $Y_k = Q_k U_k$, dalla (60) si ha

$$AY_k = Y_k D_k + \beta_k \mathbf{q}_{k+1} \mathbf{e}_k^T U_k. \tag{61}$$

Per la (57) è $\beta_k \mathbf{q}_{k+1} = \mathbf{r}_k$, tale che $\|\mathbf{r}_k\|_2 = \beta_k$. Per $i = 1, 2, \ldots, k$, indicando con $\mathbf{y}_i$ la i-esima colonna di Y_k, detto i-esimo *vettore di Ritz* del sottospazio S_k, si ha dalla (61)

$$A\mathbf{y}_i = \mu_i^{(k)} \mathbf{y}_i + \mathbf{r}_k u_{ki},$$

dove u_{ki} è l'elemento di indici (k, i) di U_k. Dalla (2), poiché $\|\mathbf{y}_i\|_2 = 1$, segue allora che esiste un autovalore λ di A tale che

$$|\lambda - \mu_i^{(k)}| \leq \|A\mathbf{y}_i - \mu_i^{(k)} \mathbf{y}_i\|_2 = |u_{ki}| \ \|r_k\|_2 = |\beta_k| \ |u_{ki}|, \quad i = 1, \ldots, k.$$

Il metodo di tridiagonalizzazione di Lanczos presenta grossi problemi di stabilità numerica. Se β_k è piccolo, nel calcolo di $\mathbf{r}_k$ si possono presentare elevati errori di cancellazione, con una conseguente perdita di ortogonalità dei vettori $\mathbf{q}_k$, per cui i risultati ottenuti possono essere del tutto inattendibili. Si può in parte ovviare a questo inconveniente utilizzando la tecnica della *riortogonalizzazione completa*, cioè riortogonalizzando il vettore calcolato $\mathbf{q}_k$ rispetto ai vettori $\mathbf{q}_j$, $j = 0, 1, \ldots, k-1$. Però questo comporta un aumento del costo computazionale che fa perdere di competitività al metodo di Lanczos, e un aumento sostanziale dell'ingombro di memoria, poiché tutti i vettori $\mathbf{q}_j$ devono essere conservati.

Una tecnica più efficiente consiste nella *riortogonalizzazione selettiva*, in cui il vettore calcolato $\mathbf{q}_k$ viene riortogonalizzato solo rispetto ad alcuni dei vettori $\mathbf{q}_j$, $j = 0, 1, \ldots, k-1$, già calcolati. È stato infatti dimostrato da Paige [14] che la perdita di ortogonalità del vettore calcolato $\mathbf{q}_k$, dovuta agli errori di arrotondamento, diventa sempre più consistente quanto più $\mathbf{q}_k$

e i corrispondenti autovalori $\mu_1^{(k)}$ e $\mu_k^{(k)}$ si avvicinano ai loro limiti e inoltre che la perdità di ortogonalità si accentua nella direzione di quei vettori di Ritz $\mathbf{y}_i$ per i quali si è già avuta convergenza. Con la riortogonalizzazione selettiva è sufficiente memorizzare i soli vettori di Ritz $\mathbf{y}_i$ per cui si è avuta convergenza e riortogonalizzare rispetto ad essi ogni vettore $\mathbf{q}_k$.

Vi è infine da riportare una variante del metodo di Lanczos che non utilizza tecniche di riortogonalizzazione e che anzi sfrutta gli effetti degli errori di arrotondamento. Infatti a causa di questi errori è molto difficile che gli elementi β_k effettivamente calcolati diventino nulli, anche se viene approssimato un sottospazio invariante, e addirittura anche se $k = n$. Con questa variante si prolunga l'applicazione dell'algoritmo per un numero di passi maggiore di n (ad esempio fino a $k = 2n$ o $3n$). Senza riortogonalizzazione il processo tende a ripartire quando si perde l'ortogonalità ad un vettore di Ritz per cui si è avuta convergenza; si vengono così a generare approssimazioni multiple allo stesso autovalore e si ottiene alla fine una matrice T_k (ad esempio di dimensioni $2n$ o $3n$) in cui vi possono essere più autovalori corrispondenti ad un autovalore di A. Esistono però dei criteri per individuare quali fra gli autovalori di T_k sono quelli corrispondenti ad autovalori di A.

In modo analogo a quanto fatto per il metodo delle potenze, è possibile definire un metodo di Lanczos a blocchi.

6.46 Esempio. Le frequenze di vibrazione di una membrana elastica, vincolata al bordo $\partial\Omega$ di un dominio $\Omega \in \mathbf{R}^2$, possono essere determinate risolvendo il seguente problema:

$$\begin{cases} \dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial^2 u}{\partial y^2} = -\lambda^2 u & \text{su } \Omega, \\ u = 0 & \text{su } \partial\Omega, \end{cases}$$

dove u è una funzione definita su $\Omega \cup \partial\Omega$ a valori in $\mathbf{R}$, e se τ e ρ sono la tensione e la densità superficiale della membrana, le frequenze f sono date da $f = \dfrac{\lambda}{2\pi}\sqrt{\dfrac{\tau}{\rho}}$. Si supponga che Ω sia un triangolo equilatero di lato 1. Procedendo in modo analogo a quanto fatto nell'esempio 5.36, e fissato un intero m che determina il passo di discretizzazione, è possibile restringere la funzione u ai punti di un reticolo a maglia triangolare di lato $\dfrac{1}{m+2}$, ottenendo il problema di autovalori

$$A_n u = -\lambda^2 u,$$

in cui la matrice $A_n \in \mathbf{R}^{n\times n}$ è simmetrica, tridiagonale a blocchi, e della forma

$$A_n = \frac{2}{3(m+2)} \begin{bmatrix} B_m & H_m & & \\ H_m^T & B_{m-1} & \ddots & \\ & \ddots & \ddots & H_2 \\ & & H_2^T & B_1 \end{bmatrix}, \quad \text{con } n = \frac{m(m+1)}{2},$$

dove le sottomatrici $B_i \in \mathbf{R}^{i \times i}$, per $i = 1, \ldots, m$, sono tridiagonali e le sottomatrici $H_i \in \mathbf{R}^{i \times (i-1)}$, per $i = 1, \ldots, m-1$, sono bidiagonali

$$B_i = \begin{bmatrix} 6 & -1 & & \\ -1 & 6 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 6 \end{bmatrix}, \qquad H_i = \begin{bmatrix} -1 & & \\ -1 & \ddots & \\ & \ddots & -1 \\ & & -1 \end{bmatrix}.$$

Per valori grandi di n le matrici A_n sono sparse: se $m = 100$ (tale valore è normale nelle applicazioni) n risulta dell'ordine di 5000, mentre ogni riga della matrice ha al massimo 7 elementi non nulli. Delle matrici A_n si conosce l'espressione esplicita degli autovalori; per esempio, per $m = 8$, la matrice A_{36} ha 19 autovalori distinti, di cui 4 semplici, 14 di molteplicità 2 e uno, $\lambda_{17} = 8$, di molteplicità 4.

Per il caso $m = 8$, si è applicato il metodo di Lanczos per approssimare gli autovalori di A_{36}, assumendo come vettore iniziale $\mathbf{q}_0 = \mathbf{e}_1$ e si è applicato il procedimento per $k = 1, 2, \ldots, 10$. Gli autovalori di ogni matrice T_k sono stati successivamente calcolati con il metodo QR e sono riportati nella figura 6.4. Con i quadratini neri sono indicati i 19 autovalori distinti di A_{36}, e per ogni k con i quadratini bianchi sono indicati i k autovalori di T_k.

Proseguendo l'applicazione del metodo fino a $k = 36$, la matrice T_k ha come autovalori tutti gli autovalori di A_{36}, oltre ad altri autovalori molto vicini ad essi, e questo non consente di determinare la molteplicità degli autovalori di A_{36}. Ad esempio, in corrispondenza all'autovalore minimo di A_{36} $\lambda_{19} = 0.7639322$, di molteplicità 1, nella matrice T_{36} compaiono i tre autovalori $\mu_{34}^{(36)} = 0.7650051$, $\mu_{35}^{(36)} = 0.7650557$, $\mu_{36}^{(36)} = 0.7643948$, mentre in corrispondenza all'autovalore $\lambda_{17} = 8$, che ha molteplicità 4, nella matrice T_{36} compaiono solo i due autovalori $\mu_{31}^{(36)} = 7.999630$, $\mu_{32}^{(36)} = 8.000580$. L'approssimazione degli autovalori estremi dello spettro è comunque buona, già per bassi valori di k.

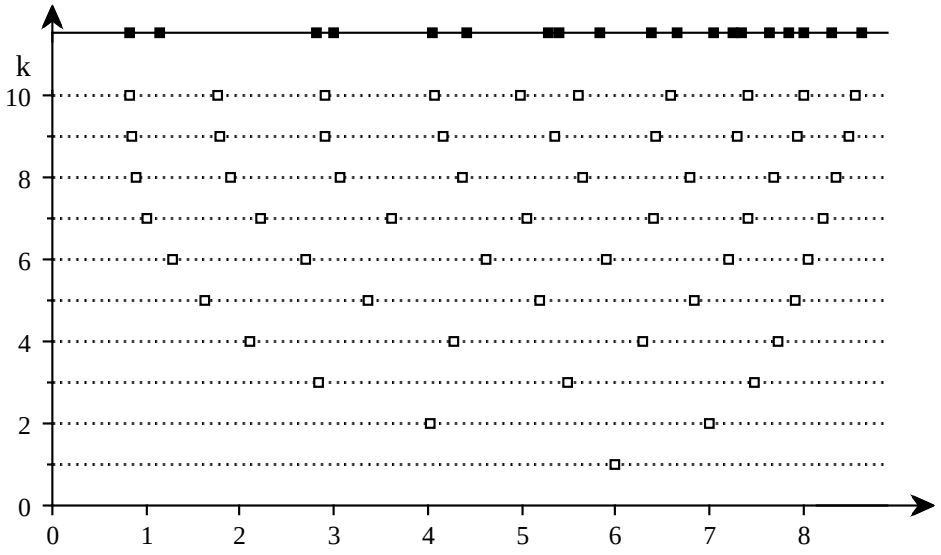


Fig. 6.4 - Metodo di Lanczos.

Nella figura 6.5 sono riportate in scala logaritmica, al crescere di k, le differenze

$$|\lambda_1 - \mu_1^{(k)}|,$$

cioè la successione degli errori di approssimazione dell'autovalore massimo. Dopo un'iniziale diminuzione, per $k > 17$ l'andamento oscillante degli errori indica che l'accumulo degli errori di arrotondamento non consente una approssimazione migliore.

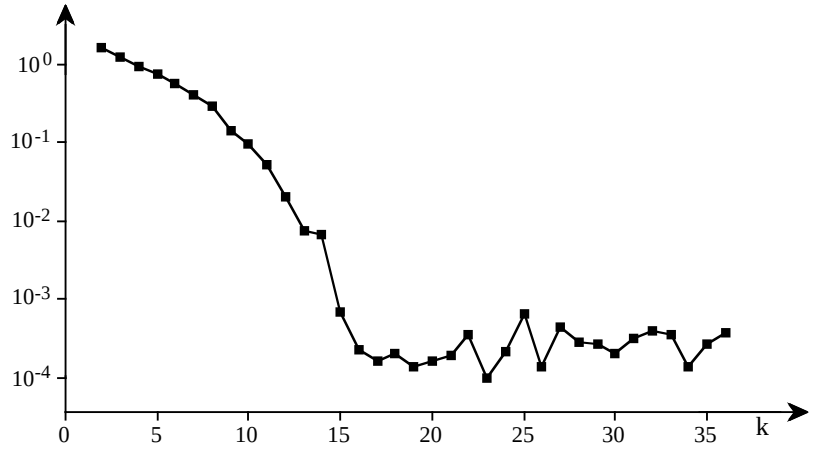


Fig. 6.5 - Errore dell'approssimazione del primo autovalore di A_{36} con il metodo di Lanczos

Per il caso $m = 18$, cioè $n = 171$, il metodo di Lanczos è stato applicato con riortogonalizzazione selettiva, usando la doppia precisione. I primi 10 autovalori sono stati calcolati con un errore assoluto massimo di 10^{-13} dopo 53 iterazioni, per un tempo totale di 3.04 secondi. Per confronto, il metodo delle iterazioni ortogonali, applicato con dimensione del sottospazio 10, fornisce i primi 10 autovalori con un errore assoluto massimo di 10^{-7} dopo 472 iterazioni per un tempo totale di 26.36 secondi. ∎

Esercizi proposti

6.1 La matrice

$$A = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 2 & 3 & 4 \\ 3 & 3 & 3 & 4 \\ 4 & 4 & 4 & 4 \end{bmatrix}$$

ha un autovalore λ approssimato da 13 e il corrispondente autovettore approssimato dal vettore $[0.675, 0.725, 0.830, 1]^T$. Si dia una limitazione superiore dell'errore da cui è affetta l'approssimazione data.

(Traccia: si applichi la (2).)

6.2 Sia $A \in \mathbf{R}^{n\times n}$ la matrice tridiagonale

$$A = \begin{bmatrix} \alpha & \beta & & \\ \beta & \alpha & \ddots & \\ & \ddots & \ddots & \beta \\ & & \beta & \alpha \end{bmatrix}.$$

Si dica se esistono autovalori λ di A tali che

$$(a) \quad |\lambda - \alpha| < |\beta| \frac{\sqrt{2}}{\sqrt{n}}\,, \qquad (b) \quad |\lambda - \alpha| > |\beta| \Big(2 - \frac{\sqrt{2}}{\sqrt{n}}\Big).$$

(Traccia: si consideri il vettore $\mathbf{x} = [1, 1, -1, -1, 1, 1, \ldots]^T$ e si applichi la (2) con $\sigma = \alpha$ per la (a), il vettore $\mathbf{x} = [1, 1, \ldots]^T$ e la (2) con $\sigma = \alpha + 2\beta$ per la (b).)

6.3 Siano $A_{11},\ A_{12},\ A_{22} \in \mathbf{C}^{n\times n}$, con A_{11} e A_{22} hermitiane e sia

$$A = \begin{bmatrix} A_{11} & A_{12} \\ \\ A_{12}^H & A_{22} \end{bmatrix}.$$

a) Si dimostri che per ogni autovalore μ di A_{11} esiste un autovalore λ di A tale che $|\lambda - \mu| \leq \|A_{12}\|_2$;

b) se

$$A_{12} = \begin{bmatrix} \mathbf{0} & O \\ \epsilon & \mathbf{0}^H \end{bmatrix},$$

si dica di quanto distano gli autovalori di A da quelli di A_{11} e A_{22}.

(Traccia: a) si applichi la (2) con $\mathbf{x} = \begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix}$, dove $A_{11}\mathbf{y} = \mu\mathbf{y}$, $\|\mathbf{y}\|_2 = 1$ e $\sigma = \mu$; b) $|\lambda - \mu| \leq |\epsilon|$.)

6.4 Sia $A \in \mathbf{C}^{n\times n}$ normale. Si dimostri che per $i = 1, \ldots, n$, esiste almeno un autovalore λ di A nel cerchio

$$\left\{ z \in \mathbf{C} \ : \ |z - a_{ii}| \leq \sqrt{\sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ji}|^2} \right\}.$$

(Traccia: si applichi il teorema 6.3 con $\mathbf{x} = \mathbf{e}_i$.)

6.5 Si calcolino gli autovalori $\lambda_i(\epsilon)$ e gli autovettori $\mathbf{x}_i(\epsilon)$ delle matrici

$$\begin{bmatrix} 1 & 1 \\ \epsilon & 1 \end{bmatrix}, \qquad \begin{bmatrix} 1 & 1 \\ 0 & 1+\epsilon \end{bmatrix}, \qquad \begin{bmatrix} 1 & 1+\epsilon \\ 0 & 1 \end{bmatrix}, \quad \text{dove } \epsilon \ll 1.$$

Si dica se il problema del calcolo degli autovalori della matrice

$$A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$

può essere mal condizionato.

(Traccia: gli autovalori sono rispettivamente $1 + \sqrt{\epsilon}$, $1 - \sqrt{\epsilon}$; 1, $1 + \epsilon$; 1. Gli autovettori corrispondenti sono $[1, -\sqrt{\epsilon}]^T$, $[1, \sqrt{\epsilon}]^T$; $[1, 0]^T$, $[1, -1]^T$; $[1, 0]^T$.)

6.6 Sono date le tre matrici

$$A_1 = \begin{bmatrix} \text{-2} & \text{-1} & 2 \\ 2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \text{-1} & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} \text{-1} & 1 & \text{-1} \\ 1 & \text{-1} & 1 \\ \text{-1} & 1 & \text{-1} \end{bmatrix}.$$

a) Si verifichi che A_1 e A_2 hanno gli stessi autovalori;

b) si dica di quanto risultano perturbati gli autovalori di A_1 e A_2 se le matrici sono perturbate nel modo seguente: $A_1 + \epsilon B$ e $A_2 + \epsilon B$, $0 < \epsilon \ll 1$ (si trascurino i termini di ordine superiore al primo in ϵ);

c) si indichi per la matrice $A_2 + \epsilon B$ la maggiorazione della perturbazione prodotta sugli autovalori che si ottiene applicando il teorema 6.14.

(Traccia: b) posto $\lambda_1 = 1,\ \lambda_2 = 0,\ \lambda_3 = -1$, gli autovalori di $A_1 + \epsilon B$ sono $\lambda_1(\epsilon) = 1 + \epsilon$, $\lambda_2(\epsilon) = -6\epsilon$, $\lambda_3(\epsilon) = -1 + 2\epsilon$, gli autovalori di $A_2 + \epsilon B$ sono $\lambda_1(\epsilon) = 1 - \epsilon$, $\lambda_2(\epsilon) = -\epsilon$, $\lambda_3(\epsilon) = -1 - \epsilon$, a meno di termini di ordine superiore al primo in ϵ; c) $-\epsilon \leq \lambda_i(\epsilon) - \lambda_i \leq 2\epsilon$.)

6.7 Una stima della perturbazione indotta sugli autovalori può essere ottenuta anche nel modo seguente: se $A = SDS^{-1}$, dove D è diagonale, gli autovalori di $A + \epsilon B$ sono anche gli autovalori di $D + \epsilon S^{-1}BS$ e possono essere localizzati con i teoremi di Gerschgorin.

a) Si utilizzi questa tecnica con le matrici dell'esercizio 6.6.

b) Sia $A \in \mathbf{C}^{n\times n}$ tale che $|a_{ij}| < \epsilon$ per $i,\ j = 1, \ldots, n,\ i \neq j$. Supponendo che ϵ sia sufficientemente piccolo, si determini un numero reale $\sigma > 0$ tale che, posto

$$S = \begin{bmatrix} \sigma & \mathbf{0}^H \\ \mathbf{0} & I_{n-1} \end{bmatrix},$$

il primo cerchio di Gerschgorin di SAS^{-1} sia il più piccolo possibile e disgiunto dagli altri. Quale risultato di perturbazione si ottiene?

(Traccia: a) per $A_1 + \epsilon B$ si ha $-4\epsilon < \lambda_1(\epsilon) - \lambda_1 < 6\epsilon$, $-12\epsilon < \lambda_2(\epsilon) - \lambda_2 < 0$, $-2\epsilon < \lambda_3(\epsilon) - \lambda_3 < 6\epsilon$; per $A_2 + \epsilon B$ si ha $-3\epsilon < \lambda_i(\epsilon) - \lambda_i < \epsilon$, $i = 1, 2, 3$; b) il primo cerchio di Gerschgorin ha centro in a_{11} e raggio minore o uguale a $(n-1)\epsilon\sigma$, l'i-esimo cerchio, per $i \neq 1$, ha centro in a_{ii} e raggio minore o uguale a $(n-2)\epsilon + \dfrac{\epsilon}{\sigma}$. Si deve pertanto determinare il minimo x per cui

$$(n-1)\epsilon x + (n-2)\epsilon + \frac{\epsilon}{x} \leq \delta, \quad \delta = \min_{i\neq 1} |a_{11} - a_{ii}|.$$

Si ha quindi

$$\sigma = \min x = \frac{\delta - (n-2)\epsilon - \sqrt{[\delta - (n-2)\epsilon]^2 - 4(n-1)\epsilon^2}}{2(n-1)\epsilon} = \frac{\epsilon}{\delta} + O(\epsilon^2)$$

e

$$|\lambda - a_{11}| \leq \frac{(n-1)\epsilon^2}{\delta} + O(\epsilon^3).\ )$$

6.8 Si determinino degli intervalli di localizzazione degli autovalori della matrice

$$A = \begin{bmatrix} 3 & 0.5 & \epsilon & 0 \\ 0.5 & 1 & -\epsilon & 0 \\ \epsilon & -\epsilon & 2 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix},$$

dove ϵ è un numero reale di modulo sufficientemente piccolo, e si determini un valore $\bar{\epsilon}$, tale che per $\epsilon \leq \bar{\epsilon}$ gli autovalori di A siano separati.

(Traccia: si scriva $A = B + \epsilon F$, si determinino gli autovalori di B e si applichi il teorema 6.14.)

6.9 Sia $A \in \mathbf{C}^{n\times n}$ e siano λ un autovalore di A di molteplicità algebrica 1 e $\mathbf{x}$ l'autovettore corrispondente, tale che $\|\mathbf{x}\|_2 = 1$.

a) se $\mathbf{y}$ è un autovettore sinistro di A normalizzato in norma 2, corrispondente a λ, cioè un vettore $\mathbf{y}$ tale che

$$\mathbf{y}^H A = \lambda \mathbf{y}^H, \quad \|\mathbf{y}\|_2 = 1,$$

si dimostri che $\mathbf{y}^H \mathbf{x} \neq 0$;

b) sia $F \in \mathbf{C}^{n\times n}$, $\|F\|_2 = 1$ e sia $\lambda(\epsilon)$ un autovalore di $A + \epsilon F$ corrispondente all'autovettore $\mathbf{x}(\epsilon)$. Sapendo che $\lambda(\epsilon)$ e le componenti di $\mathbf{x}(\epsilon)$ sono funzioni analitiche di ϵ in un opportuno intorno dello zero, si dimostri che

$$\lambda(\epsilon) - \lambda = \epsilon \, \frac{\mathbf{y}^H F \mathbf{x}}{\mathbf{y}^H \mathbf{x}} + O(\epsilon^2);$$

c) dalla relazione trovata al punto b) si ottiene

$$|\lambda(\epsilon) - \lambda| \leq \frac{1}{\mathbf{y}^H \mathbf{x}} \, |\epsilon| + O(\epsilon^2), \tag{62}$$

da cui segue che una perturbazione dell'ordine di $|\epsilon|$ sugli elementi di A genera sull'autovalore λ una perturbazione dell'ordine di $|\epsilon|/|\mathbf{y}^H \mathbf{x}|$. La quantità $1/|\mathbf{y}^H \mathbf{x}|$ viene detta *numero di condizionamento* di λ. Si determini una matrice F tale che la relazione (62) valga con il segno di uguaglianza;

d) si determini il numero di condizionamento dell'autovalore $\lambda = 1$ della matrice

$$A = \begin{bmatrix} 1 & \alpha \\ 0 & 2 \end{bmatrix}, \quad \alpha \in \mathbf{R},$$

e si valuti la perturbazione generata su λ dalla perturbazione

$$\epsilon F, \quad F = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix},$$

introdotta sulla matrice A. Se $\alpha = 10^3$ e $\epsilon = 10^{-6}$, si verifichi che la perturbazione è dell'ordine di 10^{-3};

e) si determini un *numero di condizionamento* per l'autovettore $\mathbf{x}$, cioè un γ tale che
$$\|\mathbf{x}(\epsilon) - \mathbf{x}\|_2 \leq \gamma|\epsilon|,$$
e si metta in relazione tale numero con gli autovalori di A; si esamini in particolare il caso in cui A è hermitiana;

f) si determini il numero di condizionamento dell'autovettore $\mathbf{x}$ relativo all'autovalore $\lambda = 1$ della matrice
$$A = \begin{bmatrix} 1 & 0 \\ 0 & \beta \end{bmatrix}, \quad \beta \in \mathbf{R},$$
e si valuti la perturbazione generata su $\mathbf{x}$ dalla perturbazione
$$\epsilon F, \quad F = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix},$$
introdotta sulla matrice A. Se $\beta = 1 - 10^{-3}$ e $\epsilon = 10^{-6}$, di che ordine è questa perturbazione?

(Traccia: a) Sia $U = [\,\mathbf{x} \mid U_1\,]$ una matrice unitaria, con $U_1 \in \mathbf{C}^{n\times(n-1)}$, tale che
$$U^H A U = \begin{bmatrix} \lambda & \mathbf{c}^H \\ \mathbf{0} & B \end{bmatrix},$$
in cui λ non è autovalore di B. Si verifichi che il vettore $\begin{bmatrix} 1 \\ \mathbf{z} \end{bmatrix}$, $\mathbf{z} \in \mathbf{C}^{n-1}$, è un autovettore sinistro di $U^H A U$ se $\mathbf{z}^H = \mathbf{c}^H(\lambda I - B)^{-1}$, e quindi $\mathbf{w} = U \begin{bmatrix} 1 \\ \mathbf{z} \end{bmatrix}$ è un autovettore sinistro di A, per cui
$$\mathbf{y} = \frac{\mathbf{w}}{\|\mathbf{w}\|_2} = \frac{1}{\sqrt{1 + \|\mathbf{z}\|_2^2}}\, U \begin{bmatrix} 1 \\ \mathbf{z} \end{bmatrix}.$$
Si ha poi
$$\mathbf{w}^H\mathbf{x} = [\,1 \mid \mathbf{z}^H\,]\, U^H\mathbf{x} = [\,1 \mid \mathbf{z}^H\,]\, \mathbf{e}_1 = 1, \quad \text{e} \quad \mathbf{y}^H\mathbf{x} = \frac{1}{\sqrt{1 + \|\mathbf{z}\|_2^2}}\,;$$
b) poiché $\lambda(\epsilon)$ e le componenti di $\mathbf{x}(\epsilon)$ sono funzioni analitiche di ϵ, esiste $\mu \in \mathbf{C}$ e $\mathbf{q} \in \mathbf{C}^n$ tali che $\lambda(\epsilon) = \lambda + \epsilon\mu + O(\epsilon^2)$, $\mathbf{x}(\epsilon) = \mathbf{x} + \epsilon\mathbf{q} + O(\epsilon^2)$. Dalla relazione
$$(A + \epsilon F)(\mathbf{x} + \epsilon\mathbf{q}) = (\lambda + \epsilon\mu)(\mathbf{x} + \epsilon\mathbf{q}) + O(\epsilon^2),$$

tenendo conto che $A\mathbf{x} = \lambda\mathbf{x}$ si ottiene

$$(F - \mu I)\mathbf{x} = (\lambda I - A)\mathbf{q} + O(\epsilon^2). \tag{63}$$

Premoltiplicando per $\mathbf{y}^H$, e tenendo conto che $\mathbf{y}^H A = \lambda\mathbf{y}^H$ e che $\mathbf{y}^H\mathbf{x} \neq 0$, si ottiene

$$\mu = \frac{\mathbf{y}^H F\mathbf{x}}{\mathbf{y}^H\mathbf{x}} + O(\epsilon^2);$$

c) è $|\mathbf{y}F\mathbf{x}^H| \leq \|F\|_2 \leq 1$; $F = \mathbf{y}\mathbf{x}^H$;
e) è $\mathbf{x}(\epsilon) = \mathbf{x} + \epsilon\mathbf{q} + O(\epsilon^2)$, per cui $\gamma = \|\mathbf{q}\|_2$. Inoltre, poiché si può supporre che $\mathbf{q}^H\mathbf{x} = 0$, da (63) si ha

$$U^H(F - \mu I)UU^H\mathbf{x} = \begin{bmatrix} 0 & \mathbf{c}^H \\ \mathbf{0} & B - \lambda I \end{bmatrix} U^H\mathbf{q} + O(\epsilon^2),$$

da cui

$$\begin{aligned} \|\mathbf{q}\|_2 = \|U^H\mathbf{q}\|_2 &\leq \|(B - \lambda I)^{-1}\|_2 \|(F - \mu I)\mathbf{x}\|_2 \\ &\leq \|(B - \lambda I)^{-1}\|_2 \Big(1 + \frac{\mathbf{y}^H F\mathbf{x}}{\mathbf{y}^H\mathbf{x}}\Big). \end{aligned}$$

Se A è hermitiana, è $\qquad \|(\lambda I - B)^{-1}\|_2 = \dfrac{1}{\min\limits_{\lambda \neq \lambda_i} |\lambda - \lambda_i|}$,

dove $\lambda_i,\ i = 1, \ldots, n$, sono gli autovalori di A. Quindi un autovettore è tanto meglio condizionato quanto più l'autovalore corrispondente è separato dagli altri. Questo non è vero, in generale, per una matrice non hermitiana; f)

$$\mathbf{x} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \|(\lambda I - B)^{-1}\|_2 = \frac{1}{|1 - \beta|},$$

$$\mathbf{x}(\epsilon) = \begin{bmatrix} 1 \\ \epsilon/(1 - \beta) \end{bmatrix} \frac{1}{\sqrt{1 + [\epsilon/(1 - \beta)]^2}} = \begin{bmatrix} 1 \\ \epsilon/(1 - \beta) \end{bmatrix} + O(\epsilon^2).$$

Quindi il numero di condizionamento è elevato se β è vicino a 1. Nel caso particolare è $\dfrac{1}{|1 - \beta|} = 10^3$ e $\|\mathbf{x}(\epsilon) - \mathbf{x}\|_2 \approx 10^{-3}$.)

6.10 Sia $A \in \mathbf{R}^{n \times n}$ la matrice i cui elementi sono

$$a_{ij} = \begin{cases} 1 & \text{se } i = j \neq 1, \\ -1 & \text{se } i < j, \\ 0 & \text{altrimenti.} \end{cases}$$

a) Si determini il numero di condizionamento dell'autovalore $\lambda = 0$ di A e si valuti la perturbazione generata su λ da una perturbazione di modulo ϵ sull'elemento (i, j)-esimo, con $(i, j) \neq (1, 1)$,

b) si determini una maggiorazione del numero di condizionamento per l'autovettore $\mathbf{x}$ relativo all'autovalore $\lambda = 0$.

(Traccia: a) è

$$\mathbf{x} = \mathbf{e}_1,\ \mathbf{y} = \sqrt{\frac{3}{2+4^{n-1}}}\,[1,1,2,2^2,\ldots,2^{n-2}]^T,\ \frac{1}{\mathbf{y}^T\mathbf{x}} = \sqrt{\frac{2+4^{n-1}}{3}}\ ,$$

$$F\mathbf{x} = \begin{cases} \mathbf{0} & \text{se } j \neq 1, \\ \\ \mathbf{e}_i & \text{se } j = 1, \end{cases} \qquad \frac{\mathbf{y}^T F\mathbf{x}}{\mathbf{y}^T\mathbf{x}} = \begin{cases} 0 & \text{se } j \neq 1, \\ \\ 2^{i-2} & \text{se } j = 1, i \neq 1. \end{cases}$$

Perciò la perturbazione generata su λ è dell'ordine di $2^{i-2}|\epsilon|$ se $j = 1$, di ordine superiore al primo in ϵ se $j \neq 1$; b) si ha $U = I$, quindi

$$B - \lambda I = B = \begin{bmatrix} 1 & -1 & -1 & \ldots & -1 \\ & 1 & -1 & \ddots & \vdots \\ & & 1 & \ddots & -1 \\ & & & \ddots & -1 \\ & & & & 1 \end{bmatrix}, \quad B^{-1} = \begin{bmatrix} 1 & 1 & 2 & \ldots & 2^{n-2} \\ & 1 & 1 & \ddots & \vdots \\ & & 1 & \ddots & 2 \\ & & & \ddots & 1 \\ & & & & 1 \end{bmatrix}$$

e $\|B^{-1}\|_2 \leq \sqrt{n}\,\|B^{-1}\|_\infty = 2^{n-1}\sqrt{n}$.)

6.11 Si dica quali sono i numeri di condizionamento degli autovalori della seguente matrice di ordine 20

$$A = \begin{bmatrix} 20 & 20 & & & & & \\ & 19 & 20 & & & & \\ & & 18 & 20 & & & \\ & & & \ddots & \ddots & & \\ & & & & & 2 & 20 \\ & & & & & & 1 \end{bmatrix}.$$

(Traccia: per $\lambda = 20$ si ha $\mathbf{x} = \mathbf{e}_1$, il vettore $\mathbf{y}$ soddisfa alle equazioni

$$20y_i - iy_{i+1} = 0,\ i = 1,\ldots,19,$$

e quindi, a parte il fattore di normalizzazione, si ha

$$y_1 = 1,\ y_2 = 20,\ y_3 = \frac{20^2}{2}\ ,\ldots,y_i = \frac{20^{i-1}}{(i-1)!}\ ,\ldots,y_{20} = \frac{20^{19}}{19!}$$

e

$$\|\mathbf{y}\|_2^2 = \sum_{i=1}^{n} \frac{20^{2(i-1)}}{(i-1)!^2}$$

e la radice quadrata di tale valore è il numero di condizionamento di $\lambda = 20$. Per il generico autovalore $\lambda = r$, si dimostri che

$$\mathbf{x} = \left[1,\ \frac{20-r}{-20}\ ,\ \frac{(20-r)(19-r)}{(-20)^2}\ ,\ldots,\ \frac{(20-r)!}{(-20)^{20-r}}\ ,0,\ldots,0\right]^T$$

$$\mathbf{y} = \left[0,\ldots,0,1,20,\ldots,\ \frac{20^{r-2}}{(r-2)!}\ ,\ \frac{20^{r-1}}{(r-1)!}\ \right]^T,$$

a parte i fattori di normalizzazione. I numeri di condizionamento effettivamente calcolati sono dati nella seguente tabella

r	$1/\lvert\mathbf{y}^T\mathbf{x}\rvert$	r	$1/\lvert\mathbf{y}^T\mathbf{x}\rvert$
1, 20	$8.45\ 10^7$	6, 15	$6.94\ 10^{11}$
2, 19	$1.46\ 10^9$	7, 14	$1.57\ 10^{12}$
3, 18	$1.21\ 10^{10}$	8, 13	$2.84\ 10^{12}$
4, 17	$6.39\ 10^{10}$	9, 12	$4.18\ 10^{12}$
5, 16	$2.42\ 10^{11}$	10, 11	$5.07\ 10^{12}$

6.12 Sia $A \in \mathbf{C}^{n\times n}$ normale, con autovalori $|\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_n|$ e $U,\ V \in \mathbf{C}^{n\times n}$ unitarie. Indicati con $\lambda_i(UA)$, $\lambda_i(AV)$, $\lambda_i(UAV)$ gli autovalori di UA, AV, UAV, si dimostri che

$$\left.\begin{array}{l} |\lambda_1| \geq |\lambda_i(UA)| \geq |\lambda_n| \\ |\lambda_1| \geq |\lambda_i(AV)| \geq |\lambda_n| \\ |\lambda_1| \geq |\lambda_i(UAV)| \geq |\lambda_n| \end{array}\right\}, \quad i = 1,\ldots,n.$$

(Traccia: posto $\mu = \lambda_i(UA)$, si ha $UA\mathbf{x} = \mu\mathbf{x}$, $\|\mathbf{x}\|_2 = 1$, per cui

$$\mathbf{x}^H A^H U^H UA\mathbf{x} = |\mu|^2, \text{ e quindi } |\mu|^2 = \mathbf{x}^H A^H A\mathbf{x}.$$

Essendo $A^H A$ hermitiana, per il teorema 6.7 $|\mu|^2$ è compreso fra il massimo e il minimo autovalore di $A^H A$. Poiché A è normale, esiste Z unitaria tale che $A = ZDZ^H$, D diagonale, e quindi $A^H A = ZD^H DZ^H$, per cui gli autovalori di $A^H A$ sono i quadrati dei moduli degli autovalori di A. Si dimostrino le altre due relazioni in modo analogo.)

6.13 Sia $A \in \mathbf{C}^{n\times n}$ hermitiana. Si dimostri che se A ha un autovalore λ di molteplicità algebrica m, allora ogni matrice $U^H AU$, dove $U \in \mathbf{C}^{n\times(n-1)}$ è tale che $U^H U = I_{n-1}$, ha l'autovalore λ di molteplicità almeno $m-1$.

(Traccia: si applichi il teorema 6.8, essendo $\lambda_i = \ldots = \lambda_{i+m-1} = \lambda$, risulta

$$\lambda_i = \mu_i = \lambda_{i+1} = \ldots = \mu_{i+m-2} = \lambda_{i+m-1}.\ )$$

6.14 Sia $A \in \mathbf{C}^{n\times n}$ hermitiana e per $m \leq n$ sia $U_m \in \mathbf{C}^{n\times m}$ tale che $U_m^H U_m = I_m$ e siano $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ gli autovalori di A e $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_m$ gli autovalori di $U_m^H A U_m$. Si dimostri che valgono le seguenti relazioni

$$\mu_i \leq \lambda_i,\ i = 1, \ldots, m, \qquad \lambda_{n-i+1} \leq \mu_{m-i+1},\ i = 1, \ldots, m.$$

(Traccia: si consideri una successione di matrici $V_i,\ i = 1, \ldots, n-m$, tale che $V_i \in \mathbf{C}^{(n-i+1)\times(n-i)}$, $V_i^H V_i = I_{n-i}$ e $V_1 V_2 \ldots V_{n-m} = U_m$ e si applichi il teorema 6.8 alle matrici $A_i = V_i^H A_{i-1} V_i$, con $A_0 = A$.)

6.15 Sia $A \in \mathbf{C}^{n\times n}$ hermitiana, con autovalori $\lambda_1 \geq \lambda_2 \leq \ldots \geq \lambda_n$ e sia k intero, $k \leq n$. Indicato con $\mathcal{U}_k$ l'insieme delle matrici $U \in \mathbf{C}^{n\times k}$ tali che $U^H U = I_k$, si dimostri che

$$\max_{U\in\mathcal{U}_k} \text{tr}\ (U^H A U) = \lambda_1 + \lambda_2 + \ldots + \lambda_k,$$

$$\min_{U\in\mathcal{U}_k} \text{tr}\ (U^H A U) = \lambda_n + \lambda_{n-1} + \ldots + \lambda_{n-k+1}.$$

(Traccia: per l'esercizio 6.14 è tr $(U^H A U) \leq \lambda_1 + \lambda_2 + \ldots + \lambda_k$. D'altra parte, se $U = [\ \mathbf{u}_1 \mid \ldots \mid \mathbf{u}_k\ ]$, dove $A\mathbf{u}_i = \lambda_i \mathbf{u}_i$, $\|\mathbf{u}_i\|_2 = 1$, è tr $(U^H A U) = \lambda_1 + \lambda_2 + \ldots + \lambda_k$. Si dimostri l'altra relazione in modo analogo.)

6.16 Siano A e $B \in \mathbf{C}^{n\times n}$ hermitiane e A sia definita positiva. Si dimostri che se $\|A^{-1}\|_2\ \|B\|_2 < 1$, allora $A + B$ è definita positiva.

(Traccia: si dimostri, applicando il teorema 6.14, che

$$\lambda_i - \|B\|_2 \leq \mu_i \leq \lambda_i + \|B\|_2,$$

dove λ_i, μ_i, $i = 1, \ldots, n$, sono gli autovalori di A e di $A + B$. Se λ_n è il minimo dei λ_i, è $\lambda_n - \|B\|_2 \leq \mu_i$, $i = 1, \ldots, n$, e $\lambda_n = \dfrac{1}{\|A^{-1}\|_2}$.)

6.17 Sia $A \in \mathbf{C}^{n\times n}$ hermitiana.

a) Si dimostri che la matrice

$$B = \begin{bmatrix} A & \mathbf{b} \\ \mathbf{b}^H & \alpha \end{bmatrix}$$

ha almeno un autovalore λ tale che $|\lambda - \alpha| \leq \|\mathbf{b}\|_2$;

b) si dica sotto quale ipotesi gli autovalori di A separano strettamente quelli di B.

(Traccia: a) si applichi la (2) con $\mathbf{x} = \begin{bmatrix} \mathbf{0} \\ 1 \end{bmatrix}$, $\mathbf{0} \in \mathbf{R}^n$ e $\sigma = \alpha$; b) A ha autovalori distinti e la matrice $[\, A - \mu I \mid \mathbf{b} \,]$ ha rango n per ogni autovalore μ di A.)

6.18 Siano $\lambda_1 \geq \lambda_2 \geq \lambda_3$ e $\mu_1 \geq \mu_2 \geq \mu_3$ gli autovalori delle matrici

$$A = \begin{bmatrix} 2 & 3 & -2 \\ 3 & 1 & 0 \\ -2 & 0 & -1 \end{bmatrix} \quad \text{e} \quad B = \begin{bmatrix} 2.1 & 2.9 & -2 \\ 2.9 & 0.9 & 0.1 \\ -2 & 0.1 & -1 \end{bmatrix}.$$

Si determini un numero α tale che $|\lambda_i - \mu_i| \leq \alpha$, $i = 1, \ldots, 3$.

6.19 Sia $A \in \mathbf{C}^{n \times n}$ normale e sia $A = QR$, Q unitaria, R triangolare superiore. Si dimostri che

$$\min_{i=1,\ldots,n} |\lambda_i| \leq |r_{jj}| \leq \max_{i=1,\ldots,n} |\lambda_i|, \quad j = 1, \ldots, n,$$

in cui λ_i, $i = 1, \ldots, n$, sono gli autovalori di A.

(Traccia: posto $A = UDU^H$, U unitaria, D diagonale, risulta $A^H A = UD^H DU^H = R^H R$, da cui $D^H D = U^H R^H R U$. Sia $\mathbf{x}$ tale che $U\mathbf{x} = \mathbf{e}_j$, quindi $\mathbf{x}^H \mathbf{x} = 1$ e

$$\mathbf{x}^H D^H D \mathbf{x} = \mathbf{e}_j^T R^H R \mathbf{e}_j = \sum_{i=1}^{j} |r_{ij}|^2.$$

Poiché $\mathbf{x}^H D^H D \mathbf{x} \leq \max_{i=1,\ldots,n} |\lambda_i|^2$, segue che $|r_{jj}| \leq \max_{i=1,\ldots,n} |\lambda_i|$. Per l'altra disuguaglianza se $r_{jj} \neq 0$ per ogni j, si proceda in modo analogo, essendo $(D^H D)^{-1} = U^H R^{-1} R^{-H} U$.)

6.20 Sia $A_n \in \mathbf{R}^{n \times n}$ una matrice reale e simmetrica ad albero, cioè della forma

$$A_n = \begin{bmatrix} \alpha_1 & \alpha_2 & \alpha_3 & \cdots & \alpha_n \\ \alpha_2 & \beta_2 & 0 & \cdots & 0 \\ \alpha_3 & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \beta_{n-1} & 0 \\ \alpha_n & 0 & \cdots & 0 & \beta_n \end{bmatrix},$$

con $\alpha_i \neq 0$, per $i = 1, \ldots, n$ e β_i a due a due distinti per $i = 2, \ldots, n$. Si dimostri che gli autovalori di A_{n-1} separano strettamente quelli di A_n.

(Traccia: per la separazione debole si applichi il teorema 6.8, con $U = \begin{bmatrix} I_{n-1} \\ \mathbf{0}^H \end{bmatrix}$; si supponga che $\lambda \neq \beta_i$, $i = 2, \ldots, n-1$, sia autovalore di A_{n-1} e di A_n, e si concluda che $\alpha_n = 0$; si dimostri poi che non è possibile che β_i sia autovalore di A_n e A_{n-1}.)

6.21 Siano $\mathbf{a}_i$ e $\mathbf{a}_j$ due diverse colonne di una matrice $A \in \mathbf{C}^{n\times n}$, tali che

$$\mathbf{a}_i^H \mathbf{a}_j = \epsilon \|\mathbf{a}_i\|_2 \, \|\mathbf{a}_j\|_2, \quad |\epsilon| < 1,$$

dove ϵ dà una misura dell'angolo compreso fra i due vettori $\mathbf{a}_i$ e $\mathbf{a}_j$. Si dimostri che o A è singolare, oppure

$$\mu_2(A) \geq \frac{1+|\epsilon|}{1-|\epsilon|}.$$

(Traccia: se A è non singolare, sia $B = A^H A$, allora la matrice

$$C = \begin{bmatrix} \mathbf{a}_i^H \mathbf{a}_i & \mathbf{a}_i^H \mathbf{a}_j \\ \\ \mathbf{a}_j^H \mathbf{a}_i & \mathbf{a}_j^H \mathbf{a}_j \end{bmatrix}$$

è sottomatrice principale di B. Indicando con $\lambda(B)$ e $\lambda(C)$ gli autovalori di B e di C, e con $\lambda_{\max}$ e $\lambda_{\min}$ gli autovalori massimo e minimo, vale

$$\lambda_{\min}(B) \leq \lambda(C) \leq \lambda_{\max}(B),$$

per cui

$$\mu_2^2(A) = \frac{\lambda_{\max}(B)}{\lambda_{\min}(B)} \geq \frac{\lambda_{\max}(C)}{\lambda_{\min}(C)} = \mu_2(C).$$

Si dimostri poi che il valore minimo del numero di condizionamento di C si ha quando $\mathbf{a}_i^H \mathbf{a}_i = \mathbf{a}_j^H \mathbf{a}_j$ e vale $\dfrac{1+|\epsilon|}{1-|\epsilon|}$.)

6.22 Si trasformi la matrice

$$A = \begin{bmatrix} 4 & -1 & -1 & 0 \\ -1 & 4 & 0 & -1 \\ -1 & 0 & 4 & -1 \\ 0 & -1 & -1 & 4 \end{bmatrix}$$

in forma tridiagonale con i metodi di Householder, di Givens e di Lanczos.

6.23 Si applichi il metodo di tridiagonalizzazione di Givens alla matrice

$$A = \begin{bmatrix} 1 & \sqrt{2} & \sqrt{2} & 2 \\ \sqrt{2} & -\sqrt{2} & -1 & \sqrt{2} \\ \sqrt{2} & -1 & \sqrt{2} & \sqrt{2} \\ 2 & \sqrt{2} & \sqrt{2} & -3 \end{bmatrix}.$$

6.24 Sia $A \in \mathbf{C}^{n\times n}$ hermitiana. Si dica come modificare il metodo di tridiagonalizzazione di Householder affinché la matrice $A^{(n-1)}$ tridiagonale ottenuta sia reale.

(Traccia: con la notazione del paragrafo 5, si costruisca la matrice

$$\widetilde{P}^{(k)} = \frac{\overline{\alpha}_k}{|\alpha_k|} P^{(k)} \quad \text{e} \quad T_k = \begin{bmatrix} I_k & \mathbf{0}^H \\ \mathbf{0} & \widetilde{P}^{(k)} \end{bmatrix}.)$$

6.25 Sia $A \in \mathbf{C}^{n\times n}$ antihermitiana (cioè $A^H = -A$). Si dica che struttura ha la matrice $A^{(n-1)}$ ottenuta applicando il metodo di tridiagonalizzazione di Householder ad A.

(Traccia: $A^{(n-1)}$ risulta tridiagonale antihermitiana, si sfrutti il punto b) dell'esercizio 1.21.)

6.26 Sia B_n la matrice tridiagonale hermitiana

$$B_n = \begin{bmatrix} \alpha_1 & \overline{\beta}_2 & & \\ \beta_2 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \overline{\beta}_n \\ & & \beta_n & \alpha_n \end{bmatrix}.$$

Si dimostri che se $\beta_i \neq 0$, $i = 2, \ldots, n$, allora gli autovalori di B_n sono tutti distinti.

(Traccia: se λ fosse autovalore di molteplicità maggiore di 1 di B_n, poiché gli autovalori di B_{n-1} separano quelli di B_n, λ dovrebbe essere anche autovalore di B_{n-1} e per la (21) anche di B_i, per $i = n-2, \ldots, 0$, il che è assurdo.)

6.27 Sia $A \in \mathbf{C}^{n\times n}$ in forma di Hessenberg superiore. Si dimostri che se A ha un autovalore λ di molteplicità algebrica e geometrica maggiore di 1, allora A ha un elemento sottodiagonale nullo.

(Traccia: sia λ tale che $A\mathbf{x} = \lambda\mathbf{x}$, $A\mathbf{y} = \lambda\mathbf{y}$, con $\mathbf{x}$ e $\mathbf{y}$ linearmente indipendenti. Si supponga che $x_n = y_n$; posto $\mathbf{z} = \mathbf{x} - \mathbf{y}$, si ha

$$A\mathbf{z} = \lambda\mathbf{z}, \quad \mathbf{z} \neq \mathbf{0}, \quad z_n = 0.$$

Posto

$$A - \lambda I = \begin{bmatrix} \mathbf{c}^H & \alpha \\ B & \mathbf{d} \end{bmatrix}, \quad \mathbf{z} = \begin{bmatrix} \mathbf{v} \\ 0 \end{bmatrix},$$

in cui $\mathbf{c}, \mathbf{d}, \mathbf{v} \in \mathbf{C}^{n-1}$ e $B \in \mathbf{C}^{(n-1)\times(n-1)}$ è triangolare superiore con elementi principali uguali agli elementi sottodiagonali di A. Dalla relazione $(A - \lambda I)\mathbf{z} = \mathbf{0}$ segue che $B\mathbf{v} = \mathbf{0}$, in cui $\mathbf{v} \neq \mathbf{0}$. Perciò uno almeno degli elementi principali di B deve essere nullo.)

6.28 Sia B_n la matrice tridiagonale hermitiana dell'esercizio 6.26. Si dimostri che se B_n ha un autovalore λ di molteplicità algebrica m, allora almeno $m - 1$ elementi sottodiagonali β_i di B_n sono nulli.

(Traccia: per l'esercizio 6.27 esiste un indice j, $2 \leq j \leq n$, tale che $\beta_j = 0$. Si consideri la sottomatrice di B_n ottenuta cancellando la j-esima riga e la j-esima colonna. Per l'esercizio 6.13 tale matrice ha l'autovalore λ di molteplicità almeno $m - 1$.)

6.29 Siano A e $D \in \mathbf{C}^{n\times n}$ e sia $\{Q_k\}$ una successione di matrici unitarie tali che

$$\lim_{k\to\infty} Q_k^H A Q_k = D.$$

a) Si dimostri che esiste una sottosuccessione $\{Q_{k_i}\}$ tale che

$$\lim_{i\to\infty} Q_{k_i} = Q \quad \text{e} \quad Q^H A Q = D;$$

b) si trovi una successione $\{Q_k\}$ di matrici non unitarie tali che

$$\lim_{k\to\infty} Q_k^{-1} A Q_k = D,$$

ma il $\lim_{i\to\infty} Q_{k_i}$ non esiste per nessuna sottosuccessione $\{Q_{k_i}\}$.

(Traccia: a) è sufficiente dimostrare che l'insieme delle matrici U unitarie è un compatto; ciò segue dal fatto che tale insieme è limitato essendo $\|U\|_2 = 1$ per ogni U unitaria, e chiuso, essendo l'insieme degli zeri della funzione continua $U \to \|U^H U - I\|$; b)

$$Q_k = \begin{bmatrix} 1 & 0 \\ k & 1 \end{bmatrix}, \quad A = I.)$$

6.30 Sia $\{A_k\}$ una successione di matrici tali che

$$\lim_{k\to\infty} A_k = I.$$

Indicata con $A_k = Q_k R_k$ una fattorizzazione QR della matrice A_k, esistono matrici S_k diagonali e unitarie (matrici di fase) tali che

$$\lim_{k\to\infty} Q_k S_k = \lim_{k\to\infty} S_k Q_k = I, \quad \lim_{k\to\infty} S_k^H R_k = \lim_{k\to\infty} R_k S_k^H = I.$$

(Traccia: vale $\lim_{k\to\infty}(Q_k R_k - I) = O$ e quindi, poiché gli elementi di Q_k hanno modulo limitato, si ha

$$\lim_{k\to\infty}(R_k - Q_k^H) = \lim_{k\to\infty} Q_k^H (Q_k R_k - I) = O.$$

Poiché R_k è triangolare superiore, segue che

$$\lim_{k\to\infty} \overline{q}_{ji}^{(k)} = 0, \quad \text{per } i > j.$$

Da cui, poiché Q_k è unitaria, si ottiene

$$\lim_{k\to\infty} \overline{q}_{ji}^{(k)} = 0, \quad \text{per } i < j.$$

Quindi $q_{ii}^{(k)} \neq 0$ da un certo indice k in poi, e posto

$$S_k = \begin{bmatrix} \theta_1^{(k)} & & & \\ & \theta_2^{(k)} & & \\ & & \ddots & \\ & & & \theta_n^{(k)} \end{bmatrix}, \quad \theta_i^{(k)} = \frac{\overline{q}_{ii}^{(k)}}{|q_{ii}^{(k)}|},$$

si dimostri che

$$\lim_{k\to\infty} S_k Q_k^H = \lim_{k\to\infty} Q_k^H S_k = I.)$$

6.31 Sia $A_k \in \mathbf{C}^{n\times n}$ la matrice tridiagonale hermitiana

$$A_k = \begin{bmatrix} \alpha_1 & \overline{\beta}_2 & & \\ \beta_2 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \overline{\beta}_n \\ & & \beta_n & \alpha_n \end{bmatrix},$$

in cui $\beta_i \neq 0$, per $i = 2, \ldots, n$. Si applichi ad A_k un passo del metodo QR con lo shift $\mu_k = \alpha_n$, ottenendo la matrice A_{k+1} nel modo seguente

$$A_k - \alpha_n I = Q_k R_k, \qquad A_{k+1} = R_k Q_k + \alpha_n I,$$

dove Q_k è unitaria ed R_k è triangolare superiore. Posto

$$A_{k+1} = \begin{bmatrix} \alpha'_1 & \overline{\beta}_2\,' & & \\ \beta'_2 & \alpha'_2 & \ddots & \\ & \ddots & \ddots & \overline{\beta}_n\,' \\ & & \beta'_n & \alpha'_n \end{bmatrix}.$$

si dimostri che

$$|\beta'_n| \leq \frac{|\beta_n|^3}{d^2}, \qquad |\alpha'_n - \alpha_n| \leq \frac{|\beta_n|^2}{d}, \quad d = \min_{i=1,\ldots,n-1} |\lambda_i - \alpha_n|,$$

dove $\lambda_1, \ldots, \lambda_{n-1}$, sono gli autovalori della sottomatrice principale di testa di ordine $n-1$ di A_k. Quindi, poiché A ha autovalori distinti, nel metodo QR con lo shift $\mu_k = a_{nn}^{(k)}$ la convergenza a zero dell'elemento sottodiagonale è del terzo ordine.

(Traccia: posto

$$A_k - \alpha_n I = \begin{bmatrix} B & \mathbf{g} \\ \mathbf{g}^H & 0 \end{bmatrix}, \quad \mathbf{g} = \overline{\beta}_n \mathbf{e}_{n-1},$$

siano $Q \in \mathbf{C}^{(n-1)\times(n-1)}$ una matrice unitaria tale che

$$B = QR,$$

dove R è triangolare superiore, e $G \in \mathbf{C}^{2\times 2}$ la matrice di Givens

$$G = \begin{bmatrix} c & \text{-}\overline{s} \\ s & c \end{bmatrix},$$

dove

$$c = \frac{r_{n-1,n-1}}{\sqrt{|r_{n-1,n-1}|^2 + |\beta_n|^2}}, \quad s = -\frac{\beta_n}{\sqrt{|r_{n-1,n-1}|^2 + |\beta_n|^2}}.$$

Si ha

$$A_k - \alpha_n I = \begin{bmatrix} Q & \mathbf{0} \\ \mathbf{0}^T & 1 \end{bmatrix} \begin{bmatrix} I_{n-2} & O \\ O & G^H \end{bmatrix} R_k,$$

dove R_k è triangolare superiore e risulta

$$r_{nn}^{(k)} = s\overline{\beta}_n \mathbf{e}_{n-1}^T Q^H \mathbf{e}_{n-1}.$$

Inoltre

$$A_{k+1} = R_k \begin{bmatrix} Q & \mathbf{0} \\ \mathbf{0}^T & 1 \end{bmatrix} \begin{bmatrix} I_{n-2} & O \\ O & G^H \end{bmatrix} + \alpha_n I = \begin{bmatrix} B' & \mathbf{h} \\ \mathbf{h}^H & \alpha'_n \end{bmatrix},$$

dove

$$\mathbf{h} = -\overline{s}\,\overline{r}^{(k)}_{nn}\mathbf{e}_{n-1}, \qquad \alpha'_n = cr^{(k)}_{nn} + \alpha_n.$$

Sostituendo, passando ai moduli e tenendo conto che gli elementi principali di Q sono in modulo minori di 1, si ha

$$|\beta'_n| \leq \frac{|\beta_n|^3}{|r_{n-1,n-1}|^2 + |\beta_n|^2} \leq \frac{|\beta_n|^3}{|r_{n-1,n-1}|^2}$$

e

$$|\alpha'_n - \alpha_n| \leq \frac{|\beta_n|^2 |r_{n-1,n-1}|}{|r_{n-1,n-1}|^2 + |\beta_n|^2} \leq \frac{|\beta_n|^2}{|r_{n-1,n-1}|}\ .$$

Si dimostri poi che $|r_{n-1,n-1}| \geq d$ (per l'esercizio 6.19).)

6.32 Si applichi il metodo delle potenze alla seguente matrice di *Bodewig*

$$A = \begin{bmatrix} 2 & 1 & 3 & 4 \\ 1 & -3 & 1 & 5 \\ 3 & 1 & 6 & -2 \\ 4 & 5 & -2 & -1 \end{bmatrix},$$

assumendo come vettore iniziale $\mathbf{t}_0 = [1,\ 1,\ 1,\ 1]^T$. La successione converge molto lentamente. Perché?

(Risposta: è $|\lambda_1/\lambda_2| \approx 0.998$, inoltre il vettore $\mathbf{t}_0$ è "quasi" ortogonale all'autovettore $\mathbf{x}_1$.)

6.33 Si applichi il metodo delle potenze alla matrice

$$A = \begin{bmatrix} 0 & -1 & 1 \\ 1 & 0 & -1 \\ -1 & 1 & 0 \end{bmatrix},$$

assumendo come vettore iniziale $\mathbf{t}_0 = [1,\ 1,\ -2]^T$.

(Risposta: si ottiene

$$\begin{aligned} \beta_1 &= -3, \quad \mathbf{t}_1 = [1, -1, 0]^T, \\ \beta_2 &= -2, \quad \mathbf{t}_2 = [-0.5, -0.5, 1]^T, \\ \beta_3 &= 1.5, \quad \mathbf{t}_3 = [1, -1, 0]^T, \\ \beta_{i+2} &= \beta_i, \quad \mathbf{t}_{i+2} = \mathbf{t}_i, \qquad \text{per } i \geq 2. \ ) \end{aligned}$$

6.34 Sia $A \in \mathbf{R}^{n\times n}$, tale che l'autovalore λ_1 di modulo massimo sia complesso.

a) Si esamini il comportamento della successione $\mathbf{y}_k$ ottenuta con il metodo delle potenze;

b) si dica come è possibile ricavare un'approssimazione di λ_1 da tale successione e come calcolare l'autovettore corrispondente.

(Traccia: a) siano $|\lambda_1| = |\overline{\lambda}_1| > |\lambda_3| \geq \ldots \geq |\lambda_n|$ gli autovalori di A e $\mathbf{x}_1,\ \overline{\mathbf{x}}_1,\ \mathbf{x}_3, \ldots,\ \mathbf{x}_n$ i corrispondenti autovettori. Allora $\mathbf{t}_0 \in \mathbf{R}^n$ può essere espresso come

$$\mathbf{t}_0 = \alpha_1\mathbf{x}_1 + \overline{\alpha}_1\overline{\mathbf{x}}_1 + \sum_{i=3}^{n} \alpha_i\mathbf{x}_i$$

e si ha

$$A^k\mathbf{t}_0 = |\lambda_1|^k \left[\alpha_1 \left(\frac{\lambda_1}{|\lambda_1|}\right)^k \mathbf{x}_1 + \overline{\alpha}_1 \left(\frac{\overline{\lambda}_1}{|\lambda_1|}\right)^k \overline{\mathbf{x}}_1 + \sum_{i=3}^{n} \alpha_i \left(\frac{\lambda_i}{|\lambda_i|}\right)^k \mathbf{x}_i\right],$$

da cui

$$A^k\mathbf{t}_0 = |\lambda_1|^k \left[\alpha_1 \left(\frac{\lambda_1}{|\lambda_1|}\right)^k \mathbf{x}_1 + \overline{\alpha}_1 \left(\frac{\overline{\lambda}_1}{|\lambda_1|}\right)^k \overline{\mathbf{x}}_1\right] + O\left(\left|\frac{\lambda_2}{\lambda_3}\right|^k\right).$$

Posto $\lambda_1 = |\lambda_1|(\cos\phi + \mathbf{i}\sin\phi)$, $\alpha_1 = |\alpha_1|(\cos\theta + \mathbf{i}\sin\theta)$ e indicando la j-esima componente di $\mathbf{x}_1$ con $|x_j|(\cos\xi_j + \mathbf{i}\sin\xi_j)$, si ha

$$y_j^{(k)} = (A^k\mathbf{t}_0)_j = 2|\lambda_1|^k\,|\alpha_1|\,|x_j|\cos(k\phi + \theta + \xi_j) + O\left(\left|\frac{\lambda_2}{\lambda_3}\right|^k\right);$$

quindi, a meno di termini dell'ordine di $|\lambda_2/\lambda_3|^k$, le componenti del vettore $\mathbf{y}_k$ ruotano ad ogni passo di un angolo ϕ attorno all'origine del piano complesso; b) sia $\lambda^2 + p\lambda + q = 0$ l'equazione che ha come soluzioni λ_1 e $\overline{\lambda}_1$, allora per k sufficientemente grande si ha

$$\begin{aligned}\mathbf{y}_{k+2} + p\mathbf{y}_{k+1} + q\mathbf{y}_k &= (A^{k+2} + pA^{k+1} + qA^k)\mathbf{t}_0 \\ &\approx \alpha_1(\lambda_1^{k+2} + p\lambda_1^{k+1} + q\lambda_1^k)\mathbf{x}_1 + \overline{\alpha}_1(\overline{\lambda}_1^{\,k+2} + p\overline{\lambda}_1^{\,k+1} + q\overline{\lambda}_1^{\,k})\overline{\mathbf{x}}_1 = \mathbf{0}.\end{aligned}$$

I due coefficienti p e q possono essere approssimati con le successioni p_k, q_k tali che $\mathbf{y}_{k+2} + p_k\mathbf{y}_{k+1} + q_k\mathbf{y}_k = \mathbf{0}$. Poiché in generale questi sistemi di n equazioni nelle due incognite p_k e q_k non sono risolubili, si determinano p_k e q_k con il metodo dei minimi quadrati (si veda il capitolo 7)

$$\min_{[\alpha,\beta]^T\in\mathbf{R}^2} \|\mathbf{y}_{k+2} + \alpha\mathbf{y}_{k+1} + \beta\mathbf{y}_k\|_2,$$

ricavando $[\, p_k \,,\, q_k \,]^T$ come soluzione del sistema normale

$$\begin{bmatrix} \mathbf{y}_{k+1}^T\mathbf{y}_{k+1} & \mathbf{y}_{k+1}^T\mathbf{y}_k \\ \\ \mathbf{y}_k^T\mathbf{y}_{k+1} & \mathbf{y}_k^T\mathbf{y}_k \end{bmatrix} \begin{bmatrix} \alpha \\ \\ \beta \end{bmatrix} = - \begin{bmatrix} \mathbf{y}_{k+1}^T\mathbf{y}_{k+2} \\ \\ \mathbf{y}_k^T\mathbf{y}_{k+2} \end{bmatrix}.$$

Per calcolare l'autovettore corrispondente, per k sufficientemente grande si ha

$$\mathbf{y}_k \approx \gamma\mathbf{x}_1 + \overline{\gamma}\,\overline{\mathbf{x}}_1 \quad \text{e} \quad \mathbf{y}_{k+1} = A\mathbf{y}_k \approx \gamma\lambda_1\mathbf{x}_1 + \overline{\gamma}\overline{\lambda}_1\overline{\mathbf{x}}_1,$$

per cui se $\gamma\mathbf{x}_1 = \mathbf{z} + \mathbf{i}\mathbf{w}$ e $\lambda_1 = \mu + \mathbf{i}\nu$, si ha

$$\mathbf{y}_k \approx 2\mathbf{z} \quad \text{e} \quad \mathbf{y}_{k+1} \approx 2(\mu\mathbf{z} - \nu\mathbf{w}),$$

da cui, a meno di un fattore di normalizzazione, si ricava

$$\mathbf{x}_1 \approx \nu\mathbf{y}_k + \mathbf{i}(\mu\mathbf{y}_k - \mathbf{y}_{k+1}).\ )$$

6.35 Sia $A \in \mathbf{C}^{n\times n}$ una matrice con due autovalori reali λ_1 e λ_2 di modulo massimo, tali che $\lambda_1 = -\lambda_2$ e $|\lambda_i| < |\lambda_1|$ per $i > 2$.

a) Si esamini il comportamento della successione $\mathbf{y}_k$ ottenuta con il metodo delle potenze;

b) si dica che cosa accade quando si applica il metodo delle potenze alla matrice traslata $A - \alpha I$, $\alpha \in \mathbf{R}$, $\alpha \neq 0$;

c) si studi in particolare il caso della matrice dell'esercizio 6.33.

(Traccia: a) per k pari sufficientemente grande si ha

$$y_j^{(k)} \approx \lambda_1^k(\alpha_1 x_j^{(1)} + \alpha_2 x_j^{(2)}),$$

$$y_j^{(k+1)} \approx \lambda_1^{k+1}(\alpha_1 x_j^{(1)} - \alpha_2 x_j^{(2)}),$$

$$y_j^{(k+2)} \approx \lambda_1^{k+2}(\alpha_1 x_j^{(1)} + \alpha_2 x_j^{(2)}),$$

per cui

$$\lim_{k\to\infty} \frac{y_j^{(k+2)}}{y_j^{(k)}} = \lambda_1^2;$$

b) la matrice $A - \alpha I$ non ha più due autovalori di modulo massimo, ma uno solo il cui modulo è $|\lambda_1| + |\alpha|$. Scegliendo opportunamente α si può ottenere una buona velocità di convergenza.)

6.36 Si analizzi la convergenza del metodo delle potenze nel caso di matrici con autovalori di molteplicità geometrica minore della molteplicità algebrica.

(Traccia: Si supponga dapprima che A sia un unico blocco di Jordan della forma

$$A = \begin{bmatrix} \lambda & 1 & & \\ & \lambda & \ddots & \\ & & \ddots & 1 \\ & & & \lambda \end{bmatrix},$$

e si consideri un vettore

$$\mathbf{t}_0 = \sum_{j=1}^{n} \alpha_j \mathbf{e}_j.$$

Si ha

$$A^k = \begin{bmatrix} \lambda^k & \binom{k}{1}\lambda^{k-1} & \cdots & \binom{k}{n-1}\lambda^{k-n+1} \\ & \lambda^k & \ddots & \vdots \\ & & \ddots & \binom{k}{1}\lambda^{k-1} \\ & & & \lambda^k \end{bmatrix},$$

per cui

$$A^k\mathbf{e}_j = \binom{k}{j-1}\lambda^{k-j+1}\mathbf{e}_1 + \binom{k}{j-2}\lambda^{k-j+2}\mathbf{e}_2 + \ldots + \lambda^k\mathbf{e}_j$$

$$= \binom{k}{j-1}\lambda^{k-j+1}\Big[\mathbf{e}_1 + \frac{j-1}{k-j+2}\lambda\mathbf{e}_2 + \ldots + \frac{(j-1)!}{k(k-1)\ldots(k-j+2)}\lambda^{j-1}\mathbf{e}_j\Big].$$

Per $k \to \infty$ tutti i termini fra parentesi, escluso il primo, tendono a zero, per cui, posto

$$\theta_k = \sum_{j=1}^{n} \alpha_j \binom{k}{j-1}\lambda^{k-j+1},$$

risulta

$$\lim_{k\to\infty} \frac{A^k\mathbf{t}_0}{\theta_k} = \lim_{k\to\infty} \sum_{j=1}^{n} \frac{\alpha_j}{\theta_k} A^k\mathbf{e}_j = \left[\lim_{k\to\infty} \sum_{j=1}^{n} \frac{\alpha_j}{\theta_k}\binom{k}{j-1}\lambda^{k-j+1}\right]\mathbf{e}_1 = \mathbf{e}_1,$$

e quindi si ha convergenza verso l'autovettore $\mathbf{e}_1$. Per k abbastanza grande si ha

$$(A^k\mathbf{t}_0)_1 \approx \sum_{j=1}^{n} \alpha_j \binom{k}{j-1} \lambda^{k-j+1} = \lambda^k p_{n-1}(k),$$

dove $p_{n-1}(k)$ è un polinomio in k di grado $n-1$, per cui

$$\lim_{k\to\infty} \frac{(A^{k+1}\mathbf{t}_0)_1}{(A^k\mathbf{t}_0)_1} = \lambda.$$

La convergenza è però molto lenta. Si ripeta poi il ragionamento per il caso in cui la matrice sia formata da più blocchi di Jordan.)

6.37 Si analizzi la convergenza del metodo delle potenze inverse nel caso di matrici con autovalori di molteplicità geometrica minore della molteplicità algebrica.

(Traccia: Si supponga dapprima che A sia un unico blocco di Jordan della forma

$$A = \begin{bmatrix} \lambda & 1 & & \\ & \lambda & \ddots & \\ & & \ddots & 1 \\ & & & \lambda \end{bmatrix},$$

si consideri un vettore

$$\mathbf{t}_0 = \sum_{j=1}^{n} \alpha_j \mathbf{e}_j$$

e sia μ un'approssimazione di λ. Indicato con $\mathbf{f}_j$, $j = 1, \ldots, n$, il vettore soluzione del sistema

$$(A - \mu I)\mathbf{f}_j = \mathbf{e}_j,$$

si ha

$$\mathbf{t}_1 = \sum_{j=1}^{n} \alpha_j \mathbf{f}_j,$$

dove

$$\mathbf{f}_j = \frac{(-1)^{j-1}}{(\lambda-\mu)^j} \left[1, -(\lambda-\mu), (\lambda-\mu)^2, \ldots, (-1)^{j-1}(\lambda-\mu)^{j-1}, 0, \ldots, 0\right]^T.$$

Se μ è una buona approssimazione di λ, si ha $\mathbf{t}_1 = \beta\mathbf{e}_1 + O(\lambda-\mu)$, dove

$$\beta = \sum_{j=1}^{n} (-1)^{j-1} \frac{\alpha_j}{(\lambda-\mu)^j} = (-1)^{n-1} \frac{\alpha_n}{(\lambda-\mu)^n} \left[1 + O(\lambda-\mu)\right].$$

Quindi con una sola iterazione si ottiene un vettore $\mathbf{t}_1$ che apprrossima l'autovettore $\beta\mathbf{e}_1$ con un errore dell'ordine di $\lambda-\mu$. Si ripeta il ragionamento per il caso in cui la matrice è formata da più blocchi di Jordan.)

6.38 Sia $A \in \mathbf{C}^{n\times n}$, $\mu \in \mathbf{C}$ distinto da ogni autovalore di A, e sia $\mathbf{t}_0 \in \mathbf{C}^n$ e $\mathbf{t}_1 = (A-\mu I)^{-1}\mathbf{t}_0$.

a) Si dimostri che $\mathbf{t}_1$ è autovettore, corrispondente all'autovalore μ, di una matrice $A+E$, dove $\quad \|E\|_2 = \dfrac{\|\mathbf{t}_0\|_2}{\|\mathbf{t}_1\|_1}$;

b) da questa relazione si derivi un controllo utile in fase di implementazione del metodo delle potenze inverse.

(Traccia: a) siano U_0 e U_1 le matrici di Householder tali che $U_0\mathbf{t}_0 = \alpha_0\mathbf{e}_1$ e $U_1\mathbf{t}_1 = \alpha_1\mathbf{e}_1$, dove $|\alpha_0| = \|\mathbf{t}_0\|_2$ e $|\alpha_1| = \|\mathbf{t}_1\|_2$. Allora $-\mathbf{t}_0 = E\mathbf{t}_1$, dove

$$E = -\frac{\alpha_0}{\alpha_1}\, U_0^{-1}U_1,$$

e quindi

$$(A+E)\mathbf{t}_1 = \mathbf{t}_0 + \mu\mathbf{t}_1 + E\mathbf{t}_1 = \mu\mathbf{t}_1 \quad \text{e} \quad \|E\|_2 = \frac{|\alpha_0|}{|\alpha_1|};$$

b) $\|E\|_2$ è tanto minore quanto più grande è $\|\mathbf{t}_1\|_2$ rispetto a $\|\mathbf{t}_0\|_2$, per cui se $\|\mathbf{t}_1\|_2$ è piccolo, è opportuno cambiare vettore iniziale, scegliendone uno, magari ortogonale al precedente.)

6.39 Sia $A \in \mathbf{C}^{n\times n}$ hermitiana e si supponga che A abbia autovalori distinti. Si dimostri che la successione dei μ_k calcolati con il metodo delle iterazioni del quoziente di Rayleigh ha convergenza locale del terzo ordine ad un autovalore λ di A.

(Traccia: sia $(A-\mu_{k-1}I)\mathbf{u}_k = \mathbf{t}_{k-1}$, con $\|\mathbf{t}_{k-1}\|_2 = 1$ e $\mu_{k-1} = \mathbf{t}_{k-1}^H A\mathbf{t}_{k-1}$, $\mathbf{t}_k = \dfrac{\mathbf{u}_k}{\|\mathbf{u}_k\|_2}$. Si può supporre, senza ledere la generalità, che A sia diagonale con elementi principali $d_1, d_2, \ldots, d_n$, e che $\mu_k \to d_1$. Posto $e_k = |\mu_k - d_1|$, vale

$$\begin{aligned} e_k = |\mathbf{t}_k^H(A-d_1I)\mathbf{t}_k| &\le \sum_{i=2}^{n} |t_i^{(k)}|^2\, |d_i - d_1| \\ &\le \max_{i=2,\ldots,n} |d_i-d_1| \sum_{i=2}^{n} |t_i^{(k)}|^2 = \phi \sum_{i=2}^{n} |t_i^{(k)}|^2. \end{aligned}$$

Poiché $u_i^{(k)} = \dfrac{t_i^{(k-1)}}{d_i - \mu_{k-1}}$, è $\|\mathbf{u}_k\|_2^2 \geq \dfrac{|t_1^{(k-1)}|^2}{e_{k-1}^2}$, ed essendo $t_i^{(k)} = \dfrac{u_i^{(k)}}{\|\mathbf{u}_k\|_2}$, ne segue che esiste una costante γ tale che

$$\sum_{i=2}^{n} |t_i^{(k)}|^2 \leq \gamma e_{k-1}^2 \sum_{i=2}^{n} |t_i^{(k-1)}|^2,$$

per cui

$$\sum_{i=2}^{n} |t_i^{(k)}|^2 \leq \gamma^k (e_{k-1} e_{k-2} \dots e_0)^2,$$

e quindi

$$e_k \leq \phi \gamma^k (e_{k-1} e_{k-2} \dots e_0)^2 \leq \gamma^{(3^k-1)/2} \phi^{3^{k-1}} e_0^{2 \times 3^{k-1}}, \quad k \geq 1$$

cioè il metodo ha convergenza cubica.)

6.40 Siano $A \in \mathbf{C}^{n \times n}$ hermitiana con autovalori distinti e $\mathbf{v} \in \mathbf{C}^n$. Si consideri il seguente metodo iterativo per il calcolo di un autovalore della matrice A:

$$\begin{cases} (A - \mu_{k-1} I)\mathbf{u}_k = \mathbf{v} \\ \\ \mu_k = \dfrac{\mathbf{u}_k^H A \mathbf{u}_k}{\mathbf{u}_k^H \mathbf{u}_k} \qquad\qquad k = 1, 2, \dots \end{cases}$$

a) Se $\lim\limits_{k \to \infty} \mu_k = \lambda$, si dimostri che esiste $\phi \in \mathbf{R}$ tale che, posto $e_k = |\lambda - \mu_k|$, è

$$e_k \leq \phi e_{k-1}^2,$$

e quindi la convergenza è localmente del secondo ordine;

b) si dica sotto quali condizioni sul vettore $\mathbf{v}$ e su μ_0 la successione generata dal metodo converge ad un autovalore di A.

(Traccia: si può supporre, senza ledere la generalità, che A sia diagonale con elementi principali $d_1, d_2, \dots, d_n$, e che $\lambda = d_1$. Allora

$$\lambda - \mu_k = \frac{\mathbf{u}_k^H (\lambda I - A)\mathbf{u}_k}{\mathbf{u}_k^H \mathbf{u}_k} = \frac{\sum\limits_{i=2}^{n} |u_i^{(k)}|^2 (\lambda - d_i)}{\sum\limits_{i=1}^{n} |u_i^{(k)}|^2} ,$$

da cui

$$e_k \leq \theta \, \frac{\sum\limits_{i=2}^{n} |u_i^{(k)}|^2}{\sum\limits_{i=1}^{n} |u_i^{(k)}|^2} , \quad \theta > 0.$$

Utilizzando la relazione $\mathbf{u}_k = (A - \mu_{k-1}I)^{-1}\mathbf{v}$, si dimostri che esiste una costante $\psi > 0$ tale che

$$\frac{\sum_{i=2}^{n} |u_i^{(k)}|^2}{\sum_{i=1}^{n} |u_i^{(k)}|^2} \leq \psi e_{k-1}^2,$$

e quindi, posto $\phi = \psi\theta$, ne segue che

$$e_k \leq \phi e_{k-1}^2, \quad \text{da cui} \quad e_k \leq \phi^{2^k-1} e_0^{2^k}.$$

Una condizione sufficiente di convergenza si ottiene imponendo che $\phi e_0 < 1$.)

6.41 Siano V e $W \in \mathbf{C}^{n\times n}$ due matrici unitarie, sia $1 \leq p < n$ e si considerino le decomposizioni

$$V = [\, V_1 \mid V_2 \,] \quad \text{e} \quad W = [\, W_1 \mid W_2 \,],$$

in cui $V_1,\ W_1 \in \mathbf{C}^{n\times p}$, $V_2,\ W_2 \in \mathbf{C}^{n\times(n-p)}$. La quantità

$$d = \|V_1 V_1^H - W_1 W_1^H\|_2$$

è detta *distanza* fra il sottospazio di $\mathbf{C}^n$ generato dalle colonne di V_1 e il sottospazio generato dalle colonne di W_1.

a) Si dimostri che

$$d^2 = \rho(W_1^H V_2 V_2^H W_1) = \|V_2^H W_1\|_2^2$$

(per la norma 2 delle matrici non quadrate, si veda il paragrafo 3 del capitolo 7);

b) se $d < 1$, si dimostri che la matrice $V_1^H W_1$ è non singolare e che

$$\|V_1^H W_1\|_2 \leq 1 \quad \text{e} \quad \|(V_1^H W_1)^{-1}\|_2 = \frac{1}{\sqrt{1-d^2}};$$

c) si sfruttino queste relazioni per dimostrare il teorema 6.42.

(Traccia: a) si ha

$$\|V_1 V_1^H - W_1 W_1^H\|_2 = \|V^H (V_1 V_1^H - W_1 W_1^H) W\|_2$$

e

$$S = V^H (V_1 V_1^H - W_1 W_1^H) W = \begin{bmatrix} V_1^H \\ \\ V_2^H \end{bmatrix} (V_1 V_1^H - W_1 W_1^H)[\, W_1 \mid W_2 \,]$$

$$= \begin{bmatrix} V_1^H - V_1^H W_1 W_1^H \\ \\ -V_2^H W_1 W_1^H \end{bmatrix} [\, W_1 \mid W_2 \,] = \begin{bmatrix} O & V_1^H W_2 \\ \\ -V_2^H W_1 & O \end{bmatrix}.$$

Inoltre, poiché $V_2 V_2^H = I_n - V_1 V_1^H$ e $W_2 W_2^H = I_n - W_1 W_1^H$, è

$$W_1^H V_2 V_2^H W_1 = I_p - W_1^H V_1 V_1^H W_1 \text{ e } V_1^H W_2 W_2^H V_1 = I_p - V_1^H W_1 W_1^H V_1,$$

e poiché le matrici

$$W_2^H V_1 V_1^H W_2 \quad \text{e} \quad V_1^H W_2 W_2^H V_1$$

hanno gli stessi autovalori (eccetto eventualmente l'autovalore nullo), e le matrici

$$V_1^H W_1 W_1^H V_1 \quad \text{e} \quad W_1^H V_1 V_1^H W_1$$

hanno gli stessi autovalori, ne segue che le due matrici

$$W_2^H V_1 V_1^H W_2 \quad \text{e} \quad W_1^H V_2 V_2^H W_1$$

hanno lo stesso raggio spettrale, e quindi $\rho(S^H S) = \rho(W_1^H V_2 V_2^H W_1)$;
b) poiché $W_1^H V_1 V_1^H W_1 = I_p - W_1^H V_2 V_2^H W_1$ e $d < 1$, gli autovalori di $W_1^H V_1 V_1^H W_1$ sono tutti positivi e minori o uguali a 1 e il minimo autovalore è $1 - d^2$; ne segue che $V_1^H W_1$ è non singolare e che

$$\rho[(W_1^H V_1 V_1^H W_1)^{-1}] = \frac{1}{1 - d^2}\ ;$$

c) sia $A = XTX^H$, dove $X \in \mathbf{C}^{n \times n}$ unitaria e $T \in \mathbf{C}^{n \times n}$ diagonale,

$$X = [\, U \mid Z \,] \quad \text{e} \quad T = \begin{bmatrix} T_1 & O \\ O & T_2 \end{bmatrix}, \quad U \in \mathbf{C}^{n \times p},\ T_1 \in \mathbf{C}^{p \times p},$$

in cui gli elementi principali di T_1 sono $\lambda_1,\ \lambda_2,\ \ldots,\ \lambda_p$ e gli elementi principali di T_2 sono $\lambda_{p+1},\ \ldots,\ \lambda_n$. Inoltre sia

$$H_k = [\, Q_k \mid P_k \,] \quad \text{e} \quad R_k = \begin{bmatrix} S_k \\ O \end{bmatrix},$$

dove $S_k \in \mathbf{C}^{p \times p}$ è triangolare superiore. È

$$A^k Q_0 = Q_k S_k \ldots S_1, \quad \text{per cui} \quad T^k X^H Q_0 = X^H Q_k S_k \ldots S_1,$$

e quindi

$$T_1^k U^H Q_0 = U^H Q_k S_k \ldots S_1 \quad \text{e} \quad T_2^k Z^H Q_0 = Z^H Q_k S_k \ldots S_1,$$

da cui, se $U^H Q_0$ è non singolare e $\lambda_p \neq 0$, si ricava

$$Z^H Q_k = T_2^k Z^H Q_0 (U^H Q_0)^{-1} T_1^{-k} U^H Q_k.$$

Per i punti a) e b), in cui si ponga $V_1 = U$, $V_2 = Z$, $W_1 = Q_0$ oppure $W_1 = Q_k$, si ottiene che la matrice $U^H Q_0$ è non singolare e che

$$d_0 = \|UU^H - Q_0 Q_0^H\|_2 = \|Z^H Q_0\|_2 \text{ e } d_k = \|UU^H - Q_k Q_k^H\|_2 = \|Z^H Q_k\|_2,$$

per cui passando alle norme si ha

$$d_k = \|Z^H Q_k\|_2 \leq \|T_2\|_2^k \, \|Z^H Q_0\|_2 \, \|(U^H Q_0)^{-1}\|_2 \, \|T_1^{-1}\|_2^k \, \|U^H Q_k\|_2,$$

e poiché

$$\|(U^H Q_0)^{-1}\|_2 = \frac{1}{\sqrt{1 - d_0^2}}, \ \|U^H Q_k\|_2 \leq 1, \ \|T_2\|_2^k = |\lambda_{p+1}|, \ \|T_1^{-1}\|_2 = \frac{1}{|\lambda_p|},$$

si ha

$$d_k \leq \frac{d_0}{\sqrt{1 - d_0^2}} \left| \frac{\lambda_{p+1}}{\lambda_p} \right|^k.$$

Per dimostrare la (56), si consideri la i-esima colonna $\mathbf{q}_i^{(k)}$ della matrice Q_k poiché $Q_k A Q_{k-1} = S_k$, vale

$$r_{ii}^{(k)} = {\mathbf{q}_i^{(k)}}^H A \mathbf{q}_i^{(k-1)},$$

per cui se $\mathbf{q}_i^{(k)} = \theta_i \mathbf{x}_i + \mathbf{e}_i^{(k)}$, dove θ_i è un opportuno fattore di fase, si ha $\lim_{k \to \infty} \mathbf{e}_i^{(k)} = \mathbf{0}$, e

$$r_{ii}^{(k)} = (\theta_i \mathbf{x}_i + \mathbf{e}_i^{(k)})^H A (\theta_i \mathbf{x}_i + \mathbf{e}_i^{(k-1)}) = \lambda_i + \lambda_i \theta_i {\mathbf{e}_i^{(k)}}^H \mathbf{x}_i + \overline{\lambda}_i \overline{\theta}_i \mathbf{x}_i^H \mathbf{e}_i^{(k-1)} + \mathbf{e}_i^{(k)})^H A \mathbf{e}_i^{(k-1)}.$$

Basta dunque dimostrare che

$$\|\mathbf{e}_i^{(k)}\|_2 = O\Big(\Big| \frac{\lambda_{i+1}}{\lambda_i} \Big|^k + \Big| \frac{\lambda_i}{\lambda_{i-1}} \Big|^k \Big).$$

Per questo si consideri la matrice

$$E_i = U^{(i)} {U^{(i)}}^H - Q_k^{(i)} {Q_k^{(i)}}^H,$$

dove $U^{(i)}$ e $Q_k^{(i)}$ sono matrici costituite dalle prime i colonne rispettivamente di U e di Q_k. Si osservi che qualunque sia $p \geq i$, la matrice $Q_k^{(i)}$ è sempre la stessa, e quindi

$$\|E_i\|_2 = O\Big(\Big|\frac{\lambda_{i+1}}{\lambda_i}\Big|^k\Big).$$

Si ha allora

$$E_i = E_{i-1} + \mathbf{x}_i\mathbf{x}_i^H - \mathbf{q}_i^{(k)}\mathbf{q}_i^{(k)\,H},$$

da cui

$$\|\mathbf{x}_i\mathbf{x}_i^H - \mathbf{q}_i^{(k)}\mathbf{q}_i^{(k)\,H}\|_2 \leq \|E_i\|_2 + \|E_{i-1}\|_2 = O\Big(\Big|\frac{\lambda_{i+1}}{\lambda_i}\Big|^k + \Big|\frac{\lambda_i}{\lambda_{i-1}}\Big|^k\Big).$$

Sia θ_i un fattore di fase tale che $\theta_i\mathbf{x}_i^H\mathbf{q}_i^{(k)}$ sia reale e non negativo, allora

$$\|\theta_i\mathbf{x}_i - \mathbf{q}_i^{(k)}\|_2^2 = 2(1 - |\mathbf{x}_i^H\mathbf{q}_i^{(k)}|).$$

Gli autovalori della matrice $\mathbf{x}_i\mathbf{x}_i^H - \mathbf{q}_i^{(k)}\mathbf{q}_i^{(k)\,H}$ sono uguali, a parte lo zero, a quelli della matrice

$$\begin{bmatrix} 1 & \mathbf{x}_i^H\mathbf{q}_i^{(k)} \\ -\mathbf{q}_i^{(k)\,H}\mathbf{x}_i & -1 \end{bmatrix},$$

(si veda l'esercizio 2.24) e quindi

$$\|\mathbf{x}_i\mathbf{x}_i^H - \mathbf{q}_i^{(k)}\mathbf{q}_i^{(k)\,H}\|_2 = 1 - |\mathbf{x}_i^H\mathbf{q}_i^{(k)}|^2 = \frac{1}{2}\,(1 + |\mathbf{x}_i^H\mathbf{q}_i^{(k)}|)\|\theta_i\mathbf{x}_i - \mathbf{q}_i^{(k)}\|_2^2$$

per cui

$$\|\theta_i\mathbf{x}_i - \mathbf{q}_i^{(k)}\|_2^2 \leq 2\|\mathbf{x}_i\mathbf{x}_i^H - \mathbf{q}_i^{(k)}\mathbf{q}_i^{(k)\,H}\|_2^2.\ )$$

6.42 Siano $A,\ B \in \mathbf{C}^{n\times n}$. Il *problema generalizzato agli autovalori* è il problema di determinare i numeri λ tali che

$$A\mathbf{x} = \lambda B\mathbf{x}, \quad \mathbf{x} \neq \mathbf{0}. \tag{64}$$

I λ sono detti *autovalori del problema generalizzato* o *autovalori generalizzati* e le soluzioni $\mathbf{x}$ sono dette *autovettori*. Gli autovalori del problema generalizzato sono le soluzioni dell'*equazione caratteristica*

$$\det(A - \lambda B) = 0.$$

Se la matrice B è non singolare, il problema (64) è equivalente al problema

$$B^{-1}A\mathbf{x} = \lambda\mathbf{x}, \quad \mathbf{x} \neq \mathbf{0},$$

se la matrice A è non singolare, il problema (64) è equivalente al problema

$$A^{-1}B\mathbf{x} = \frac{1}{\lambda}\,\mathbf{x}, \quad \mathbf{x} \neq \mathbf{0}.$$

a) Si dica sotto quali ipotesi il problema (64) ha n autovalori (contati con la loro molteplicità);

b) si risolva il problema (64) nei seguenti casi

$$(1) \quad A = \begin{bmatrix} 1 & 1 \\ 1 & \text{-}1 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix},$$

$$(2) \quad A = \begin{bmatrix} 1 & 1 \\ 1 & \text{-}1 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix},$$

$$(3) \quad A = \begin{bmatrix} 1 & 1 \\ \text{-}1 & \text{-}1 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix};$$

c) si dimostri che se A è hermitiana e B è definita positiva, posto $B = LL^H$, vale

$$L^{-1}AL^{-H}\mathbf{y} = \lambda\mathbf{y}, \ \mathbf{y} = L^H\mathbf{x},$$

la matrice $L^{-1}AL^{-H}$ è hermitiana e quindi gli autovalori generalizzati sono reali;

d) se le matrici A e B sono reali e simmetriche, e B è definita positiva, allora la funzione definita su $\mathbf{R}^n$ $f(\mathbf{x}) = \dfrac{\mathbf{x}^H A\mathbf{x}}{\mathbf{x}^H B\mathbf{x}}$ è stazionaria nel punto $\mathbf{v} \in \mathbf{R}^n$ se e solo se

$$\lambda = \frac{\mathbf{v}^H A\mathbf{v}}{\mathbf{v}^H B\mathbf{v}}\,, \tag{65}$$

si mostri con un controesempio che la (65) non vale se B non è definita positiva;

e) si dimostri il seguente *teorema del minimax* per il problema (64): siano A hermitiana e B definita positiva e siano $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ gli autovalori del problema (64), allora

$$\lambda_{n-k+1} = \min_{V_k} \max_{\substack{\mathbf{x}\neq\mathbf{0} \\ \mathbf{x}\in V_k}} \frac{\mathbf{x}^H A\mathbf{x}}{\mathbf{x}^H B\mathbf{x}}\,, \quad \lambda_k = \max_{V_k} \min_{\substack{\mathbf{x}\neq\mathbf{0} \\ \mathbf{x}\in V_k}} \frac{\mathbf{x}^H A\mathbf{x}}{\mathbf{x}^H B\mathbf{x}}\,,$$

dove V_k è un qualunque sottospazio di $\mathbf{C}^n$ di dimensione k, per $k = 1, \ldots, n$;

f) siano A hermitiana e B definita positiva e siano $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ e $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_n$ rispettivamente gli autovalori dei problemi generalizzati

$$A\mathbf{x} = \lambda B\mathbf{x}$$

e

$$A_1\mathbf{z} = \mu B_1\mathbf{z},$$

dove A_1 e B_1 sono le sottomatrici principali di testa di ordine $n-1$ di A e B. Si dimostri che

$$\lambda_1 \geq \mu_1 \geq \lambda_2 \geq \mu_2 \geq \ldots \geq \mu_{n-1} \geq \lambda_n;$$

g) siano A hermitiana e B definita positiva, allora

$$\mathbf{x}^H B\mathbf{x} > |\mathbf{x}^H A\mathbf{x}| \text{ per ogni } \mathbf{x} \in \mathbf{C}^n, \mathbf{x} \neq \mathbf{0} \text{ se e solo se } \rho(B^{-1}A) < 1;$$

h) si risolva il problema (64) nel caso in cui A, $B \in \mathbf{R}^n$ sono le matrici tridiagonali simmetriche

$$A = \begin{bmatrix} \alpha & \beta & & \\ \beta & \alpha & \ddots & \\ & \ddots & \ddots & \beta \\ & & \beta & \alpha \end{bmatrix}, \quad B = \begin{bmatrix} \gamma & \delta & & \\ \delta & \gamma & \ddots & \\ & \ddots & \ddots & \delta \\ & & \delta & \gamma \end{bmatrix}.$$

(Traccia: a) si dica che cosa accade se $N(A) \cap N(B) \neq \{\mathbf{0}\}$ o se B è singolare; d) come controesempio si consideri il caso

$$A = \begin{bmatrix} 1 & 1 \\ 1 & \text{-}1 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix};$$

e) si applichi il teorema 6.7 alla matrice $L^{-1}AL^{-H}$; g) per il punto e) l'autovalore di modulo massimo di $B^{-1}A$ è uguale a

$$\max_{\mathbf{x}\neq\mathbf{0}} \frac{\mathbf{x}^H A\mathbf{x}}{\mathbf{x}^H B\mathbf{x}} \quad \text{oppura} \quad \min_{\mathbf{x}\neq\mathbf{0}} \frac{\mathbf{x}^H A\mathbf{x}}{\mathbf{x}^H B\mathbf{x}};$$

h) si noti che

$$A - \lambda B = \begin{bmatrix} \alpha - \lambda\gamma & \beta - \lambda\delta & & \\ \beta - \lambda\delta & \alpha - \lambda\gamma & \ddots & \\ & \ddots & \ddots & \beta - \lambda\delta \\ & & \beta - \lambda\delta & \alpha - \lambda\gamma \end{bmatrix},$$

e si sfrutti l'esercizio 2.40.)

Commento bibliografico

Il problema del calcolo degli autovalori e degli autovettori è uno dei più importanti problemi computazionali dell'algebra lineare: il calcolo degli autovalori è richiesto in molti problemi, spesso di grandi dimensioni, in cui le matrici sono sparse e generalmente dotate di struttura.

Il testo fondamentale su cui si basano gran parte degli studi sui metodi per il calcolo degli autovalori è il libro di Wilkinson [28], che riporta lo stato dell'arte in questo settore al 1965. Sistematiche presentazioni dei metodi per calcolare gli autovalori e gli autovettori di matrici sono riportate anche nei libri di Golub e Van Loan [7], Parlett [18], Schwarz, Rutishauser e Stiefel [21], Stewart [23]. I programmi in Algol dei principali metodi sono riportati nel libro di Wilkinson e Reinsch [30]. Un'esposizione elementare sul calcolo degli autovalori è riportata nel libro di Atkinson [1].

In questo libro l'esposizione della teoria della perturbazione è stata fatta tenendo conto di quanto riportato nel libro di Stoer e Bulirsch [26], l'ordinamento dei metodi segue lo schema riportato in Stewart [24], e la descrizione dei metodi segue la presentazione fatta nel più recente libro di Golub e Van Loan [7].

Una trattazione sistematica della teoria della perturbazione nel calcolo degli autovalori è stata elaborata da Wilkinson [28] sulla traccia di risultati sia classici (il teorema sulla separazione degli autovalori delle sottomatrici principali delle matrici simmetriche è stato dato da Cauchy nel 1823), sia più recenti (il teorema del minimax è riportato da Courant e Hilbert nel 1953 e il teorema di Bauer-Fike è del 1960).

La riduzione in forma tridiagonale delle matrici hermitiane e in forma di Hessenberg superiore utilizza le matrici elementari di Householder, introdotte nel 1958. Una trattazione completa di queste matrici è riportata nel libro di Householder [9] del 1964; per un'analisi accurata dell'errore delle trasformazioni di Householder si veda, oltre a [28], il libro di Lawson e Hanson [13] del 1980.

Il metodo di Givens, presentato nel 1954 [6], che può essere considerato un precursore del metodo di Householder, utilizza le matrici di rotazione introdotte da Jacobi nel 1846. In [6] viene anche proposta per la prima volta l'utilizzazione della successione di Sturm per la separazione degli autovalori di matrici tridiagonali simmetriche. Per rappresentare una successione di trasformazioni di Givens in forma compatta, analoga a quella usata per le matrici di Householder, si può utilizzare una tecnica, descritta da Stewart [25]. Una variante del metodo di Givens, detta *fast Givens transformations*, che non richiede il calcolo di radici quadrate, e che ha un costo computazionale confrontabile con quello del metodo di Householder, è stata proposta da Gentleman [5].

La riduzione di una matrice in forma di Hessenberg superiore con ma-

trici elementari di Gauss è esposta in Businger [2], con lo studio della stabilità numerica. Il metodo di Hyman è descritto in [10].

Il metodo QR è stato sviluppato da Francis nel 1961 [4] e rappresenta uno sviluppo del metodo LR di Rutishauser [20] del 1958. In molti lavori successivi la convergenza del metodo QR è stata studiata nei suoi aspetti sia teorici che pratici. Una descrizione sistematica delle tecniche di shift è stata fatta da Wilkinson [29], e lo studio della stabilità numerica è riportato in Wilkinson [28].

Il metodo di Jacobi, presentato da Jacobi nel 1846, è stato il metodo principe per il calcolo degli autovalori delle matrici hermitiane non sparse fino all'introduzione del metodo QR. Nel 1949 ne fu suggerita da Goldstine l'utilizzazione sui primi calcolatori. Nel 1953 il metodo è stato implementato sul calcolatore ILLIAC presso l'Università dell'Illinois. Vari autori, sulla traccia dei primi lavori di Henrici [8] del 1958 e di Wilkinson [27] del 1962, hanno individuato la convergenza quadratica del metodo di Jacobi, sia con la strategia classica che con quella ciclica. Sono stati fatti dei tentativi, senza grossi successi, per estendere il metodo anche a classi di matrici non hermitiane. A questo proposito si veda la bibliografia riportata in Golub e Van Loan [7]. La tecnica, descritta in questo testo, per la costruzione della matrice di rotazione, è dovuta a Rutishauser (1971).

Il metodo delle potenze si basa sulle proprietà asintotiche delle potenze di matrici ed è stato suggerito nel 1913 da Müntz. Il metodo delle potenze inverse è stato proposto da Wielandt nel 1944. Sulle proprietà delle potenze delle matrici si basa anche il metodo di Leverrier (1840) che è il più vecchio metodo pratico per il calcolo dei coefficienti del polinomio caratteristico. Per il metodo delle potenze e delle sue varianti si veda il libro di Wilkinson [28], in cui viene anche studiato estensivamente l'uso del metodo per approssimare gli autovalori complessi. La convergenza cubica del metodo di Rayleigh nel caso di matrici hermitiane è stata dimostrata da Parlett [17]. La dimostrazione del teorema di convergenza del metodo delle iterazioni ortonormali è riportata in [7]. L'accelerazione di Ritz per il caso delle matrici hermitiane è stata suggerita da Stewart [22].

Il metodo di Lanczos è stato presentato nel 1950 [12]. Questo metodo nella versione originale presenta instabilità numerica. L'algoritmo di tridiagonalizzazione qui descritto, introdotto da Paige [15] nel 1972, è una delle varianti più stabili. Lo studio della velocità di convergenza del metodo di Lanczos si basa sui risultati della teoria di Kaniel-Paige, elaborata nei due lavori di Kaniel [11] e di Paige [14]. Nei successivi lavori di Paige, in particolare [16], sono esposte le tecniche di ortogonalizzazione che rendono più stabile il metodo. Vari tentativi sono stati fatti per individuare varianti del metodo che non richiedano la riortogonalizzazione; per una trattazione generale del metodo di Lanczos e in particolare per la variante del prolunga-

mento del metodo oltre le dimensioni della matrice, si veda il libro di Cullum e Willoughby [3], in cui si danno delle indicazioni per individuare quali degli autovalori che si vengono a determinare sono accettabili oppure no. Per un confronto critico dei metodi di Lanczos e delle iterazioni ortogonali, con la descrizione del software esistente, si veda [19].

Bibliografia

[1] K. E. Atkinson, *An Introduction to Numerical Analysis*, John Wiley, New York, 1978.

[2] P. Businger, "Reducing a Matrix to Hessenberg Form", *Math. Comp.*, 23, 1969, pp. 819-821.

[3] J. Cullum, R. A. Willoughby, *Lanczos Algorithms for Large Symmetric Eigenvalue Computations*, vol. I, Theory, Progress in Scientific Computing, 3., Birkhäuser, Boston, 1985.

[4] J. G. F. Francis, "The QR Transformation: A Unitary Analogue to the LR Transformation", *Comp. J.*, 4, 1961, pp. 265-271, 332-334.

[5] M. Gentleman, "Least Squares Computations by Givens Transformations without Square Roots", *J. Inst. Math. Appl.*, 12, 1973, pp. 329-336.

[6] W. Givens, "Numerical Computation of Characteristic Values of a Real Symmetric Matrix", Oak Ridge National Laboratory, *ORN* L-1574, 1954.

[7] G. H. Golub, C. F. Van Loan, *Matrix Computations*, 2nd Edition, The Johns Hopkins University Press, Baltimore, Maryland, 1989.

[8] P. Henrici, "On the Speed of Convergence of Cyclic and Quasicyclic JacobiMethods for Computing the Eigenvalues of Hermitian Matrice s", *SIAM J. Applied Math.*, 6, 1958, pp. 144-162.

[9] A. S. Householder, *The Theory of Matrices in Numerical Analysis*, Blaisdell, Boston, Mass., 1964.

[10] M. Hyman, "Eigenvalues and Eigenvectors of General Matrices", *Twelfth National Meeting A. C. M.*, Houston, Texas, 1957.

[11] S. Kaniel, "Estimates for Some Computational Techniques in Linear Algebra", *Math. Comp.*, 20, 1966, pp. 369-378.

[12] C. Lanczos, "An Iteration Method for the Solution of the Eigenvalue Problem of Linear Differential and Integral Operators", *J. Res. Nat. Bur. Stand.*, 45, 1950, pp. 255-282.

[13] C. L. Lawson, R. J. Hanson, *Solving Least Squares Problems*, Prentice-Hall, Englewood Cliffs, 1980.

[14] C. C. Paige, *The Computation of Eigenvalues and Eigenvectors of Very Large Sparse Matrices*, Ph. D. thesis, London University, 1961.

[15] C. C. Paige, "Computational Variants of the Lanczos Method for the Eigenproblem", *J. Inst. Math. Applic.*, 10, 1972, pp. 373-381.

[16] C. C. Paige, "Practical Use of Symmetric Lanczos Process with Reorthogonalization", *BIT* 10, 1976, pp. 183-195.

[17] B. N. Parlett, "The Rayleigh Quotient Iteration and Some Generalizations for Nonnormal Matrices", *Math. Comp.*, 28, 1974, pp. 679-693.

[18] B. N. Parlett, *The Symmetric Eigenvalue Problem*, Prentice-Hall, Englewood Cliffs, 1980.

[19] B. N. Parlett, "The Software Scene in the Extraction of Eigenvalues from Sparse Matrices", *SIAM J. Sci. Stat. Comput.*, 5, 1984, pp. 590-604.

[20] H. Rutishauser, "Solution of Eigenvalue Problems with the LR Transformation", *Nat. Bur. Stand. App. Math. Ser.*, 49, 1958, pp. 47-81.

[21] H. R. Schwarz, H. Rutishauser, E. Stiefel, *Numerical Analysis of Symmetric Matrices*, Prentice-Hall, Englewood Cliffs, 1973.

[22] G. W. Stewart, "Accelerating the Orthogonal Iteration for the Eigenvalues of a Hermitian Matrix", *Numer. Math.*, 13, 1969, pp. 362-376.

[23] G. W. Stewart, *Introduction to Matrix Computation*, Academic Press, New York, 1973.

[24] G. W. Stewart, "The Numerical Treatment of Large Eigenvalue Problems", *Proc. IFIP Congress 74*, North-Holland, 1974, pp. 666-672.

[25] G. W. Stewart, "The Economical Storage of Plane Rotations", *Numer. Math.*, 25, 1976, pp. 137-138.

[26] J. Stoer, R. Bulirsch, *Introduction to Numerical Analysis*, Springer-Verlag, New York, 1980.

[27] J. H. Wilkinson, "Note on the Quadratic Convergence of the Cyclic Jacobi Process", *Numer. Math.*, 4, 1962, pp. 296-300.

[28] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.

[29] J. H. Wilkinson, "Global Convergence of Tridiagonal QR Algorithm With Origin Shift", *Lin. Alg. and Its Applic.*, 1, 1968, pp. 409-420.

[30] J. H. Wilkinson, C. Reinsch, *Handbook for Automatic Computation, vol. 2, Linear Algebra*, Springer-Verlag, New York, 1971.

Capitolo 7

IL PROBLEMA LINEARE DEI MINIMI QUADRATI

1. Le equazioni normali

Sia

$$A\mathbf{x} = \mathbf{b} \tag{1}$$

un sistema lineare in cui la matrice $A \in \mathbf{C}^{m\times n}$ dei coefficienti è tale che $m \geq n$. Se $m > n$, il sistema (1) ha più equazioni che incognite e si dice *sovradeterminato*. Se il sistema (1) non ha soluzione, fissata una norma vettoriale $\| \ . \ \|$, si ricercano i vettori $\mathbf{x} \in \mathbf{C}^n$ che minimizzano la quantità $\|A\mathbf{x} - \mathbf{b}\|$. In norma 2, il problema diventa quello di determinare un vettore $\mathbf{x} \in \mathbf{C}^n$ tale che

$$\|A\mathbf{x} - \mathbf{b}\|_2 = \min_{\mathbf{y}\in\mathbf{C}^n} \|A\mathbf{y} - \mathbf{b}\|_2 = \gamma. \tag{2}$$

Tale problema viene detto *problema dei minimi quadrati*.

Il seguente teorema caratterizza l'insieme X dei vettori $\mathbf{x} \in \mathbf{C}^n$ che soddisfano alla (2).

7.1 Teorema. *Valgono le seguenti proprietà:*

a) $\mathbf{x} \in X$ *se e solo se*

$$A^H A\mathbf{x} = A^H\mathbf{b}. \tag{3}$$

Il sistema (3) viene detto sistema delle equazioni normali o sistema normale.

b) X *è un insieme non vuoto, chiuso e convesso.*

c) L'insieme X *si riduce ad un solo elemento* $\mathbf{x}^*$ *se e solo se la matrice* A *ha rango massimo.*

d) Esiste $\mathbf{x}^* \in X$ *tale che*

$$\|\mathbf{x}^*\|_2 = \min_{\mathbf{x}\in X} \|\mathbf{x}\|_2. \tag{4}$$

Il vettore $\mathbf{x}^*$ *è l'unico vettore di* X *che appartiene a* $N(A^H A)^\perp$ *ed è detto soluzione di minima norma.*

Dim. a) Siano

$$S(A) = \{\ \mathbf{y} \in \mathbf{C}^m \ :\ \mathbf{y} = A\mathbf{x},\ \mathbf{x} \in \mathbf{C}^n\ \}$$

e

$$S(A)^{\perp} = \{\ \mathbf{z} \in \mathbf{C}^m\ :\ \mathbf{z}^H\mathbf{y} = 0,\ \text{per ogni } \mathbf{y} \in S(A)\ \}$$

il sottospazio di $\mathbf{C}^m$ immagine di A, e il sottospazio ortogonale a $S(A)$ (si vedano i paragrafi 2 e 6 del capitolo 1). Il vettore $\mathbf{b}$ può essere così decomposto

$$\mathbf{b} = \mathbf{b}_1 + \mathbf{b}_2, \quad \text{dove } \mathbf{b}_1 \in S(A) \text{ e } \mathbf{b}_2 \in S(A)^{\perp},$$

per cui per il residuo

$$\mathbf{r} = \mathbf{b}_1 - A\mathbf{x} + \mathbf{b}_2 = \mathbf{y} + \mathbf{b}_2, \quad \text{dove } \mathbf{y} = \mathbf{b}_1 - A\mathbf{x} \in S(A) \text{ e } \mathbf{b}_2 \in S(A)^{\perp}$$

vale

$$\|\mathbf{r}\|_2^2 = (\mathbf{y} + \mathbf{b}_2)^H(\mathbf{y} + \mathbf{b}_2) = \|\mathbf{y}\|_2^2 + \|\mathbf{b}_2\|_2^2,$$

in quanto $\mathbf{y}^H\mathbf{b}_2 = \mathbf{b}_2^H\mathbf{y} = 0$. Poiché solo $\mathbf{y}$ dipende da $\mathbf{x}$, si ha che $\|\mathbf{r}\|_2^2$ è minimo se e solo se $\mathbf{b}_1 = A\mathbf{x}$, cioè se e solo se il vettore $\mathbf{r}$ appartiene a $S(A)^{\perp}$ ed è quindi ortogonale alle colonne di A, cioè

$$A^H\mathbf{r} = A^H(\mathbf{b} - A\mathbf{x}) = \mathbf{0}.$$

Ne segue quindi che $\mathbf{x} \in X$ se e solo se $\mathbf{x}$ è soluzione di (3). Inoltre risulta $\gamma^2 = \|\mathbf{b}_2\|_2^2$.

Nel caso di $\mathbf{R}^2$ con una matrice A di rango 1 si può dare la seguente interpretazione geometrica, illustrata nella figura 7.1. Il vettore $\mathbf{b} = \mathbf{r} - A\mathbf{x}$ risulta decomposto in un sol modo nel vettore $\mathbf{b}_2 = \mathbf{r} \in S(A)^{\perp}$ e nel vettore $\mathbf{b}_1 = A\mathbf{x} \in S(A)$. Il vettore $A\mathbf{x}$ è quindi la proiezione ortogonale del vettore $\mathbf{b}$ sul sottospazio generato dalle colonne di A.

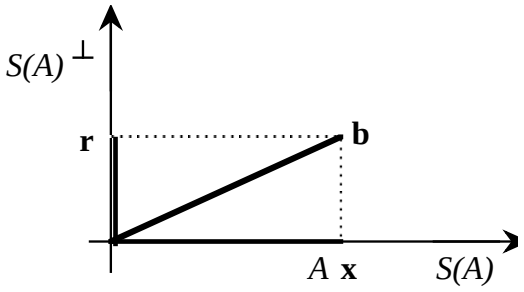


Fig. 7.1 - Proiezione ortogonale del vettore $\mathbf{b}$.

b) Da quanto detto precedentemente, segue che l'insieme X è non vuoto. Se $\mathbf{x}_0$ è tale che

$$A^HA\mathbf{x}_0 = A^H\mathbf{b}$$

allora risulta

$$X = \{\ \mathbf{x} \in \mathbf{C}^n\ :\ \mathbf{x} = \mathbf{x}_0 + \mathbf{v},\ \mathbf{v} \in N(A^H A)\ \}.$$

Quindi X è una *varietà lineare affine*, parallela ad $N(A^H A)$, passante per $\mathbf{x}_0$, e poiché $N(A^H A)$ è chiuso e convesso, X è un insieme chiuso e convesso.

c) La matrice A ha rango massimo se e solo se la matrice $A^H A$ è non singolare (si veda il paragrafo 6 del capitolo 1) e quindi se e solo se il sistema (3) ha una e una sola soluzione $\mathbf{x}^*$. Perciò l'insieme X è costituito dal solo elemento $\mathbf{x}^*$ se e solo se la matrice A ha rango massimo. In tal caso l'insieme $N(A^H A)$ è costituito dal solo elemento nullo.

d) l'esistenza della soluzione di minima norma è ovvia nel caso in cui X si riduce al solo elemento $\mathbf{x}^*$. Se X non si riduce al solo elemento $\mathbf{x}^*$, sia $\mathbf{x}_0 \in X$ e si consideri l'insieme

$$B = \{\ \mathbf{x} \in \mathbf{C}^n\ :\ \|\mathbf{x}\|_2 \leq \|\mathbf{x}_0\|_2\ \}.$$

Poiché, se $\mathbf{x} \in X$, ma $\mathbf{x} \notin B$, risulta $\|\mathbf{x}\|_2 > \|\mathbf{x}_0\|_2$, allora

$$\min_{\mathbf{x} \in X} \|\mathbf{x}\|_2 = \min_{\mathbf{x} \in X \cap B} \|\mathbf{x}\|_2.$$

L'insieme $X \cap B$ è un insieme non vuoto, limitato e chiuso, in quanto intersezione di insiemi chiusi, e quindi compatto; essendo la norma una funzione continua, esiste un $\mathbf{x}^* \in X \cap B$ per cui vale la (4).

Inoltre $\mathbf{x}^*$ è l'unico vettore di X appartenente a $N(A^H A)^\perp$. Infatti esistono e sono unici $\mathbf{y} \in N(A^H A)$ e $\mathbf{z} \in N(A^H A)^\perp$ tali che $\mathbf{x}^* = \mathbf{y} + \mathbf{z}$. Poiché $\mathbf{x}^*$ è soluzione di (3), è $A^H A(\mathbf{y} + \mathbf{z}) = A^H \mathbf{b}$, da cui $A^H A\mathbf{z} = A^H \mathbf{b}$ e quindi $\mathbf{z}$ è soluzione del problema (2). Se $\mathbf{y}$ non fosse uguale a $\mathbf{0}$, $\mathbf{z}$ avrebbe norma 2 minore di $\|\mathbf{x}^*\|_2$, ciò che è assurdo perché $\mathbf{x}^*$ è la soluzione di minima norma: ne segue che $\mathbf{x}^* = \mathbf{z} \in N(A^H A)^\perp$.

Nel caso di $\mathbf{R}^2$ con una matrice A di rango 1, si può dare l'interpretazione geometrica illustrata nella figura 7.2, in cui è riportata la varietà X, parallela alla varietà $N(A^H A)$. Il punto $\mathbf{x}_0$ è un qualunque punto di X. Il punto $\mathbf{x}^*$ è quello di minima norma, e quindi quello più vicino all'origine O dello spazio $\mathbf{C}^n$. ∎

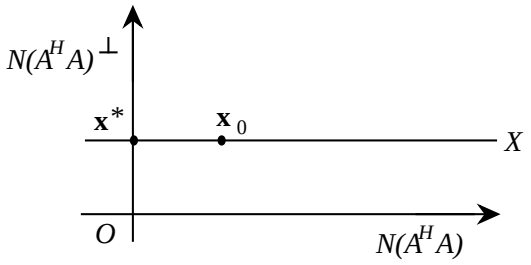


Fig. 7.2 - $\mathbf{x}^*$ è la soluzione di minima norma.

Se la matrice A ha rango massimo, allora la soluzione del problema dei minimi quadrati può essere ottenuta risolvendo il sistema (3). In tal caso, poiché la matrice $A^H A$ è definita positiva, si può utilizzare per la risoluzione il metodo di Cholesky. Determinata la matrice L triangolare inferiore tale che

$$LL^H = A^H A,$$

la soluzione $\mathbf{x}^*$ di (3) viene calcolata risolvendo successivamente i due sistemi di ordine n con matrice dei coefficienti triangolare

$$L\mathbf{y} = A^H \mathbf{b}$$
$$L^H \mathbf{x} = \mathbf{y}.$$

Il costo computazionale è di $n^2 m/2$ operazioni moltiplicative per la costruzione della matrice hermitiana $A^H A$ e di $n^3/6$ operazioni moltiplicative per il calcolo della soluzione del sistema (3). Quindi in totale il calcolo della soluzione del problema dei minimi quadrati per mezzo della risoluzione del sistema (3) con il metodo di Cholesky ha un costo computazionale di

$$f_1(n,m) = \frac{n^2}{2}(m + \frac{n}{3}) \quad \text{operazioni moltiplicative.} \tag{5}$$

7.2 Esempio. Si consideri il problema dei minimi quadrati (2) con

$$A = \frac{1}{45}\begin{bmatrix} 14 & 32 & -38 \\ -44 & 58 & 8 \\ -18 & 96 & 51 \\ 63 & -36 & 54 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}.$$

Per determinare una soluzione di (2) si costruiscono

$$A^T A = \frac{1}{81}\begin{bmatrix} 257 & -244 & 64 \\ -244 & 596 & 88 \\ 64 & 88 & 281 \end{bmatrix}, \quad A^T \mathbf{b} = \frac{1}{3}\begin{bmatrix} 1 \\ 10 \\ 5 \end{bmatrix}.$$

Poiché la matrice $A^T A$ è non singolare, si applica il metodo di Cholesky, ottenendo la fattorizzazione LL^T, dove

$$L = \frac{1}{9\sqrt{257}}\begin{bmatrix} 257 & 0 & 0 \\ -244 & 306 & 0 \\ 64 & \frac{2124}{17} & \frac{243}{17}\sqrt{257} \end{bmatrix}.$$

La soluzione $\mathbf{x}^*$ risulta

$$\mathbf{x}^* = \frac{1}{54}\begin{bmatrix} 46 \\ 43 \\ 2 \end{bmatrix}.$$

Sostituendo nella (2) si ha che

$$A\mathbf{x}^* - \mathbf{b} = \frac{1}{5}\,[\text{-}1,\ \text{-}4,\ 2,\ \text{-}2]^T,$$

e quindi $\gamma = \|A\mathbf{x}^* - \mathbf{b}\|_2 = 1.$ ∎

Se la matrice A non ha rango massimo, non si può risolvere il sistema (3) con il metodo di Cholesky, ma si può applicare il metodo di Gauss con la variante del massimo pivot.

7.3 Esempio. Si consideri il problema dei minimi quadrati (2) con

$$A = \frac{1}{45}\begin{bmatrix} 6 & 12 & \text{-}72 \\ \text{-}16 & \text{-}7 & \text{-}8 \\ 58 & 16 & 104 \\ 87 & 24 & 156 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}.$$

Si ha

$$A^TA = \frac{1}{81}\begin{bmatrix} 449 & 128 & 772 \\ 128 & 41 & 184 \\ 772 & 184 & 1616 \end{bmatrix}, \quad A^T\mathbf{b} = \begin{bmatrix} 3 \\ 1 \\ 4 \end{bmatrix},$$

e la matrice A^TA è singolare. Calcolando la fattorizzazione LU della matrice A^TA con il metodo di Gauss si ottiene

$$L = \begin{bmatrix} 1 & 0 & 0 \\ \dfrac{128}{449} & 1 & 0 \\ \dfrac{772}{449} & \text{-}8 & 1 \end{bmatrix}, \quad U = \frac{1}{81}\begin{bmatrix} 449 & 128 & 772 \\ 0 & \dfrac{2025}{449} & \text{-}\dfrac{16200}{449} \\ 0 & 0 & 0 \end{bmatrix}.$$

Ne segue che la matrice A ha rango 2. L'insieme X risulta formato dai vettori

$$\mathbf{x} = \begin{bmatrix} -\dfrac{1}{5} - 4h \\ \dfrac{13}{5} + 8h \\ h \end{bmatrix}, \quad h \in \mathbf{C}.$$

Il vettore di minima norma $\mathbf{x}^*$ può essere ricavato calcolando il valore di h per cui la funzione

$$f(h) = \|\mathbf{x}\|_2^2$$

è minima. Tale valore è $h = -\dfrac{4}{15}$, a cui corrisponde il vettore

$$\mathbf{x}^* = \frac{1}{15}\,[13,\ 7,\ -4]^T.$$

Sostituendo nella (2) si ha che

$$A\mathbf{x}^* - \mathbf{b} = \frac{1}{3}\,[-1,\ -4,\ -1,\ 0]^T,$$

e quindi $\gamma = \|A\mathbf{x}^* - \mathbf{b}\|_2 = \sqrt{2}$. ∎

Operando in aritmetica finita il sistema (3) può risultare non risolubile, e in tal caso non può essere utilizzato per risolvere il problema (2).

7.4 Esempio. Sia u la precisione di macchina con cui si eseguono i calcoli. Si consideri il problema (2) con

$$A = \begin{bmatrix} 3 & 3 \\ 4 & 4 \\ 0 & \alpha \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix},$$

dove α è un numero tale che $u \le \alpha < 1$ e $\alpha^2 < u$. Si ha

$$A^TA = \begin{bmatrix} 25 & 25 \\ 25 & 25+\alpha^2 \end{bmatrix}, \quad A^T\mathbf{b} = \begin{bmatrix} 7 \\ 7+\alpha \end{bmatrix}.$$

Applicando il metodo di Cholesky, si calcola la matrice L della fattorizzazione LL^T di A^TA:

$$L = \begin{bmatrix} 5 & 0 \\ 5 & \alpha \end{bmatrix},$$

da cui si ricava la soluzione del problema (2) :

$$\mathbf{x}^* = \left[\frac{7}{25} - \frac{1}{\alpha}\,,\ \frac{1}{\alpha}\right]^T.$$

Però operando con precisione u, poiché $\alpha^2 < u$, la matrice effettivamente calcolata al posto della A^TA è

$$\begin{bmatrix} 25 & 25 \\ 25 & 25 \end{bmatrix}$$

e ha rango 1. Con tale matrice al posto della A^TA il sistema (3) non sarebbe risolubile. ∎

2. Metodo QR per il calcolo della soluzione del problema dei minimi quadrati

Si esamina ora un altro procedimento, detto metodo QR, che opera direttamente sulla matrice A fattorizzandola nella forma QR.

Si supponga dapprima che la matrice $A \in \mathbf{C}^{m\times n}$ abbia rango massimo $k = n \leq m$. Si applica il metodo di Householder alla matrice A, ottenendo una successione di matrici di Householder

$$P^{(k)} \in \mathbf{C}^{m\times m},\ k = 1, \ldots, n.$$

Posto $Q^H = P^{(1)}P^{(2)}\ldots P^{(n)}$, risulta

$$A = QR, \tag{6}$$

dove la matrice $R \in \mathbf{C}^{m\times n}$ ha la forma

$$R = \begin{bmatrix} R_1 \\ \\ O \end{bmatrix} \quad \begin{matrix} \} & n \text{ righe} \\ \\ \} & m-n \text{ righe} \end{matrix} \tag{7}$$

ed R_1 è una matrice triangolare superiore non singolare in quanto A ha rango massimo. Dalla (6) si ha

$$\begin{aligned} \|A\mathbf{x} - \mathbf{b}\|_2 &= \|QR\mathbf{x} - \mathbf{b}\|_2 = \|Q(R\mathbf{x} - Q^H\mathbf{b})\|_2 = \|R\mathbf{x} - Q^H\mathbf{b}\|_2 \\ &= \|R\mathbf{x} - \mathbf{c}\|_2. \end{aligned} \tag{8}$$

dove $\mathbf{c} = Q^H\mathbf{b}$. Partizionando il vettore $\mathbf{c}$ nel modo seguente

$$\mathbf{c} = \begin{bmatrix} \mathbf{c}_1 \\ \\ \mathbf{c}_2 \end{bmatrix} \quad \begin{matrix} \} & n \text{ componenti} \\ \\ \} & m-n \text{ componenti} \end{matrix}$$

per la (7) vale

$$R\mathbf{x} - \mathbf{c} = \begin{bmatrix} R_1\mathbf{x} - \mathbf{c}_1 \\ \\ -\mathbf{c}_2 \end{bmatrix}.$$

Per la (8) risulta

$$\begin{aligned} \min_{\mathbf{x}\in\mathbf{C}^n} \|A\mathbf{x} - \mathbf{b}\|_2^2 &= \min_{\mathbf{x}\in\mathbf{C}^n} \|R\mathbf{x} - \mathbf{c}\|_2^2 = \min_{\mathbf{x}\in\mathbf{C}^n} \left[\|R_1\mathbf{x} - \mathbf{c}_1\|_2^2 + \|\mathbf{c}_2\|_2^2\right] \\ &= \|\mathbf{c}_2\|_2^2 + \min_{\mathbf{x}\in\mathbf{C}^n} \|R_1\mathbf{x} - \mathbf{c}_1\|_2^2. \end{aligned}$$

Poiché R_1 è non singolare, la soluzione $\mathbf{x}^*$ del sistema lineare

$$R_1\mathbf{x} = \mathbf{c}_1 \tag{9}$$

è tale che

$$\min_{\mathbf{x}\in\mathbf{C}^n} \|R_1\mathbf{x} - \mathbf{c}_1\|_2 = \|R_1\mathbf{x}^* - \mathbf{c}_1\|_2 = 0.$$

Ne segue che $\mathbf{x}^*$ è la soluzione del problema (2) e

$$\gamma = \|\mathbf{c}_2\|_2.$$

Analogamente al caso della risoluzione dei sistemi lineari, il metodo può essere applicato senza calcolare effettivamente né le matrici $P^{(k)}, k = 1, \ldots, n$, né la matrice Q. Si può procedere infatti nel modo seguente: sia

$$P^{(k)} = I - \beta_k \mathbf{v}_k \mathbf{v}_k^H, \quad k = 1, \ldots, n,$$

secondo la notazione del paragrafo 12 del capitolo 4. Si considera la matrice

$$T^{(1)} = [A \mid \mathbf{b}]$$

e si costruisce le successione delle matrici $T^{(k)}$ tali che

$$T^{(k+1)} = P^{(k)}T^{(k)} = T^{(k)} - \beta_k \mathbf{v}_k \mathbf{v}_k^H T^{(k)}.$$

Al termine dopo n passi si ottiene la matrice

$$T^{(n+1)} = [R \mid \mathbf{c}].$$

Il costo computazionale di questa fattorizzazione è di $n^2(m - n/3)$ operazioni moltiplicative (si veda il paragrafo 13 del capitolo 4); il costo computazionale della risoluzione del sistema triangolare (9) è di $n^2/2$ operazioni moltiplicative. Quindi il costo computazionale del calcolo della soluzione del problema dei minimi quadrati è di

$$f_2(n, m) = n^2(m - \frac{n}{3}) \qquad \text{operazioni moltiplicative.} \tag{10}$$

Confrontando questo costo computazionale con quello riportato nella (5) risulta che

$$f_2(n, m) \geq f_1(n, m) \qquad \text{per} \quad m \geq n,$$

per cui il metodo QR richiede in generale più operazioni di quante sono richieste risolvendo con il metodo di Cholesky il sistema normale (3). Se $m = n$ i due metodi hanno lo stesso costo computazionale.

La matrice R_1 ottenuta con il metodo QR e la matrice L^H ottenuta applicando il metodo di Cholesky alla matrice $A^H A$ sono uguali a meno della moltiplicazione per una matrice di fase. Infatti dalla (6) si ha

$$LL^H = A^H A = R^H Q^H QR = R^H R = R_1^H R_1,$$

dove sia la L^H che la R_1 sono triangolari superiori. Quindi

$$R_1 = DL^H,$$

dove $D \in \mathbf{C}^{n\times n}$ è una matrice diagonale unitaria.

7.5 Esempio. Per calcolare la soluzione del problema dei minimi quadrati (2) dell'esempio 7.2, si applica il metodo QR. Si considera la matrice

$$T^{(1)} = [A \mid \mathbf{b}] = \frac{1}{45}\begin{bmatrix} 14 & 32 & \text{-}38 & 45 \\ \text{-}44 & 58 & 8 & 45 \\ \text{-}18 & 96 & 51 & 45 \\ 63 & \text{-}36 & 54 & 45 \end{bmatrix},$$

e dopo 3 passi si ottiene la matrice

$$T^{(4)} = \begin{bmatrix} R_1 & \mathbf{c}_1 \\ O & \mathbf{c}_2 \end{bmatrix} = \frac{1}{9\sqrt{257}}\begin{bmatrix} \text{-}257 & 244 & \text{-}64 & \text{-}27 \\ 0 & \text{-}306 & -\dfrac{2124}{17} & -\dfrac{4221}{17} \\ 0 & 0 & \dfrac{243}{17}\sqrt{257} & -\dfrac{9}{17}\sqrt{257} \\ 0 & 0 & 0 & 9\sqrt{257} \end{bmatrix}.$$

Risolvendo il sistema (9) si ricava la soluzione $\mathbf{x}^*$ già trovata nell'esempio 7.2. Inoltre è

$$\mathbf{c}_2 = [1],$$

e quindi $\gamma = \|\mathbf{c}_2\|_2 = 1$. ∎

Se la matrice A non ha rango massimo, la matrice R_1 ottenuta ha almeno un elemento diagonale nullo e quindi non è possibile calcolare la soluzione del sistema (9). Questa difficoltà viene superata applicando il metodo QR con la tecnica del *massimo pivot per colonne* nel modo seguente: al k-esimo passo, costruita la matrice $A^{(k)}$ della forma

$$A^{(k)} = \begin{bmatrix} C^{(k)} & D^{(k)} \\ O & B^{(k)} \end{bmatrix} \quad \begin{matrix} \} & k-1 \text{ righe} \\ \} & m-k+1 \text{ righe} \end{matrix}$$

si determina la colonna di $B^{(k)}$ la cui norma 2 è massima. Sia j, $1 \leq j \leq n-k+1$, l'indice di tale colonna. Se $j \neq 1$, si scambiano fra loro la k-esima e la $(k+j-1)$-esima colonna della matrice $A^{(k)}$. Quindi si applica la matrice elementare $P^{(k)}$ alla matrice con le colonne così permutate. Se il rango di A è $r < m$, questo procedimento termina dopo r passi, e si ottiene una decomposizione del tipo

$$A\Pi = QR, \tag{11}$$

dove $\Pi \in \mathbf{R}^{n\times n}$ è una matrice di permutazione, $Q \in \mathbf{C}^{m\times m}$ è una matrice unitaria ed R è della forma

$$R = \begin{bmatrix} R_1 & S \\ \\ O & O \end{bmatrix} \begin{matrix} \} & r \text{ righe} \\ \\ \} & m-r \text{ righe} \end{matrix} \tag{12}$$

in cui $R_1 \in \mathbf{C}^{r\times r}$ è triangolare superiore non singolare e $S \in \mathbf{C}^{r\times(n-r)}$. Gli elementi diagonali di R_1 risultano positivi e ordinati in ordine di modulo non crescente. Dalla (11) si ottiene

$$\|A\mathbf{x} - \mathbf{b}\|_2 = \|R\Pi^T\mathbf{x} - \mathbf{c}\|_2, \quad \text{dove} \quad \mathbf{c} = Q^H\mathbf{b}.$$

Partizionando i vettori $\mathbf{y} = \Pi^T\mathbf{x}$ e $\mathbf{c}$ nel modo seguente

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \\ \mathbf{y}_2 \end{bmatrix} \begin{matrix} \} & r \text{ componenti} \\ \} & n-r \text{ componenti} \end{matrix} \qquad \mathbf{c} = \begin{bmatrix} \mathbf{c}_1 \\ \\ \mathbf{c}_2 \end{bmatrix} \begin{matrix} \} & r \text{ componenti} \\ \} & m-r \text{ componenti} \end{matrix}$$

per la (12) risulta

$$R\Pi^T\mathbf{x} - \mathbf{c} = \begin{bmatrix} R_1\mathbf{y}_1 + S\mathbf{y}_2 - \mathbf{c}_1 \\ \\ -\mathbf{c}_2 \end{bmatrix}.$$

Poiché per ogni vettore $\mathbf{y}_2 \in \mathbf{C}^{n-r}$, esiste un vettore $\mathbf{y}_1 \in \mathbf{C}^r$, tale che

$$R_1\mathbf{y}_1 = \mathbf{c}_1 - S\mathbf{y}_2,$$

la soluzione del problema

$$\min_{\substack{\mathbf{y}_1\in\mathbf{C}^r \\ \mathbf{y}_2\in\mathbf{C}^{n-r}}} \|R_1\mathbf{y}_1 + S\mathbf{y}_2 - \mathbf{c}_1\|$$

non è unica e risulta

$$\begin{bmatrix} \mathbf{y}_1 \\ \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} R_1^{-1}(\mathbf{c}_1 - S\mathbf{y}_2) \\ \\ \mathbf{y}_2 \end{bmatrix},$$

e

$$\mathbf{x} = \Pi \begin{bmatrix} \mathbf{y}_1 \\ \\ \mathbf{y}_2 \end{bmatrix}.$$

Si osservi che se la matrice S è nulla, la soluzione $\mathbf{x}^*$ di minima norma si ottiene ponendo $\mathbf{y}_2 = \mathbf{0}$, mentre ciò non è vero nel caso generale.

Dal punto di vista pratico si fissa una costante ϵ, che dipende dalla precisione con cui si eseguono i calcoli, e il procedimento termina quando tutte le colonne di $B^{(k)}$ hanno norma minore di ϵ.

7.6 Esempio. Applicando il metodo QR al problema dell'esempio 7.3, si ha

$$T^{(1)} = [A \mid \mathbf{b}] = \frac{1}{45} \begin{bmatrix} 6 & 12 & -72 & 45 \\ -16 & -7 & -8 & 45 \\ 58 & 16 & 104 & 45 \\ 87 & 24 & 156 & 45 \end{bmatrix},$$

e dopo 3 passi si ottiene la matrice

$$T^{(4)} = \begin{bmatrix} R_1 & S & \mathbf{c}_1 \\ O & O & \mathbf{c}_2 \end{bmatrix} = \frac{1}{9\sqrt{449}} \begin{bmatrix} -449 & -128 & -772 & -243 \\ 0 & -45 & 360 & -117 \\ 0 & 0 & 0 & 9\sqrt{449} \\ 0 & 0 & 0 & 9\sqrt{449} \end{bmatrix}.$$

Quindi la soluzione $\mathbf{x}$ non è unica e dipende da un parametro h. Se si pone uguale ad h la terza componente x_3 di $\mathbf{x}$, le altre due si ottengono risolvendo il sistema

$$\begin{bmatrix} -449 & -128 \\ 0 & -45 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -243 \\ -117 \end{bmatrix} - h \begin{bmatrix} -772 \\ 360 \end{bmatrix}.$$

Inoltre è

$$\mathbf{c}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix},$$

e quindi $\gamma = \|\mathbf{c}_2\|_2 = \sqrt{2}$. ∎

La fattorizzazione QR può essere ricavata anche utilizzando il metodo di Givens, che può essere conveniente quando la matrice A è sparsa.

Il metodo QR consente infine di risolvere alcuni problemi come quello dell'esempio 7.4, in cui il sistema normale non può essere effettivamente risolto, a causa degli errori di rappresentazione degli elementi della matrice $A^H A$.

7.7 Esempio. Applicando il metodo QR al problema dell'esempio 7.4 si ha

$$T^{(1)} = \begin{bmatrix} 3 & 3 & 1 \\ 4 & 4 & 1 \\ 0 & \alpha & 1 \end{bmatrix},$$

e dopo 3 passi, operando con la precisione di macchina u, si ottiene

$$T^{(4)} = \begin{bmatrix} -5 & -5 & -\frac{7}{5} \\ 0 & -\alpha & -1 \\ 0 & 0 & \frac{1}{5} \end{bmatrix},$$

da cui si ricava

$$\mathbf{x}^* = \left[\frac{7}{25} - \frac{1}{\alpha}\ ,\ \frac{1}{\alpha} \right]^T . \qquad \blacksquare$$

3. Norme di matrici non quadrate

Il concetto di norma può essere esteso anche a matrici non quadrate. In particolare, se $A \in \mathbf{C}^{m\times n}$, dove m e n sono interi qualsiasi, le norme matriciali indotte, considerate nel capitolo 3, vengono definite per mezzo della relazione

$$\|A\| = \max_{\|\mathbf{x}\|'=1} \|A\mathbf{x}\|'',$$

dove $\|\ .\ \|'$ e $\|\ .\ \|''$ sono norme vettoriali rispettivamente su $\mathbf{C}^n$ e su $\mathbf{C}^m$.

Si può dimostrare, in modo analogo a quanto fatto nel caso delle matrici quadrate, che nel caso in cui le due norme vettoriali coincidano con la norma 2, la norma matriciale indotta che si ottiene è data da

$$\|A\|_2 = \sqrt{\rho(A^H A)}.$$

Anche per matrici non quadrate si definisce la norma di Frobenius di A nel modo seguente

$$\|A\|_F = \sqrt{\text{tr } (A^H A)}.$$

Inoltre se $U \in \mathbf{C}^{m\times m}$ e $V \in \mathbf{C}^{n\times n}$ sono matrici unitarie, poiché

$$(U^H AV)^H (U^H AV) = V^H A^H AV,$$

risulta

$$\|U^H AV\|_2 = \sqrt{\rho(A^H A)} = \|A\|_2$$

e

$$\|U^H AV\|_F = \sqrt{\text{tr } (A^H A)} = \|A\|_F.$$

4. Decomposizione ai valori singolari di una matrice

Lo studio della soluzione del problema dei minimi quadrati può essere condotto anche utilizzando la decomposizione ai valori singolari di una matrice.

7.8 Teorema. *Sia $A \in \mathbf{C}^{m\times n}$. Allora esistono una matrice unitaria $U \in \mathbf{C}^{m\times m}$ e una matrice unitaria $V \in \mathbf{C}^{n\times n}$ tali che*

$$A = U\Sigma V^H, \tag{13}$$

dove la matrice $\Sigma \in \mathbf{R}^{m\times n}$ ha elementi σ_{ij} nulli per $i \neq j$ e per $i = j$ ha elementi $\sigma_{ii} = \sigma_i$, con

$$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_p \geq 0, \quad p = \min\{m, n\}.$$

La decomposizione (13) è detta *decomposizione ai valori singolari* della matrice A, mentre i valori σ_i, per $i = 1, \ldots, p$, sono detti i *valori singolari* di A. Indicate con $\mathbf{u}_i$, $i = 1, \ldots, m$, e $\mathbf{v}_i$, $i = 1, \ldots, n$, le colonne rispettivamente di U e di V, i vettori $\mathbf{u}_i$ e $\mathbf{v}_i$, $i = 1, \ldots, p$, sono detti rispettivamente *vettori singolari sinistri e vettori singolari destri* di A. La matrice Σ è univocamente determinata, anche se le matrici U e V non lo sono.

Dim. Si considera per semplicità il caso $m \geq n$ (se $m < n$, si sostituisce A con A^H). Si procede dimostrando per induzione su n che la tesi vale per ogni $m \geq n$.

Per $n = 1$ è $A = \mathbf{a} \in \mathbf{C}^m$. Si pone $\sigma_1 = \|\mathbf{a}\|_2$ e si considera come matrice U la matrice di Householder tale che $U\mathbf{a} = \sigma_1 \mathbf{e}_1$. La matrice V è la matrice $V = [1]$.

Per $n > 1$ si dimostra che se la tesi vale per le matrici di $\mathbf{C}^{k\times(n-1)}$, con $k \geq n-1$, allora vale per le matrici di $\mathbf{C}^{m\times n}$, con $m \geq n$. Sia $\mathbf{x} \in \mathbf{C}^n$, tale che $\|\mathbf{x}\|_2 = 1$ e $\|A\|_2 = \|A\mathbf{x}\|_2$. Si consideri il vettore

$$\mathbf{y} = \frac{A\mathbf{x}}{\|A\mathbf{x}\|_2} \in \mathbf{C}^m.$$

Allora $\|\mathbf{y}\|_2 = 1$ e $A\mathbf{x} = \sigma_1 \mathbf{y}$, con $\sigma_1 = \|A\|_2$. Siano poi $V_1 \in \mathbf{C}^{n\times n}$ e $U_1 \in \mathbf{C}^{m\times m}$ matrici unitarie le cui prime colonne sono uguali rispettivamente a $\mathbf{x}$ e $\mathbf{y}$. Poiché

$$U_1^H AV_1 \mathbf{e}_1 = U_1^H A\mathbf{x} = U_1^H \sigma_1 \mathbf{y} = \sigma_1 [1, 0, \ldots, 0]^T,$$

è

$$A_1 = U_1^H A V_1 = \begin{bmatrix} \sigma_1 & \mathbf{w}^H \\ \mathbf{0} & B \end{bmatrix} \begin{matrix} \} & 1 \text{ riga} \\ \} & m-1 \text{ righe} \end{matrix}$$

in cui $\mathbf{w} \in \mathbf{C}^{n-1}$, $B \in \mathbf{C}^{(m-1)\times(n-1)}$ e $\mathbf{0} \in \mathbf{C}^{m-1}$. Si dimostra ora che $\mathbf{w} = \mathbf{0}$. Si supponga per assurdo che $\mathbf{w} \neq \mathbf{0}$ e si consideri il vettore $\mathbf{z} = \begin{bmatrix} \sigma_1 \\ \mathbf{w} \end{bmatrix} \neq \mathbf{0}$, per cui

$$A_1 \mathbf{z} = \begin{bmatrix} \sigma_1^2 + \mathbf{w}^H \mathbf{w} \\ B\mathbf{w} \end{bmatrix} = \begin{bmatrix} \|\mathbf{z}\|_2^2 \\ B\mathbf{w} \end{bmatrix}.$$

' Si ha

$$\|A_1 \mathbf{z}\|_2^2 = \|\mathbf{z}\|_2^4 + \|B\mathbf{w}\|_2^2 \geq \|\mathbf{z}\|_2^4,$$

da cui, dividendo per $\|\mathbf{z}\|_2^2$ si ottiene

$$\frac{\|A_1 \mathbf{z}\|_2^2}{\|\mathbf{z}\|_2^2} \geq \|\mathbf{z}\|_2^2,$$

e quindi

$$\|A_1\|_2^2 \geq \sigma_1^2 + \mathbf{w}^H \mathbf{w}. \tag{14}$$

D'altra parte è $\|A_1\|_2 = \|A\|_2$, in quanto A_1 è ottenuta da A con trasformazioni unitarie e quindi

$$\|A_1\|_2 = \sigma_1. \tag{15}$$

Dal confronto fra la (14) e la (15) segue l'assurdo. Quindi

$$A_1 = \begin{bmatrix} \sigma_1 & \mathbf{0}^H \\ \mathbf{0} & B \end{bmatrix}.$$

Dalla (15) segue che

$$\sigma_1 \geq \|B\|_2. \tag{16}$$

Infatti

$$\sigma_1^2 = \|A_1\|_2^2 = \rho(A_1^H A_1) = \rho\Big(\begin{bmatrix} \sigma_1^2 & \mathbf{0}^H \\ \mathbf{0} & B^H B \end{bmatrix} \Big) = \max\left[\sigma_1^2, \rho(B^H B)\right]$$
$$\geq \rho(B^H B) = \|B\|_2^2.$$

Poiché $B \in \mathbf{C}^{(m-1)\times(n-1)}$ e $m-1 \geq n-1$, per l'ipotesi induttiva si ha

$$U_2^H B V_2 = \Sigma_2,$$

dove le matrici $U_2 \in \mathbf{C}^{(m-1)\times(m-1)}$ e $V_2 \in \mathbf{C}^{(n-1)\times(n-1)}$ sono unitarie e $\Sigma_2 \in \mathbf{R}^{(m-1)\times(n-1)}$ ha elementi $\sigma_2 \geq \ldots \geq \sigma_p$. Poiché $\sigma_2 = \|B\|_2 \leq \sigma_1$ per la (16), la tesi segue con

$$U = U_1 \begin{bmatrix} 1 & \mathbf{0}^H \\ \mathbf{0} & U_2 \end{bmatrix}, \quad V = V_1 \begin{bmatrix} 1 & \mathbf{0}^H \\ \mathbf{0} & V_2 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \sigma_1 & \mathbf{0}^H \\ \mathbf{0} & \Sigma_2 \end{bmatrix}.$$

Le matrici U e V non sono univocamente determinate: infatti se

$$A = U \Sigma V^H$$

è una decomposizione ai valori singolari di A, e se $S \in \mathbf{C}^{n\times n}$ è una matrice di fase e $Z \in \mathbf{C}^{(m-n)\times(m-n)}$ è una matrice unitaria, anche

$$A = U \begin{bmatrix} S & O \\ O & Z \end{bmatrix} \Sigma S^H V^H$$

è una decomposizione ai valori singolari di A. Inoltre se $\sigma_i = \sigma_{i+1} = \ldots = \sigma_{i+j}$, per $j \geq 1$, detta P una qualunque matrice unitaria di ordine $j+1$ e considerata la matrice diagonale a blocchi

$$Q = \begin{bmatrix} I_{i-1} & O & O \\ O & P & O \\ O & O & I_{n-j-i} \end{bmatrix},$$

si ha che

$$A = U \begin{bmatrix} Q & O \\ O & I_{m-n} \end{bmatrix} \Sigma\, Q^H V^H$$

è una decomposizione ai valori singolari di A. ∎

Dal teorema 7.8 segue che

$$A = U \Sigma V^H = \sum_{i=1}^{p} \sigma_i \mathbf{u}_i \mathbf{v}_i^H \tag{17}$$

$$\left.\begin{aligned} A\mathbf{v}_i &= \sigma_i \mathbf{u}_i \\ A^H \mathbf{u}_i &= \sigma_i \mathbf{v}_i \end{aligned}\right\}, \quad i = 1, \ldots, p.$$

7.9 Esempio. La matrice A dell'esempio 7.2 ha la decomposizione ai valori singolari

$$A = U\Sigma V^H,$$

dove

$$U = \frac{1}{15}\begin{bmatrix} 2 & -4 & 14 & 3 \\ 8 & -1 & -4 & 12 \\ 11 & 8 & 2 & -6 \\ -6 & 12 & 3 & 6 \end{bmatrix}, \quad V = \frac{1}{9}\begin{bmatrix} -4 & 4 & 7 \\ 8 & 1 & 4 \\ 1 & 8 & -4 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}.$$

La matrice A dell'esempio 7.3 ha la decomposizione ai valori singolari

$$A = U\Sigma V^H,$$

dove

$$U = \frac{1}{15}\begin{bmatrix} -4 & 14 & 2 & 3 \\ -1 & -4 & 8 & 12 \\ 8 & 2 & 11 & -6 \\ 12 & 3 & -6 & 6 \end{bmatrix}, \quad V = \frac{1}{9}\begin{bmatrix} 4 & 7 & -4 \\ 1 & 4 & 8 \\ 8 & -4 & 1 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 5 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

∎

Dal teorema 7.8 segue il

7.10 Teorema. *Sia $A \in \mathbf{C}^{m\times n}$ e sia*

$$A = U\Sigma V^H$$

la sua decomposizione ai valori singolari, dove

$$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_k > \sigma_{k+1} = \ldots = \sigma_p = 0.$$

Allora valgono le seguenti proprietà

a) $$A = U_k \Sigma_k V_k^H = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^H, \quad \textit{dove}$$

$U_k \in \mathbf{C}^{m\times k}$ è la matrice le cui colonne sono $\mathbf{u}_1, \ldots, \mathbf{u}_k$,

$V_k \in \mathbf{C}^{n\times k}$ è la matrice le cui colonne sono $\mathbf{v}_1, \ldots, \mathbf{v}_k$,

$\Sigma_k \in \mathbf{R}^{k\times k}$ è la matrice diagonale i cui elementi principali sono

$$\sigma_1, \ldots, \sigma_k.$$

b) *Il nucleo di A è generato dai vettori $\mathbf{v}_{k+1}, \ldots, \mathbf{v}_n$.*

c) *L'immagine di A è generata dai vettori $\mathbf{u}_1, \ldots, \mathbf{u}_k$, e quindi*

$$\textit{rango di } A = k.$$

d) *σ_i^2, $i = 1, \ldots, p$, sono gli autovalori della matrice $A^H A$ (se $m < n$ i restanti autovalori sono nulli) e quindi*

$$\|A\|_F^2 = \sum_{i=1}^{k} \sigma_i^2,$$

$$\|A\|_2 = \sigma_1.$$

e) *Se $m = n$ e A è normale, allora $\sigma_i = |\lambda_i|$, $i = 1, \ldots, n$, dove i λ_i sono gli autovalori di A, e i vettori singolari destri e sinistri coincidono con gli autovettori di A.*

Dim. Si supponga per semplicità che $p = n \leq m$ (se $n > m$ si sostituisce A con A^H).

a) La matrice Σ ha la forma

$$\Sigma = \begin{bmatrix} \Sigma_k & O \\ O & O \end{bmatrix} \begin{matrix} \} & k \text{ righe} \\ \} & m-k \text{ righe} \end{matrix}$$

per cui, partizionando le matrici U e V nel modo seguente

$$U = [U_k \mid U'_{m-k}], \qquad V = [V_k \mid V'_{n-k}],$$

dalla (17) risulta che

$$A = U_k \Sigma_k V_k^H. \tag{18}$$

b) Se $\mathbf{x} \in \mathbf{C}^n$, la condizione $A\mathbf{x} = \mathbf{0}$ per la (17) è equivalente alla condizione

$$U \Sigma V^H \mathbf{x} = \mathbf{0},$$

e, poiché U è non singolare, è equivalente a

$$\Sigma V^H \mathbf{x} = \mathbf{0}. \tag{19}$$

Il vettore $\mathbf{z} = \Sigma V^H \mathbf{x}$ può essere partizionato nel modo seguente

$$\mathbf{z} = \begin{bmatrix} \Sigma_k V_k^H \mathbf{x} \\ \mathbf{0} \end{bmatrix} \begin{matrix} \} & k \text{ componenti,} \\ \} & m-k \text{ componenti,} \end{matrix} \tag{20}$$

per cui la (19) può essere scritta come $\Sigma_k V_k^H \mathbf{x} = \mathbf{0}$, ossia $V_k^H \mathbf{x} = \mathbf{0}$. Quindi $A\mathbf{x} = \mathbf{0}$ se e solo se $\mathbf{x}$ è ortogonale alle prime k colonne di V, ed essendo V unitaria, se e solo se $\mathbf{x}$ è generato dalle restanti colonne di V.

c) Dalla (18) si ha

$$\mathbf{y} = A\mathbf{x} = U_k \Sigma_k V_k^H \mathbf{x} = U_k \mathbf{z}, \tag{21}$$

dove $\mathbf{z} = \Sigma_k V_k^H \mathbf{x} \in \mathbf{C}^k$. Quindi $\mathbf{y}$ è generato dalle colonne di U_k. Viceversa dalla (20) si ha che, poiché la matrice $\Sigma_k V_k^H$ è di rango massimo, per ogni $\mathbf{x} \in \mathbf{C}^n, \mathbf{x} \neq \mathbf{0}$, esiste uno $\mathbf{z} \neq \mathbf{0}$ per cui vale la (21).

d) Dalla (17) si ha che

$$A^H A = V \Sigma^H \Sigma V^H,$$

dove $\Sigma^H \Sigma \in \mathbf{R}^{n \times n}$ è la matrice diagonale i cui elementi principali sono $\sigma_1^2, \ldots, \sigma_p^2$. Poiché la traccia e il raggio spettrale di due matrici simili sono uguali, si ha

$$\|A\|_F^2 = \text{tr } (A^H A) = \sum_{i=1}^{p} \sigma_i^2 = \sum_{i=1}^{k} \sigma_i^2$$

e

$$\|A\|_2^2 = \rho(A^H A) = \sigma_1^2,$$

e poiché $\sigma_1 > 0$, risulta $\|A\|_2 = \sigma_1$.

e) Se A è normale, dalla forma normale di Schur di A

$$A = UDU^H,$$

segue che

$$A^H A = UD^H DU^H,$$

perciò gli autovalori σ_i^2 di $A^H A$ sono tali che

$$\sigma_i^2 = \overline{\lambda}_i \lambda_i = |\lambda_i|^2, \quad \text{per } i = 1, \ldots, n.$$

▪

7.11 Esempio. La matrice A dell'esempio 7.2 ha rango 3 (infatti la matrice $A^H A$ è non singolare); come risulta dall'esempio 7.9 i suoi valori singolari sono $\sigma_1 = 3, \sigma_2 = 2, \sigma_3 = 1$. Per il punto d) del teorema 7.10 risulta

$$\|A\|_F = \sqrt{14}, \qquad \|A\|_2 = 3.$$

La matrice A dell'esempio 7.3 ha rango 2: infatti $\sigma_1 = 5, \sigma_2 = 1, \sigma_3 = 0$, come risulta dall'esempio 7.9. Per il punto d) del teorema 7.10 risulta

$$\|A\|_F = \sqrt{26}, \qquad \|A\|_2 = 1,$$

e l'insieme degli $\mathbf{x} \in \mathbf{R}^3$ tali che $A\mathbf{x} = \mathbf{0}$ è generato dal vettore

$$\mathbf{v}_3 = \frac{1}{9}\,[\text{-}4,\ 8,\ 1]^T. \qquad \blacksquare$$

7.12 Esempio. La matrice hermitiana

$$A = \frac{1}{81}\begin{bmatrix} \text{-}65 & 76 & 104 \\ 76 & \text{-}206 & 8 \\ 104 & 8 & 109 \end{bmatrix}$$

è tale che $A = UDU^H$, dove U, matrice unitaria, e D, matrice diagonale, sono

$$U = \frac{1}{9}\begin{bmatrix} \text{-}4 & 4 & 7 \\ 8 & 1 & 4 \\ 1 & 8 & \text{-}4 \end{bmatrix}, \quad D = \begin{bmatrix} \text{-}3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & \text{-}1 \end{bmatrix}.$$

Poiché gli autovalori di A sono $\lambda_1 = \text{-}3, \lambda_2 = 2, \lambda_3 = \text{-}1$, i valori singolari di A sono $\sigma_1 = 3, \sigma_2 = 2, \sigma_3 = 1$ e la decomposizione ai valori singolari di A è data da

$$A = U\Sigma V^H,$$

dove

$$V = \frac{1}{9}\begin{bmatrix} 4 & 4 & \text{-}7 \\ \text{-}8 & 1 & \text{-}4 \\ \text{-}1 & 8 & 4 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

$\blacksquare$

Dal teorema 7.10 si può ricavare anche un procedimento per calcolare i valori e i vettori singolari di A. Per semplicità si suppone $m \geq n$ (se fosse $m < n$ basta riferirsi alla matrice A^H). Questo procedimento si articola nei seguenti passi:

a) si calcolano gli autovalori e gli autovettori, normalizzati in norma 2, della matrice $A^H A$ e si considera la seguente decomposizione in forma normale di Schur della matrice $A^H A$

$$A^H A = QDQ^H, \quad D \in \mathbf{R}^{n\times n},\ Q \in \mathbf{C}^{n\times n}, \tag{22}$$

in cui gli elementi principali di D sono gli autovalori in ordine non crescente di $A^H A$ e Q è la corrispondente matrice degli autovettori (un

metodo stabile per calcolare la decomposizione (22) sarà descritto nel paragrafo 9);

b) si calcola la matrice

$$C = AQ \in \mathbf{C}^{m\times n}, \tag{23}$$

e si determina, utilizzando la tecnica del massimo pivot per colonne, la fattorizzazione QR della matrice

$$C\Pi = UR = U \begin{bmatrix} R_1 \\ \\ O \end{bmatrix}, \tag{24}$$

dove $\Pi \in \mathbf{R}^{n\times n}$ è una matrice di permutazione, $U \in \mathbf{C}^{m\times m}$ è una matrice unitaria ed $R_1 \in \mathbf{C}^{n\times n}$ è una matrice triangolare superiore con gli elementi principali reali non negativi e ordinati in modo non crescente. Le condizioni imposte sull'ordinamento degli elementi principali di R_1 rendono unica questa fattorizzazione se gli elementi principali di R_1 sono tutti distinti.

Da (23) e (24) si ottiene

$$A = UR\Pi^T Q^H, \tag{25}$$

da cui

$$A^H A = Q\Pi R^H U^H UR\Pi^T Q^H = Q\Pi R^H R\Pi^T Q^H = Q\Pi R_1^H R_1 \Pi^T Q^H,$$

e quindi per la (22) risulta

$$R_1^H R_1 = \Pi^T D\Pi.$$

Poiché la matrice $\Pi^T D\Pi$ risulta essere diagonale, ne segue che $R_1^H R_1$ è diagonale, e quindi R_1 non può che essere diagonale. Inoltre, poiché gli elementi principali di R_1 e di D sono ordinati in modo non crescente, se gli autovalori di $A^H A$ sono tutti distinti è $\Pi = I$. Quindi la (25), se si pone $\Sigma = R$ e $V = Q\Pi$, rappresenta la decomposizione ai valori singolari di A.

La decomposizione ai valori singolari di una matrice consente anche di risolvere il seguente problema di minimo: data una matrice $A \in \mathbf{C}^{m\times n}$ di rango k, e fissato un intero $r < k$, qual è la matrice $B \in \mathbf{C}^{m\times n}$ di rango r più "vicina" ad A? Vale infatti il seguente

7.13 Teorema. *Sia $A \in \mathbf{C}^{m\times n}$ e sia*

$$A = U\Sigma V^H$$

la decomposizione ai valori singolari di A, dove

$$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_k > \sigma_{k+1} = \ldots = \sigma_p = 0,$$

e sia r un intero positivo minore o uguale a k. Indicando con

$$A_r = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^H$$

e con

$$S = \{\ B \in \mathbf{C}^{m\times n}\ :\ \textit{rango di } B = r\ \},$$

si ha

$$\min_{B\in S} \|A - B\|_2 = \|A - A_r\|_2 = \sigma_{r+1}.$$

Dim. Sia $\Sigma_r \in \mathbf{R}^{m\times n}$ la matrice i cui elementi sono $\sigma_{ii} = \sigma_i$ per $i = 1, \ldots, r$ e $\sigma_{ij} = 0$ altrimenti. Allora vale

$$U^H A_r V = \Sigma_r$$

e quindi per il punto c) del teorema 7.10 è rango di $A_r = r$. Risulta inoltre

$$\|A - A_r\|_2 = \|U^H(A - A_r)V\|_2 = \|\Sigma - \Sigma_r\|_2 = \sigma_{r+1}, \tag{26}$$

in quanto σ_{r+1} è il massimo degli elementi non nulli di $\Sigma - \Sigma_r$. Sia $B \in S$. Il nucleo di B ha dimensione $n - r$ perché B ha rango r. Poiché l'intersezione fra $N(B)$ e il sottospazio T di $\mathbf{C}^n$ generato dai vettori $\mathbf{v}_1, \ldots, \mathbf{v}_r, \mathbf{v}_{r+1}$ non può ridursi al solo vettore nullo, in quanto dim $T+$ dim $N(B) = n + 1$, esiste un elemento $\mathbf{z} \in N(B) \cap T,\ \mathbf{z} \neq \mathbf{0}$. Si supponga che $\|\mathbf{z}\|_2 = 1$; essendo $\mathbf{z}$ elemento di T, si può scrivere

$$\mathbf{z} = \sum_{j=1}^{r+1} \alpha_j \mathbf{v}_j. \tag{27}$$

Per il punto a) del teorema 7.10 si ha

$$A\mathbf{z} = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i (\mathbf{v}_i^H \mathbf{z}) = \sum_{i=1}^{r+1} \sigma_i \mathbf{u}_i (\mathbf{v}_i^H \mathbf{z}), \tag{28}$$

in quanto, essendo $\mathbf{z} \in T$ e V unitaria, è $\mathbf{v}_i^H \mathbf{z} = 0$ per $i = r+2, \ldots, k$. Poiché $\mathbf{z} \in N(B)$, è $B\mathbf{z} = \mathbf{0}$ e si ha

$$\|A - B\|_2^2 \geq \|(A - B)\mathbf{z}\|_2^2 = \|A\mathbf{z}\|_2^2. \tag{29}$$

D'altra parte per la (28), poiché i vettori $\mathbf{u}_i$, $i = 1, \ldots, r+1$, sono ortonormali, si ha

$$\|A\mathbf{z}\|_2^2 = \sum_{i=1}^{r+1} \sigma_i^2 |\mathbf{v}_i^H \mathbf{z}|^2.$$

Poiché $\sigma_i^2 \geq \sigma_{r+1}^2$, $i = 1, \ldots, r+1$, si ha

$$\|A\mathbf{z}\|_2^2 \geq \sigma_{r+1}^2 \sum_{i=1}^{r+1} |\mathbf{v}_i^H \mathbf{z}|^2. \tag{30}$$

Per la (27) è

$$\begin{aligned} \sum_{i=1}^{r+1} |\mathbf{v}_i^H \mathbf{z}|^2 = \sum_{i=1}^{r+1} \left| \mathbf{v}_i^H \sum_{j=1}^{r+1} \alpha_j \mathbf{v}_j \right|^2 = \sum_{i=1}^{r+1} \left| \sum_{j=1}^{r+1} \alpha_j \mathbf{v}_i^H \mathbf{v}_j \right|^2 = \sum_{i=1}^{r+1} |\alpha_i|^2 \\ = \|\mathbf{z}\|_2^2 = 1, \end{aligned}$$

e quindi dalla (30) segue

$$\|A\mathbf{z}\|_2 \geq \sigma_{r+1}. \tag{31}$$

Confrontando la (31) e la (29) si ha che

$$\|A - B\|_2 \geq \sigma_{r+1},$$

e poiché per la (26) è $\|A - A_r\|_2 = \sigma_{r+1}$, ne segue che

$$\min_{B \in S} \|A - B\|_2 = \|A - A_r\|_2 = \sigma_{r+1}. \qquad \blacksquare$$

L'importanza del teorema 7.13 risiede nel fatto che esso consente di quantificare esattamente, tramite il valore singolare σ_{r+1}, la distanza in norma 2 della matrice A dalla "più vicina" matrice di rango r, e quindi di stimare l'errore che si commette quando la matrice A, a seguito di operazioni eseguite in aritmetica finita, viene sostituita con una matrice di rango r. Questa stima non è ottenibile facilmente utilizzando le fattorizzazioni LU e QR, che hanno lo scopo di trasformare la matrice A in una matrice della forma

$$\begin{bmatrix} B \\ \\ O \end{bmatrix},$$

in cui B è triangolare superiore: il rango di A viene ricavato mediante il numero di elementi principali di B che sono diversi da zero a meno di una precisione prefissata. Questo modo di procedere non garantisce assolutamente un buon risultato perché è molto difficile stimare bene il rango di una matrice triangolare. Si esamini a questo proposito il seguente esempio dovuto a Wilkinson.

7.14 Esempio. Si consideri la matrice triangolare superiore B di ordine n i cui elementi sono

$$b_{ij} = \begin{cases} 1 & \text{se } i = j, \\ -1 & \text{se } i < j, \\ 0 & \text{se } i > j \end{cases}$$

(la matrice B^T è stata usata nell'esercizio 4.19). La matrice ha rango n. Se però l'elemento di indici $(n, 1)$ viene perturbato della quantità $\epsilon = -2^{2-n}$, la matrice così ottenuta ha rango $n - 1$. Quindi una piccola perturbazione introdotta su un elemento non principale altera il rango della matrice triangolare B. Infatti se si calcolano i valori singolari di B si ottiene:

n	σ_{n-1}	σ_n
5	1.509442	$0.9298509\ 10^{-1}$
10	1.502146	$0.2929687\ 10^{-2}$
15	1.500909	$0.9170198\ 10^{-4}$
20	1.500494	$0.2969867\ 10^{-5}$

Quindi al crescere di n la matrice B è sempre più vicina ad una matrice di rango $n - 1$, anche se questo non appare evidente dagli elementi principali, che sono gli autovalori di B. ∎

5. Risoluzione del problema dei minimi quadrati con i valori singolari

Utilizzando il teorema 7.10 è possibile dare una formulazione esplicita della soluzione $\mathbf{x}^*$ di minima norma del problema dei minimi quadrati e del corrispondente γ, anche nel caso in cui la matrice A non sia di rango massimo.

7.15 Teorema. *Sia $A \in \mathbf{C}^{m\times n}$ di rango k, con $m \geq n \geq k$, e sia*

$$A = U\Sigma V^H$$

la decomposizione ai valori singolari di A. Allora la soluzione di minima norma del problema (2) è data da

$$\mathbf{x}^* = \sum_{i=1}^{k} \frac{\mathbf{u}_i^H \mathbf{b}}{\sigma_i} \mathbf{v}_i$$

e

$$\gamma^2 = \sum_{i=k+1}^{m} |\mathbf{u}_i^H \mathbf{b}|^2.$$

Dim. Poiché la norma 2 è invariante per trasformazioni unitarie, si ha

$$\|A\mathbf{x} - \mathbf{b}\|_2^2 = \|U^H(A\mathbf{x} - \mathbf{b})\|_2^2 = \|U^H AVV^H\mathbf{x} - U^H\mathbf{b}\|_2^2,$$

e posto $\mathbf{y} = V^H\mathbf{x}$, si ha

$$\|A\mathbf{x} - \mathbf{b}\|_2^2 = \|\Sigma\mathbf{y} - U^H\mathbf{b}\|_2^2 = \sum_{i=1}^{n} |\sigma_i y_i - \mathbf{u}_i^H\mathbf{b}|^2 + \sum_{i=n+1}^{m} |\mathbf{u}_i^H\mathbf{b}|^2$$

$$= \sum_{i=1}^{k} |\sigma_i y_i - \mathbf{u}_i^H\mathbf{b}|^2 + \sum_{i=k+1}^{m} |\mathbf{u}_i^H\mathbf{b}|^2, \tag{32}$$

dove y_i , $i = 1, \ldots, n$, sono le componenti di $\mathbf{y}$. Il minimo della (32) viene raggiunto per

$$y_i = \frac{\mathbf{u}_i^H\mathbf{b}}{\sigma_i}, \quad i = 1, \ldots, k. \tag{33}$$

Fra tutti i vettori $\mathbf{y} \in \mathbf{C}^n$ per cui vale la (33), il vettore di minima norma $\mathbf{y}^*$ è quello per cui

$$y_i^* = \begin{cases} \dfrac{\mathbf{u}_i^H\mathbf{b}}{\sigma_i}, & \text{per } i = 1, \ldots, k, \\ 0, & \text{per } i = k+1, \ldots, n. \end{cases}$$

Poiché $\mathbf{x} = V\mathbf{y}$, è $\|\mathbf{x}^*\|_2 = \|\mathbf{y}^*\|_2$ e quindi

$$\mathbf{x}^* = V\mathbf{y}^* = \sum_{i=1}^{k} y_i^*\mathbf{v}_i = \sum_{i=1}^{k} \frac{\mathbf{u}_i^H\mathbf{b}}{\sigma_i}\ \mathbf{v}_i,$$

e dalla (32) risulta

$$\gamma^2 = \|A\mathbf{x} - \mathbf{b}\|_2^2 = \sum_{i=k+1}^{m} |\mathbf{u}_i^H \mathbf{b}|^2. \qquad \blacksquare$$

7.16 Esempio. Dalla decomposizione ai valori singolari della matrice A dell'esempio 7.2

$$A = U\Sigma V^H,$$

dove U, V e Σ sono quelle riportate nell'esempio 7.9, si ha

$$\mathbf{x}^* = \frac{\mathbf{u}_1^H \mathbf{b}}{\sigma_1}\mathbf{v}_1 + \frac{\mathbf{u}_2^H \mathbf{b}}{\sigma_2}\mathbf{v}_2 + \frac{\mathbf{u}_3^H \mathbf{b}}{\sigma_3}\mathbf{v}_3,$$

e poiché $\mathbf{u}_i^H \mathbf{b} = 1$, per $i = 1, \ldots, 4$, risulta

$$\mathbf{x}^* = \frac{1}{3}\mathbf{v}_1 + \frac{1}{2}\mathbf{v}_2 + \mathbf{v}_3 = \frac{1}{54}\,[46,\ 43,\ 2]^T,$$

e $\quad \gamma = |\mathbf{u}_4^H \mathbf{b}| = 1.$

Dalla decomposizione ai valori singolari della matrice A dell'esempio 7.3

$$A = U\Sigma V^H,$$

dove U, V e Σ sono quelle riportate nell'esempio 7.9, si ha

$$\mathbf{x}^* = \frac{\mathbf{u}_1^H \mathbf{b}}{\sigma_1}\mathbf{v}_1 + \frac{\mathbf{u}_2^H \mathbf{b}}{\sigma_2}\mathbf{v}_2,$$

e poiché $\mathbf{u}_i^H \mathbf{b} = 1$, per $i = 1, \ldots, 4$, risulta

$$\mathbf{x}^* = \frac{1}{5}\mathbf{v}_1 + \mathbf{v}_2 = \frac{1}{15}\,[13,\ 7,\ \text{-}4]^T,$$

e $\quad \gamma = \sqrt{|\mathbf{u}_3^H \mathbf{b}|^2 + |\mathbf{u}_4^H \mathbf{b}|^2} = \sqrt{2}.$ $\blacksquare$

6. Pseudoinversa di Moore-Penrose

Se la matrice A è quadrata e non singolare, la soluzione del sistema (1) e del problema dei minimi quadrati (2) coincidono e possono essere espresse nella forma

$$\mathbf{x}^* = A^{-1}\mathbf{b},$$

per mezzo della matrice inversa A^{-1}. Il concetto di matrice inversa può essere esteso anche al caso di matrici A per cui A^{-1} non esiste. In questo caso si definisce una matrice pseudoinversa di A, indicata con il simbolo A^+, che consente di scrivere la soluzione di minima norma del problema (2) nella forma

$$\mathbf{x}^* = A^+\mathbf{b}.$$

7.17 Definizione. Sia $A \in \mathbf{C}^{m\times n}$ una matrice di rango k. La matrice $A^+ \in \mathbf{C}^{n\times m}$ tale che

$$A^+ = V\Sigma^+U^H,$$

dove $\Sigma^+ \in \mathbf{R}^{n\times m}$ è la matrice che ha elementi σ_{ij} nulli per $i \neq j$ e per $i = j$ ha elementi

$$\sigma_{ii}^+ = \begin{cases} \dfrac{1}{\sigma_i}, & \text{per } i = 1,\ldots,k, \\ 0, & \text{per } i = k+1,\ldots,p, \end{cases}$$

è detta *pseudoinversa di Moore-Penrose* di A. ∎

Valgono le seguenti proprietà (si veda l'esercizio 7.11):

a) La matrice $X = A^+$ è l'unica matrice di $\mathbf{C}^{n\times m}$ che soddisfa alle seguenti *equazioni di Moore-Penrose*:
   1) $AXA = A$,
   2) $XAX = X$,
   3) $(AX)^H = AX$,
   4) $(XA)^H = XA$.

b) Se il rango di A è massimo, allora

$$\begin{array}{ll} \text{se } m \geq n, & A^+ = (A^HA)^{-1}A^H, \\ \text{se } m \leq n, & A^+ = A^H(AA^H)^{-1}, \\ \text{se } m = n = \text{ rango di } A, & A^+ = A^{-1}. \end{array}$$

È immediato verificare che per il teorema 7.15 risulta

$$\mathbf{x}^* = A^+\mathbf{b}$$

e

$$\gamma = \|(I - AA^+)\mathbf{b}\|_2.$$

Inoltre A^+ è la soluzione dei seguenti problemi (si veda l'esercizio 7.29)

$$\min_{X\in\mathbf{C}^{n\times m}} \|AX - I_m\|_2,$$

$$\min_{X\in\mathbf{C}^{n\times m}} \|AX - I_m\|_F.$$

Utilizzando la matrice A^+ è anche possibile estendere il concetto di condizionamento alle matrici quadrate singolari e alle matrici non quadrate.

7.18 Definizione. Sia $A \in \mathbf{C}^{m\times n}$ una matrice di rango k. Si definisce *numero di condizionamento* di A il numero

$$\mu(A) = \|A\| \, \|A^+\|$$

dove $\| \, . \, \|$ è una qualsiasi norma matriciale. ∎

Si osservi che se la norma usata è la norma 2, per il teorema 7.10 d) è

$$\mu_2(A) = \frac{\sigma_1}{\sigma_k}, \tag{34}$$

ed inoltre, poiché la matrice $(A^H A)^+$ ha per valori singolari non nulli le quantità

$$\frac{1}{\sigma_i^2}, \quad i = 1, \ldots, k,$$

è

$$\mu_2(A^H A) = \frac{\sigma_1^2}{\sigma_k^2}, \tag{35}$$

Confrontando le (34) e (35), ne segue che

$$\mu_2(A^H A) = [\mu_2(A)]^2. \tag{36}$$

L'importanza della matrice pseudoinversa A^+ di A è essenzialmente di carattere teorico, in quanto il calcolo della pseudoinversa di una matrice può risultare molto instabile: infatti, se una piccola perturbazione degli elementi di A modifica il rango della matrice, si può generare una grossa perturbazione degli elementi di A^+.

7.19 Esempio. Siano

$$A = \frac{1}{15}\begin{bmatrix} 6 & -8 \\ 6 & -8 \\ 3 & -4 \end{bmatrix}, \quad \delta A = \frac{\epsilon}{15}\begin{bmatrix} 4 & 3 \\ -8 & -6 \\ 8 & 6 \end{bmatrix}, \quad \epsilon \neq 0.$$

La decomposizione ai valori singolari di A è

$$A = U\Sigma V^T, \quad U = \frac{1}{3}\begin{bmatrix} 2 & 1 & -2 \\ 2 & -2 & 1 \\ 1 & 2 & 2 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad V = \frac{1}{5}\begin{bmatrix} 3 & 4 \\ -4 & 3 \end{bmatrix},$$

e quindi la matrice A ha rango 1 e risulta

$$A^+ = V\Sigma^+U^T = V\begin{bmatrix}1 & 0 & 0\\ 0 & 0 & 0\end{bmatrix}U^T = \frac{1}{15}\begin{bmatrix}6 & 6 & 3\\ -8 & -8 & -4\end{bmatrix}.$$

La decomposizione ai valori singolari di $A+\delta A$ è

$$A+\delta A = U\Sigma' V^T, \quad \Sigma' = \begin{bmatrix}1 & 0\\ 0 & \epsilon\\ 0 & 0\end{bmatrix}, \quad U \text{ e } V \text{ come sopra,}$$

e quindi la matrice $A+\delta A$ ha rango 2 e risulta

$$(A+\delta A)^+ = V{\Sigma'}^+U^T = V\begin{bmatrix}1 & 0 & 0\\ 0 & 1/\epsilon & 0\end{bmatrix}U^T = A^+ + \frac{1}{15\epsilon}\begin{bmatrix}4 & -8 & 8\\ 3 & -6 & 6\end{bmatrix}.$$

La perturbazione δA è tale che $\|\delta A\|_2 = |\epsilon|$ e genera sugli elementi di A^+ una perturbazione la cui norma 2 è

$$\|(A+\delta A)^+ - A^+\|_2 = \frac{1}{15|\epsilon|}\left\|\begin{bmatrix}4 & -8 & 8\\ 3 & -6 & 6\end{bmatrix}\right\|_2 = \frac{1}{|\epsilon|}.$$

Se per esempio $\epsilon = 10^{-6}$, una perturbazione δA tale che $\|\delta A\|_2 = 10^{-6}$ produce una perturbazione sugli elementi della pseudoinversa tale che $\|(A+\delta A)^+ - A^+\|_2 = 10^6$. ∎

Nel caso in cui il rango della matrice $A+\delta A$ non sia diverso da quello di A, la matrice A^+ risulta affetta da una perturbazione assai minore che nel caso precedente. Vale infatti il seguente teorema, per la cui dimostrazione si veda [12].

7.20 Teorema. *Siano* $A,\ \delta A \in \mathbf{C}^{m\times n}$, *tali che* $\|A^+\|_2\|\delta A\|_2 < 1$ *e rango di* $(A+\delta A) \le$ *rango di* A. *Allora*

$$\textit{rango di } (A+\delta A) = \textit{ rango di } A$$

e

$$\frac{\|(A+\delta A)^+ - A^+\|_2}{\|A^+\|_2} \le \alpha\frac{\mu_2(A)\epsilon_A}{1-\mu_2(A)\epsilon_A},$$

dove $\epsilon_A = \dfrac{\|\delta A\|_2}{\|A\|_2}$ *e* α *è una costante positiva minore di 2.*

∎

7. Condizionamento del problema dei minimi quadrati

Nel caso in cui la matrice A sia di rango massimo, vale il seguente teorema di perturbazione del problema dei minimi quadrati (per la dimostrazione si veda [12]).

7.21 Teorema. *Siano $m \geq n$, $A \in \mathbf{C}^{m\times n}$ una matrice di rango massimo, $\delta A \in \mathbf{C}^{m\times n}$, tale che $\|A^+\|\,\|\delta A\| < 1, \mathbf{b} \in \mathbf{C}^m, \mathbf{b} \neq \mathbf{0}$ e $\delta\mathbf{b} \in \mathbf{C}^m$. Allora la matrice $A+\delta A$ è ancora di rango massimo e, indicata con $\mathbf{x}+\delta\mathbf{x}$ la soluzione del problema dei minimi quadrati perturbato*

$$\min_{\mathbf{y}\in\mathbf{C}^n} \|(A+\delta A)\mathbf{y} - (\mathbf{b}+\delta\mathbf{b})\|_2,$$

risulta

$$\frac{\|\delta\mathbf{x}\|_2}{\|\mathbf{x}\|_2} \leq \frac{\mu_2(A)}{1-\epsilon_A\mu_2(A)}\left[\left(1+\mu_2(A)\frac{\gamma}{\|A\|_2\|\mathbf{x}\|_2}\right)\epsilon_A + \frac{\|\mathbf{b}\|_2}{\|A\|_2\|\mathbf{x}\|_2}\,\epsilon_b\right], \quad (37)$$

dove γ è il residuo del problema definito in (2) e

$$\epsilon_A = \frac{\|\delta A\|_2}{\|A\|_2}, \qquad \epsilon_b = \frac{\|\delta\mathbf{b}\|_2}{\|\mathbf{b}\|_2}.$$ ∎

Questo teorema consente di trarre alcune indicazioni riguardo al metodo da scegliere per risolvere il problema dei minimi quadrati, in modo da non pregiudicare l'accuratezza della soluzione. Quando la matrice A ha rango massimo con il metodo QR, effettuata la fattorizzazione (6), si risolve il sistema (9), la cui matrice R_1 è tale che

$$\mu_2(R_1) = \mu_2(R) = \mu_2(A),$$

cioè la stabilità del metodo è legata a $\mu_2(A)$. Se invece la soluzione del problema (2) viene ottenuta attraverso il sistema delle equazioni normali, la cui matrice è A^HA, la stabilità del metodo è legata a $\mu_2(A^HA)$, che per la (36) è uguale a $[\mu_2(A)]^2$.

D'altra parte dalla (37), che si può anche assumere come maggiorazione dell'errore inerente, segue che l'errore ϵ_A di perturbazione della matrice A è amplificato dal fattore

$$\frac{c_A}{1-\epsilon_A\mu_2(A)}, \quad \text{dove } c_A = \mu_2(A) + [\mu_2(A)]^2\frac{\gamma}{\|A\|_2\|\mathbf{x}\|_2}.$$

Se il residuo γ è così piccolo, che il secondo termine di c_A risulta minore del primo, allora la maggiorazione dell'errore inerente è dominata da $\mu_2(A)$. In questo caso non è conveniente ricorrere alla risoluzione del sistema normale, il cui condizionamento è $[\mu_2(A)]^2$, e conviene usare il metodo QR. Se invece il residuo γ è tale che il secondo termine di c_A risulta dominante, la maggiorazione dell'errore inerente è dominata dal termine $[\mu_2(A)]^2$. Allora dal punto di vista della stabilità, il metodo QR e i metodi basati sulla risoluzione del sistema normale, sono equivalenti. In questo caso, e nel caso in cui non sia possibile determinare a priori se per un dato problema di

minimi quadrati il residuo γ è piccolo oppure no, per la scelta del metodo occorre tenere presente:

a) il costo computazionale, che in generale è maggiore per il metodo QR;

b) eventuali proprietà di struttura delle matrici A e $A^H A$, per esempio eventuale sparsità della matrice A che non si trasmette alla matrice $A^H A$, per cui il costo computazionale del metodo QR è minore;

c) la disponibilità di memoria centrale del calcolatore che, quando m sia molto più grande di n, può essere sufficiente per contenere la matrice $A^H A$ ma non la matrice A;

d) la precisione di macchina, che può essere sufficiente a rappresentare gli elementi della matrice A ma non quelli della matrice $A^H A$ (si veda l'esempio 7.4, in cui nel calcolo di $A^H A$ si perdono informazioni fondamentali). Inoltre gli elementi di $A^H A$ potrebbero non essere rappresentabili a causa di overflow o di underflow.

7.22 Esempio. Sia $A \in \mathbf{R}^{m\times n}$ la matrice con elementi

$$a_{ij} = \lambda_i^{j-1}, \quad j = 1, \ldots, n,$$

dove i numeri $\lambda_i,\ i = 1, \ldots, m$ sono a due a due distinti. Questa matrice, che interviene nell'interpolazione di funzioni, ha rango massimo e se $m = n$ è detta matrice di *Vandermonde*.

Gli elementi della matrice $C = A^T A$ sono

$$c_{kj} = \sum_{i=1}^{m} \lambda_i^{k+j-2}, \quad k,\ j = 1, \ldots, n.$$

Tale matrice appartiene alla classe delle matrici di *Hankel*, i cui elementi sono definiti da $2n-1$ parametri $\alpha_k,\ k = 1, \ldots, 2n-1$, nel modo seguente

$$h_{ij} = \alpha_{i+j-1},\ i = 1, \ldots, n,\ j = 1, \ldots, n$$

(la matrice dell'esempio 4.21 è una matrice di Hankel). In questo caso per risolvere il problema (2) è conveniente, dal punto di vista del costo computazionale, passare attraverso la risoluzione del sistema normale, perché la costruzione della matrice $A^T A$ e del vettore $A^T \mathbf{b}$ richiede un numero di operazioni moltiplicative dell'ordine di $2mn$, ed inoltre esistono metodi per risolvere sistemi lineari con matrici di Hankel che hanno un costo computazionale dell'ordine di $n \log_2^2 n$.

Se però i numeri λ_i sono troppo piccoli o troppo grandi, è possibile che gli elementi c_{kj} non siano rappresentabili a causa di overflow o di underflow. In tal caso è necessario ricorrere al metodo QR. ∎

Se la matrice A non ha rango massimo, allora la risoluzione del problema (2), e in particolare il calcolo della soluzione di minima norma, è molto più delicata. La difficoltà più grossa si incontra nella determinazione del rango della matrice A. Quando la matrice A è fortemente mal condizionata, solo la determinazione dei valori singolari di A consente un'effettiva comprensione della natura del problema. Infatti:

a) se la matrice A e il vettore $\mathbf{b}$ sono ottenuti da dati sperimentali, per l'effetto congiunto dell'incertezza sui dati e del mal condizionamento della matrice si possono ottenere più soluzioni del problema (2) che, pur essendo molto diverse fra di loro, hanno tutte norma molto vicina alla norma minima e quindi sono tutte accettabili come approssimazioni della soluzione di minima norma.

b) Se il malcondizionamento della matrice A è generato dalla presenza di un certo numero di valori singolari vicini fra di loro e molto più piccoli degli altri, può essere utile considerare il problema approssimato che si ottiene eliminando i valori singolari più piccoli ed esprimere la soluzione utilizzando solo i valori singolari più grossi e i corrispondenti vettori singolari. Il valore del residuo γ che si ottiene operando in questo modo può dare una misura di quanto la soluzione calcolata è accettabile.

c) il calcolo dei valori e dei vettori singolari di una matrice è un problema ben posto, come si vedrà nel prossimo paragrafo, e il metodo per calcolare i valori e i vettori singolari che fa uso dell'algoritmo di Golub e Reinsch, descritto nel paragrafo 9 è stabile [9].

È opportuno rilevare però che il costo computazionale del calcolo dei valori e dei vettori singolari risulta in generale superiore a quello richiesto dai metodi diretti, come il metodo di Cholesky applicato al sistema normale o il metodo QR. Lawson e Hanson [12] stimano che il costo computazionale richiesto per la risoluzione del problema (2) tramite il calcolo dei valori singolari sia circa il doppio di quello del metodo QR quando m è molto più grande di n, e fino a 9 volte quello del metodo QR quando m è poco più grande di n.

8. Teoremi di perturbazione per i valori singolari

A differenza di quanto avviene per il problema del calcolo degli autovalori di una matrice, il calcolo dei valori singolari è un problema sempre ben condizionato, perché, come si vedrà, piccole perturbazioni degli elementi della matrice inducono nei risultati perturbazioni non superiori a quelle dei dati.

Sia $A \in \mathbf{C}^{m\times n}$, con $m \geq n$ (se $m < n$ le considerazioni che seguono si

applicano ad A^H) e si consideri la matrice hermitiana

$$B = \begin{bmatrix} O & A \\ A^H & O \end{bmatrix} \quad \in \mathbf{C}^{(m+n)\times(m+n)}. \tag{38}$$

Sia $A = U\Sigma V^H$ la decomposizione ai valori singolari di A e siano

$$U = [U_1 \mid U_2], \quad \Sigma = \begin{bmatrix} \Sigma_1 \\ O \end{bmatrix},$$

dove $U_1 \in \mathbf{C}^{m\times n}$, $U_2 \in \mathbf{C}^{m\times(m-n)}$ e $\Sigma_1 \in \mathbf{R}^{n\times n}$. La matrice

$$Z = \frac{1}{\sqrt{2}} \begin{bmatrix} U_1 & U_1 & \sqrt{2}\,U_2 \\ V & -V & O \end{bmatrix} \quad \in \mathbf{C}^{(m+n)\times(m+n)},$$

è unitaria e tale che

$$B = \begin{bmatrix} O & A \\ A^H & O \end{bmatrix} = Z \begin{bmatrix} \Sigma_1 & O & O \\ O & -\Sigma_1 & O \\ O & O & O \end{bmatrix} Z^H,$$

e quindi le colonne di Z costituiscono un insieme ortonormale di autovettori di B. La matrice B ha come autovalori i numeri σ_i e $-\sigma_i$, $i = 1, \ldots, n$, dove σ_i sono i valori singolari di A, oltre ad $m-n$ autovalori nulli, poiché $m \geq n$; se il rango k di A è minore di n, la molteplicità algebrica dell'autovalore nullo è $m + n - 2k$. I seguenti teoremi derivano dai corrispondenti teoremi di perturbazione degli autovalori.

7.23 Teorema. *Siano $A \in \mathbf{C}^{m\times n}$ e $\widehat{A} \in \mathbf{C}^{m\times(n-1)}$ la matrice ottenuta da A eliminando una colonna. Se σ_i, $i = 1, \ldots, \min\{m, n\}$, sono i valori singolari di A e τ_i, $i = 1, \ldots, \min\{m, n-1\}$, sono i valori singolari di $\widehat{A}$, risulta*

$$\text{se } m \geq n \qquad \sigma_1 \geq \tau_1 \geq \sigma_2 \geq \ldots \geq \tau_{n-1} \geq \sigma_n \geq 0,$$

$$\text{se } m < n \qquad \sigma_1 \geq \tau_1 \geq \sigma_2 \geq \ldots \geq \sigma_m \geq \tau_m \geq 0.$$

Dim. Si consideri la matrice $\widehat{B} \in \mathbf{C}^{(m+n-1)\times(m+n-1)}$ ottenuta a partire da $\widehat{A}$, in modo analogo alla matrice B della (38). Quindi $\widehat{B}$ si può ottenere da B eliminando una colonna e la riga corrispondente. Per il teorema 6.10 gli autovalori di $\widehat{B}$ separano quelli di B. ∎

7.24 Teorema. *Siano A e $\delta A \in \mathbf{C}^{m\times n}$. Se σ_i, τ_i e ψ_i, $i = 1, \ldots, n$, sono i valori singolari di A, di δA e di $A + \delta A$, risulta*

$$|\psi_i - \sigma_i| \leq \tau_1 = \|\delta A\|_2, \quad i = 1, \ldots, n.$$

Dim. Se $m \geq n$, la matrice

$$\begin{bmatrix} O & \delta A \\ \delta A^H & O \end{bmatrix},$$

ha autovalori

$$\tau_1 \geq \tau_2 \geq \ldots \geq \tau_n \geq -\tau_n \geq \ldots \geq -\tau_1,$$

oltre a $m - n$ autovalori nulli. Per il teorema 6.14 segue che

$$\sigma_i - \tau_1 \leq \psi_i \leq \sigma_i + \tau_1,$$

da cui

$$|\psi_i - \sigma_i| \leq \tau_1 = \|\delta A\|_2, \quad i = 1, \ldots, n.$$

Se $m < n$, la dimostrazione può essere condotta in modo analogo. ∎

Da questo teorema risulta che la perturbazione generata sui valori singolari da una perturbazione δA sugli elementi della matrice A è limitata superiormente da $\|\delta A\|_2$.

7.25 Esempio. Siano

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad \delta A = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \epsilon & 0 & 0 & 0 \end{bmatrix}.$$

Per il teorema 7.24 i valori singolari σ_i e ψ_i, $i = 1, \ldots, 4$ rispettivamente delle matrici A e $A + \delta A$ verificano la relazione

$$|\psi_i - \sigma_i| \leq |\epsilon|, \quad i = 1, \ldots, 4.$$

Gli autovalori λ_i e μ_i, $i = 1, \ldots, 4$ di A e di $A + \delta A$ verificano invece la relazione (si veda l'esempio 6.6)

$$|\lambda_i - \mu_i| \leq \sqrt[4]{|\epsilon|}.$$

In effetti i valori singolari di A sono $\sigma_1 = \sigma_2 = \sigma_3 = 1$, $\sigma_4 = 0$ e quelli di $A + \delta A$ sono $\psi_1 = \psi_2 = \psi_3 = 1$, $\psi_4 = |\epsilon|$, mentre gli autovalori di A sono nulli e gli autovalori di $A + \delta A$ sono i μ_i tali che $\mu_i^4 = \epsilon$, $i = 1, \ldots, 4$.

Ad esempio, nel caso che $\epsilon = 10^{-8}$, introducendo una perturbazione sull'elemento a_{41} di modulo pari ad ϵ, i valori singolari risultano affetti da un perturbazione uguale, mentre gli autovalori risultano affetti da una perturbazione di modulo pari ad 10^{-2}. ∎

Questo esempio illustra come il problema del calcolo dei valori singolari di una matrice risulti essere ben posto, anche se l'associato problema del calcolo degli autovalori è mal posto, caso che può presentarsi quando la matrice non è diagonalizzabile.

Se la matrice è normale autovalori e valori singolari hanno lo stesso modulo. Se la matrice non è normale la differenza fra i moduli degli autovalori e i valori singolari può essere arbitrariamente grande, come risulta anche dall'esempio 7.25. Vale il seguente teorema.

7.26 Teorema. *Sia $A \in \mathbf{C}^{n\times n}$. Per ogni autovalore λ di A vale*

$$\sigma_n \leq |\lambda| \leq \sigma_1.$$

Dim. Sia $A\mathbf{x} = \lambda\mathbf{x}$, $\mathbf{x} \neq \mathbf{0}$. Allora

$$\mathbf{x}^H A^H A\mathbf{x} = |\lambda|^2 \|\mathbf{x}\|_2^2. \tag{39}$$

Dalla decomposizione ai valori singolari di A segue che

$$A^H A = V \Sigma^2 V^H,$$

in cui $V \in \mathbf{C}^{n\times n}$ è unitaria, e quindi posto $\mathbf{y} = V^H\mathbf{x}$, è

$$\mathbf{x}^H A^H A\mathbf{x} = \mathbf{y}^H \Sigma^2 \mathbf{y} = \sum_{i=1}^{n} \sigma_i^2 |y_i|^2,$$

e poiché

$$\sigma_n^2 \|\mathbf{y}\|_2^2 \leq \sum_{i=1}^{n} \sigma_i^2 |y_i|^2 \leq \sigma_1^2 \|\mathbf{y}\|_2^2$$

e $\|\mathbf{y}\|_2^2 = \|\mathbf{x}\|_2^2$, per la (39) risulta

$$\sigma_n \leq |\lambda| \leq \sigma_1.$$ ∎

Se la matrice $A \in \mathbf{C}^{n\times n}$ è non singolare, il suo numero di condizionamento in norma 2 è per la (34)

$$\mu_2(A) = \frac{\sigma_1}{\sigma_n},$$

mentre per il teorema 7.26 risulta in generale

$$\mu_2(A) \geq \frac{\lambda_1}{\lambda_n},$$

con il segno di uguaglianza se A è normale. Quindi una matrice non normale può essere malcondizionata anche se il rapporto fra il massimo e il minimo modulo degli autovalori non è grande.

Queste considerazioni hanno suggerito studi e ricerche per individuare matrici "quasi normali", cioè con autovalori "vicini" in modulo ai valori singolari, e per migliorare quindi il condizionamento del problema del calcolo degli autovalori.

7.27 Esempio. La matrice triangolare B dell'esempio 7.14 ha gli autovalori tutti uguali a 1. Però al cresce di n il minimo valore singolare tende a zero, per cui la matrice risulta mal condizionata, anche per valori non troppo grossi di n. Al variare di n il numero di condizionamento $\mu_2(A) = \sigma_1/\sigma_n$ è riportato nella seguente tabella

n	$\mu_2(A)$
5	$2.942748\ 10^1$
10	$1.918453\ 10^3$
15	$9.512388\ 10^4$
20	$3.996832\ 10^6$

∎

9. Calcolo della forma normale di Schur di A^HA

Per quanto visto nel paragrafo 4, gli autovalori di A^HA sono i quadrati dei valori singolari $\sigma_1, \ldots, \sigma_n$ di A e il calcolo della decomposizione ai valori singolari di A richiede, come primo passo, che si calcolino gli autovalori e gli autovettori di A^HA. Per semplificare questo primo passo conviene trasformare prima la matrice A in una matrice B bidiagonale superiore a elementi reali, con il seguente algoritmo di *Golub e Reinsch*, che utilizza una successione di trasformazioni unitarie. Se la matrice A è sparsa può essere

conveniente utilizzare trasformazioni di Givens al posto di quelle di Householder. Per semplicità si suppone $m \geq n$ (se $m < n$ si applica l'algoritmo alla matrice A^H).

Posto $A^{(1)} = A$, si costruisce la prima matrice elementare di Householder $P^{(1)}$ in modo che la matrice

$$A^{(2)} = P^{(1)} A^{(1)}$$

abbia nulli gli elementi della prima colonna al di sotto di quello principale. La matrice $A^{(2)}$ risulta

$$A^{(2)} = \begin{bmatrix} \alpha & \mathbf{c}^H \\ \mathbf{0} & B^{(2)} \end{bmatrix},$$

in cui $\mathbf{c} \in \mathbf{C}^{n-1}$. Si costruisce poi la matrice elementare di Householder $K^{(1)} \in \mathbf{C}^{(n-1)\times(n-1)}$ tale che il vettore $K^{(1)}\mathbf{c}$ abbia nulle tutte le componenti di indice maggiore o uguale a 2. Indicata con $H^{(1)}$ la matrice elementare di Householder

$$H^{(1)} = \begin{bmatrix} 1 & \mathbf{0}^H \\ \mathbf{0} & K^{(1)} \end{bmatrix},$$

la matrice $A^{(3)} = A^{(2)} H^{(1)}$ ha nulli gli elementi della prima riga che hanno indice di colonna maggiore o uguale a 3 e gli elementi della prima colonna che hanno indice di riga maggiore o uguale a 2. Si ripete il procedimento per $n-2$ volte, annullando alternativamente gli elementi delle colonne e delle righe, e proseguendo poi sull'ultima colonna se $m > n$, come indicato nella figura 7.3 per il caso particolare $n = 4,\ m > 4$.

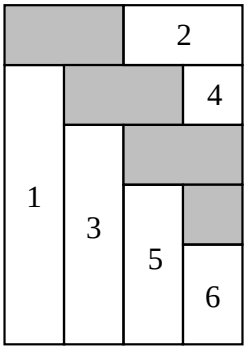


Fig. 7.3 - Riduzione a forma bidiagonale.

La successione $A^{(1)} = A, A^{(2)}, \ldots, A^{(2n-1)}$ viene costruita nel modo seguente:

$$\left.\begin{aligned} A^{(2k)} &= P^{(k)} A^{(2k-1)} \\ A^{(2k+1)} &= A^{(2k)} H^{(k)} \end{aligned}\right\} \quad k = 1, 2, \ldots, n-2$$

$$A^{(2n-2)} = P^{(n-1)} A^{(2n-3)}$$

$$A^{(2n-1)} = P^{(n)} A^{(2n-2)} \quad \text{se } m > n,$$

dove $P^{(k)} \in \mathbf{C}^{m\times m}$, $k = 1, 2, \ldots, n$, è una matrice elementare di Householder che annulla gli elementi della k-esima colonna di $A^{(2k-1)}$ con indice di riga maggiore o uguale a $k+1$ (se $m = n$ è $P^{(n)} = I$), e $H^{(k)} \in \mathbf{C}^{n\times n}$, $k = 1, 2, \ldots, n-2$, è una matrice elementare di Householder che annulla gli elementi della k-esima riga di $A^{(2k)}$ con indice di colonna maggiore o uguale a $k+2$. Dopo $2n-2$ passi risulta quindi

$$P^{(n)} \ldots P^{(2)} P^{(1)} A H^{(1)} H^{(2)} \ldots H^{n-2}$$

$$= \begin{bmatrix} \alpha_1 & \beta_1 & & & \\ & \alpha_2 & \beta_2 & & \\ & & \ddots & \ddots & \\ & & & \ddots & \beta_{n-1} \\ & & & & \alpha_n \\ & & & & \end{bmatrix} \begin{matrix} \left.\begin{matrix} \\ \\ \\ \\ \\ \end{matrix}\right\} n \text{ righe} \\ \left.\begin{matrix} \\ \end{matrix}\right\} m-n \text{ righe} \end{matrix}$$

in cui α_i, $i = 1, \ldots, n$ e β_i, $i = 1, \ldots, n-1$, sono in generale numeri complessi.

Se gli α_i e i β_i non sono tutti reali, esistono due matrici di fase S e T (si veda l'esercizio 7.35) tali che

$$S \begin{bmatrix} \alpha_1 & \beta_1 & & & \\ & \alpha_2 & \beta_2 & & \\ & & \ddots & \ddots & \\ & & & \ddots & \beta_{n-1} \\ & & & & \alpha_n \end{bmatrix} T = \begin{bmatrix} |\alpha_1| & |\beta_1| & & & \\ & |\alpha_2| & |\beta_2| & & \\ & & \ddots & \ddots & \\ & & & \ddots & |\beta_{n-1}| \\ & & & & |\alpha_n| \end{bmatrix}.$$

Posto

$$P = \begin{bmatrix} S & O \\ O & I_{m-n} \end{bmatrix} P^{(n)} \ldots P^{(2)} P^{(1)} \in \mathbf{C}^{m\times m}$$

e

$$H = H^{(1)} H^{(2)} \ldots H^{n-2} T \in \mathbf{C}^{n\times n},$$

risulta allora

$$PAH = \begin{bmatrix} B \\ \\ O \end{bmatrix},$$

dove $B \in \mathbf{R}^{n\times n}$ è una matrice bidiagonale superiore.

L'algoritmo di Golub e Reinsch richiede

$$2mn^2 - \frac{2}{3}n^3$$

operazioni moltiplicative [9].

Se $m > \frac{5}{3}n$ si riduce il costo computazionale se, prima di applicare il metodo di Golub e Reinsch, la matrice A viene fattorizzata nella forma QR [3].

7.28 Esempio. Sia

$$A = \frac{1}{100}\begin{bmatrix} \text{-}50 & 230 & 235 \\ 50 & \text{-}142 & 81 \\ 50 & 38 & \text{-}159 \\ 100 & \text{-}4 & 122 \\ \text{-}150 & 126 & \text{-}343 \end{bmatrix}.$$

Posto $A^{(1)} = A$, con l'algoritmo di Golub e Reinsch si ha

$$P^{(1)} = I - \beta_1 \mathbf{v}_1 \mathbf{v}_1^T, \text{ con } \mathbf{v}_1 = \frac{1}{2}\,[\text{-}5,\ 1,\ 1,\ 2,\ \text{-}3]^T,\ \beta_1 = \frac{1}{5},$$

$$A^{(2)} = P^{(1)}A^{(1)} = \frac{1}{5}\begin{bmatrix} 10 & \text{-}9 & 12 \\ 0 & \text{-}3 & 4 \\ 0 & 6 & \text{-}8 \\ 0 & 8 & 6 \\ 0 & \text{-}6 & \text{-}17 \end{bmatrix},$$

$$H^{(1)} = I - \gamma_1 \mathbf{w}_1 \mathbf{w}_1^T, \text{ con } \mathbf{w}_1 = \frac{12}{5}\,[0,\ \text{-}2,\ 1]^T,\ \gamma_1 = \frac{5}{72},$$

$$A^{(3)} = A^{(2)}H^{(1)} = \begin{bmatrix} 2 & 3 & 0 \\ 0 & 1 & 0 \\ 0 & \text{-}2 & 0 \\ 0 & 0 & 2 \\ 0 & \text{-}2 & \text{-}3 \end{bmatrix},$$

$$P^{(2)} = I - \beta_2 \mathbf{v}_2 \mathbf{v}_2^T, \text{ con } \mathbf{v}_2 = [0,\ 4,\ \text{-}2,\ 0,\ \text{-}2]^T,\ \beta_2 = \frac{1}{12},$$

$$A^{(4)} = P^{(2)}A^{(3)} = \begin{bmatrix} 2 & 3 & 0 \\ 0 & -3 & -2 \\ 0 & 0 & 1 \\ 0 & 0 & 2 \\ 0 & 0 & -2 \end{bmatrix},$$

$$P^{(3)} = I - \beta_3 \mathbf{v}_3 \mathbf{v}_3^T, \text{ con } \mathbf{v}_3 = [0,\ 0,\ 4,\ 2,\ -2]^T,\ \beta_3 = \frac{1}{12},$$

$$A^{(5)} = P^{(3)}A^{(4)} = \begin{bmatrix} 2 & 3 & 0 \\ 0 & -3 & -2 \\ 0 & 0 & -3 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

da cui si ha

$$A = P^T \begin{bmatrix} B \\ \\ O \end{bmatrix} H^T,$$

dove

$$P = P^{(3)}P^{(2)}P^{(1)} = \frac{1}{180} \begin{bmatrix} -45 & 45 & 45 & 90 & -135 \\ -75 & -45 & 135 & 30 & 75 \\ -145 & 69 & -51 & -62 & 13 \\ -35 & -33 & -93 & 134 & 59 \\ 50 & 150 & 30 & 40 & 70 \end{bmatrix},$$

$$B = \begin{bmatrix} 2 & 3 & 0 \\ 0 & -3 & -2 \\ 0 & 0 & -3 \end{bmatrix},$$

$$H = H^{(1)} = \frac{1}{5} \begin{bmatrix} 5 & 0 & 0 \\ 0 & -3 & 4 \\ 0 & 4 & 3 \end{bmatrix}.$$

■

Dopo aver applicato l'algoritmo di Golub e Reinsch la matrice A risulta quindi della forma

$$A = P^H \begin{bmatrix} B \\ \\ O \end{bmatrix} H^H,$$

in cui P e H sono matrici unitarie e $B \in \mathbf{R}^{n\times n}$ è bidiagonale superiore e poiché

$$A^H A = H B^T B H^H, \tag{40}$$

gli autovalori di B^TB sono ancora $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$. Per calcolare $\sigma_1, \sigma_2, \ldots,$ σ_n basta quindi calcolare gli autovalori della matrice simmetrica, semidefinita positiva e tridiagonale B^TB.

Con il metodo QR per il calcolo degli autovalori descritto nel capitolo 6 si ottiene la decomposizione

$$B^TB = WDW^T, \tag{41}$$

in cui $D \in \mathbf{R}^{n\times n}$ è la matrice diagonale avente come elementi principali i σ_i^2, $i = 1, \ldots, n$, ordinati in modo non crescente, e $W \in \mathbf{R}^{n\times n}$ è una matrice ortogonale. Da (40) e (41) si ottiene

$$A^HA = HWDW^TH^H,$$

che coincide con la (22) se si pone $Q = HW$.

7.29 Esempio. Per la matrice A dell'esempio 7.28 si ha

$$A^TA = HB^TBH^T,$$

dove

$$H = \frac{1}{5}\begin{bmatrix} 5 & 0 & 0 \\ 0 & -3 & 4 \\ 0 & 4 & 3 \end{bmatrix},$$

$$B^TB = \begin{bmatrix} 4 & 6 & 0 \\ 6 & 18 & 6 \\ 0 & 6 & 13 \end{bmatrix}.$$

Gli autovalori di B^TB calcolati con il metodo QR sono

$$\sigma_1^2 = 23.34213, \quad \sigma_2^2 = 10.31179, \quad \sigma_3^2 = 1.346078,$$

e quindi i valori singolari di A sono

$$\sigma_1 = 4.831369, \quad \sigma_2 = 3.211198, \quad \sigma_3 = 1.160206,$$

e la matrice A risulta di rango 3.

La matrice ortogonale degli autovettori di B^TB è

$$H = \begin{bmatrix} 0.2591518 & -0.3622751 & 0.8953192 \\ 0.8354232 & -0.3810986 & -0.3960196 \\ 0.4846731 & 0.8505999 & 0.2038906 \end{bmatrix}.$$

Posto

$$C = AQ = AHW = \begin{bmatrix} 1.863326 & 2.755032 & 0.01697129 \\ 1.067666 & -1.305669 & -0.2788946 \\ -1.438590 & -0.1623822 & 0.9091571 \\ 1.433844 & -0.1479529 & 0.6420239 \\ -3.821606 & 0.9841209 & -0.1710005 \end{bmatrix},$$

si fattorizza la matrice C nella forma QR con il metodo di Householder; risulta

$$C = UR,$$

dove

$$R = \begin{bmatrix} 4.831369 & 0 & 0 \\ 0 & 3.211198 & 0 \\ 0 & 0 & 1.160206 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

e

$$U = \begin{bmatrix} 0.3856735 & 0.8579459 & 0.01462799 & -0.2087997 & 0.2671558 \\ 0.2209868 & -0.4065989 & -0.2403848 & -0.2269533 & 0.8225245 \\ -0.2977611 & -0.05056683 & 0.7836169 & -0.5247251 & 0.1392329 \\ 0.2967787 & -0.04607503 & 0.5533710 & 0.7317119 & 0.2611086 \\ -0.7910007 & 0.3064681 & -0.1473861 & 0.3068481 & 0.4056067 \end{bmatrix}.$$

La decomposizione ai valori singolari di A è allora data da

$$A = URV^T,$$

dove

$$V = Q = HW = \begin{bmatrix} 0.2591518 & -0.3622751 & 0.8953192 \\ -0.1135125 & 0.9091362 & 0.4007223 \\ 0.9591415 & 0.2054818 & -0.1944808 \end{bmatrix}.$$

In questo caso la matrice A ha rango massimo e il problema dei minimi quadrati (2) con

$$\mathbf{b} = [1,\ 1,\ 1,\ 1,\ 1]^T$$

ha un'unica soluzione, che per il teorema 7.15 è data da

$$\mathbf{x}^* = \frac{\mathbf{q}_1^T\mathbf{b}}{\sigma_1}\,\mathbf{t}_1 + \frac{\mathbf{q}_2^T\mathbf{b}}{\sigma_2}\,\mathbf{t}_2 + \frac{\mathbf{q}_3^T\mathbf{b}}{\sigma_3}\,\mathbf{t}_3,$$

in cui i vettori $\mathbf{q}_j$ e $\mathbf{t}_j, j = 1, 2, 3$ sono rispettivamente le prime tre colonne della matrice U e della matrice V. Risulta allora

$$\mathbf{x}^* = [0.6592600,\ 0.5244430,\ \text{-}0.1560473]^T. \qquad \blacksquare$$

Il metodo QR per il calcolo degli autovalori di B^TB, matrice tridiagonale, richiede ad ogni iterazione un numero di operazioni moltiplicative lineare in n. Poiché B è bidiagonale, anche il calcolo di B^TB richiede un numero di operazioni moltiplicative lineare in n. In [9] è esposta una particolare tecnica che consente di utilizzare il metodo QR per il calcolo degli autovalori di B^TB senza calcolare esplicitamente gli elementi di B^TB. Anche questa tecnica richiede ad ogni iterazione un numero di operazioni moltiplicative lineare in n.

Come è già stato rilevato, uno dei problemi più delicati è quello del calcolo del rango della matrice A. In pratica, fissata una tolleranza ϵ, si assume come rango di A il numero degli elementi principali b_{ii} di B tali che $|b_{ii}| \geq \epsilon$. La scelta di ϵ assume quindi un ruolo molto importante nella risoluzione del problema.

In generale, prima di calcolare gli autovalori e gli autovettori della matrice B^TB, è opportuno esaminare se uno degli elementi α_i o β_i di B è nullo, perché in tal caso il problema del calcolo degli autovalori di B^TB è ricondotto al calcolo degli autovalori di due matrici di ordine inferiore.

a) Se esiste un indice $i,\ 1 \leq i \leq n-1$, per cui $\beta_i = 0$, allora la matrice B ha la forma

$$B = \begin{bmatrix} B_1 & O \\ O & B_2 \end{bmatrix} \begin{matrix} \} & i \text{ righe} \\ \} & n-i \text{ righe} \end{matrix}$$

dove B_1 e B_2 sono blocchi quadrati, e risulta

$$B^TB = \begin{bmatrix} B_1^TB_1 & O \\ O & B_2^TB_2 \end{bmatrix}, \tag{42}$$

in cui le matrici $B_1^TB_1$ e $B_2^TB_2$ hanno rispettivamente ordine i e $n-i$;

b) se esiste un indice $i,\ 2 \leq i \leq n-1$, per cui $\alpha_i = 0$ e $\beta_i \neq 0$, allora la matrice B ha la forma

$$B = \begin{bmatrix} B_1 & O \\ O & B_2 \end{bmatrix},$$

dove $B_1 \in \mathbf{R}^{(i-1)\times i}$ e $B_2 \in \mathbf{R}^{(n-i+1)\times(n-i)}$, per cui B^TB è ancora della forma (42).

7.30 Esempio. La matrice bidiagonale

$$B = \begin{bmatrix} 3 & -1 & 0 & 0 & 0 \\ 0 & 0 & -2 & 0 & 0 \\ 0 & 0 & 2 & -1 & 0 \\ 0 & 0 & 0 & \frac{1}{6}\sqrt{6} & 2 \\ 0 & 0 & 0 & 0 & \sqrt{3} \end{bmatrix},$$

ha l'elemento $b_{22} = 0$. Il problema del calcolo degli autovalori di $B^T B$ viene ricondotto al calcolo degli autovalori delle due matrici

$$B_1^T B_1 = \begin{bmatrix} 9 & -3 \\ -3 & 1 \end{bmatrix}, \quad \text{e} \quad B_2^T B_2 = \begin{bmatrix} 8 & -2 & 0 \\ -2 & \frac{7}{6} & \frac{1}{3}\sqrt{6} \\ 0 & \frac{1}{3}\sqrt{6} & 7 \end{bmatrix}. \qquad \blacksquare$$

10. Calcolo della soluzione di minima norma con il metodo del gradiente coniugato

Il metodo del gradiente coniugato descritto nel capitolo 5 per la risoluzione dei sistemi lineari può essere utilizzato anche per calcolare la soluzione di minima norma del problema dei minimi quadrati

$$\min_{\mathbf{y} \in \mathbf{R}^n} \|A\mathbf{y} - \mathbf{b}\|_2^2, \quad \text{dove } A \in \mathbf{R}^{m \times n} \text{ e } \mathbf{b} \in \mathbf{R}^m, \quad \text{con} \quad m \geq n.$$

In questo caso il funzionale che si deve minimizzare è dato da

$$\|A\mathbf{y} - \mathbf{b}\|_2^2 = 2(\tfrac{1}{2}\mathbf{y}^T A^T A\mathbf{y} - \mathbf{b}^T A\mathbf{y}) + \mathbf{b}^T\mathbf{b}, \quad \mathbf{y} \in \mathbf{R}^n$$

e soluzione del problema dei minimi quadrati è un vettore $\mathbf{x}$ tale che

$$\Phi(\mathbf{x}) = \min_{\mathbf{y} \in \mathbf{R}^n} \Phi(\mathbf{y}),$$

dove

$$\Phi(\mathbf{y}) = \tfrac{1}{2}\|A\mathbf{y} - \mathbf{b}\|_2^2 - \mathbf{b}^T\mathbf{b} = \tfrac{1}{2}\mathbf{y}^T A^T A\mathbf{y} - \mathbf{b}^T A\mathbf{y}.$$

Si può quindi applicare il metodo del gradiente coniugato utilizzando l'algoritmo esposto nel paragrafo 7 del capitolo 5 con l'ovvia modifica che la direzione del gradiente negativo di $\Phi(\mathbf{x})$ nel punto $\mathbf{x}_k$ è data da

$$-\nabla\Phi(\mathbf{x}_k) = A^T(\mathbf{b} - A\mathbf{x}_k) = A^T\mathbf{r}_k,$$

dove $\mathbf{r}_k$ è il residuo del punto $\mathbf{x}_k$. L'algoritmo risulta allora il seguente:

1. $k = 0,\ \mathbf{x}_0$ arbitrario, $\mathbf{r}_0 = \mathbf{b} - A\mathbf{x}_0$
2. se $\mathbf{r}_k = \mathbf{0}$, stop
3. altrimenti si calcoli

$$\mathbf{s}_k = A^T\mathbf{r}_k,$$

$$\beta_k = \frac{\|\mathbf{s}_k\|_2^2}{\|\mathbf{s}_{k-1}\|_2^2} \quad (\beta_0 = 0, \text{ per } k = 0),$$

$$\mathbf{p}_k = \mathbf{s}_k + \beta_k\mathbf{p}_{k-1} \quad (\mathbf{p}_0 = \mathbf{s}_0, \text{ per } k = 0),$$

$$\alpha_k = \frac{\|\mathbf{s}_k\|_2^2}{\|A\mathbf{p}_k\|_2^2},$$

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k\mathbf{p}_k,$$

$$\mathbf{r}_{k+1} = \mathbf{r}_k - \alpha_k A\mathbf{p}_k,$$

$k = k + 1$ e si vada al punto 2.

Come condizione di arresto si può usare la stessa condizione di arresto usata nel capitolo 5, cioè

$$\|\mathbf{s}_k\|_2 < \epsilon\,\|\mathbf{b}\|_2,$$

dove ϵ è una tolleranza prefissata.

7.31 Esempio. Si applica il metodo del gradiente coniugato al problema di minimi quadrati (2) con

$$A = \frac{1}{100}\begin{bmatrix} \text{-}50 & 230 & 235 \\ 50 & \text{-}142 & 81 \\ 50 & 38 & \text{-}159 \\ 100 & \text{-}4 & 122 \\ \text{-}150 & 126 & \text{-}343 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix},$$

(la soluzione è stata calcolata per mezzo dei valori singolari negli esempi 7.28 e 7.29). Fissato $\mathbf{x}_0 = \mathbf{0}$, con l'algoritmo descritto si ottiene la successione

k	$\mathbf{x}_k$			$\|\mathbf{s}_k\|_2$
1	0.	0.2432537	-0.06277519	1.508581
2	0.1100399	0.3457329	0.001886844	1.271655
3	0.6592587	0.5244448	-0.1560494	$0.8422623\ 10^{-5}$

Si assume quindi $\mathbf{x}_3$ come approssimazione di $\mathbf{x}^*$. ∎

Se la matrice A ha rango massimo, con questo algoritmo si ottiene un'approssimazione della soluzione $\mathbf{x}^*$ del problema (2); se A non ha rango massimo si ottiene un'approssimazione di una soluzione $\mathbf{x}$ del problema (2), che dipende dalla scelta del punto iniziale $\mathbf{x}_0$. Scegliendo $\mathbf{x}_0 = \mathbf{0}$ si ottiene un'approssimazione della soluzione $\mathbf{x}^*$ di minima norma. Infatti per ogni $\mathbf{x} \in N(A)$, si ha

$$\mathbf{x}^T \mathbf{s}_k = \mathbf{x}^T A^T \mathbf{r}_k = 0 \quad \text{per ogni } k,$$

e poiché $\mathbf{p}_0 = \mathbf{s}_0$ e $\mathbf{x}_0 = \mathbf{0}$, procedendo per induzione su k, si ha

$$\mathbf{x}^T \mathbf{p}_k = \mathbf{x}^T \mathbf{s}_k + \beta_k \mathbf{x}^T \mathbf{p}_{k-1} = 0,$$

$$\mathbf{x}^T \mathbf{x}_{k+1} = \mathbf{x}^T \mathbf{x}_k + \alpha_k \mathbf{x}^T \mathbf{p}_k = 0,$$

cioè $\mathbf{x}_{k+1}$ appartiene allo spazio $N(A)^\perp$. Poiché in aritmetica esatta la successione $\{\mathbf{x}_k\}$, dopo al più n passi, raggiunge una soluzione del problema (2), questa deve appartenere allo spazio $N(A)^\perp$, ed essendo $N(A)^\perp = N(A^H A)^\perp$ (si veda l'esercizio 1.38), coincide con $\mathbf{x}^*$, che è l'unica soluzione di (2) appartenente a $N(A^H A)^\perp$.

7.32 Esempio. Si applica il metodo del gradiente coniugato al problema dei minimi quadrati (2) con

$$A = \frac{1}{45}\begin{bmatrix} 6 & 12 & -72 \\ -16 & -7 & -8 \\ 58 & 16 & 104 \\ 87 & 24 & 156 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix},$$

già studiato nell'esempio 7.3, la cui soluzione di minima norma è stata calcolata ricorrendo ai valori singolari nell'esempio 7.16. Fissato $\mathbf{x}_0 = [1,\ 1,\ 1]^T$, si ottiene la successione

k	$\mathbf{x}_k$			$\|\mathbf{s}_k\|_2$
1	0.4538005	0.8656725	-0.1101770	0.2133383
2	0.6197463	0.9604924	-0.2049520	$0.3888539\ 10^{-3}$
3	0.6197532	0.9604941	-0.2049382	$0.5444772\ 10^{-6}$

e $\mathbf{x}_3$ approssima la soluzione (si veda l'esempio 7.3)

$$\mathbf{x} = \begin{bmatrix} -\dfrac{1}{5} - 4h \\ \dfrac{13}{5} + 8h \\ h \end{bmatrix}, \quad \text{per } h = -0.2049382,$$

che non è di minima norma, con l'errore effettivo

$$\|\mathbf{x}_3 - \mathbf{x}\|_2 = 0.1013279\ 10^{-5}.$$

Fissato $\mathbf{x}_0 = \mathbf{0}$, la successione che si ottiene è

k	$\mathbf{x}_k$			$\|\mathbf{s}_k\|_2$
1	0.1246007	0.04153361	0.1661344	0.9774478
2	0.8666654	0.4666667	-0.2666695	$0.8394349\ 10^{-4}$
3	0.8666668	0.4666670	-0.2666665	$0.1644842\ 10^{-6}$

e $\mathbf{x}_3$ approssima la soluzione di minima norma $\mathbf{x}^*$ con l'errore effettivo

$$\|\mathbf{x}_3 - \mathbf{x}^*\|_2 = 0.4525244\ 10^{-6}.$$ ∎

7.33 Esempio (approssimazione polinomiale). Sia $f(x)$ una funzione reale e siano $x_i, i = 1, 2, \ldots, m,\ m$ punti, a due a due distinti, appartenenti al dominio di $f(x)$. Fra tutti i polinomi

$$p_{n-1}(x) = \alpha_0 + \alpha_1 x + \ldots + \alpha_{n-1} x^{n-1}$$

di grado minore o uguale a $n-1$, quello per cui lo *scarto quadratico*

$$\sum_{i=1}^{m} [p_{n-1}(x_i) - f(x_i)]^2$$

è minimo viene detto *polinomio di approssimazione ai minimi quadrati* di $f(x)$. I coefficienti del polinomio di approssimazione ai minimi quadrati possono essere determinati con una delle tecniche descritte in questo capitolo, risolvendo il problema (2) in cui la matrice $A \in \mathbf{R}^{m\times n}$ e i vettori $\mathbf{y} \in \mathbf{R}^n$ e $\mathbf{b} \in \mathbf{R}^m$ hanno gli elementi

$$a_{ij} = x_i^{j-1}, \quad y_j = \alpha_{j-1}, \quad b_i = f(x_i), \quad i = 1, \ldots, m,\ j = 1, \ldots, n.$$

Poiché i punti x_i sono a due a due distinti, la matrice A (come si è rilevato nell'esempio 7.22) ha rango massimo e quindi il problema dei minimi quadrati (2) ha una e una sola soluzione, che fornisce i coefficienti del polinomio di approssimazione dei minimi quadrati di grado $n-1$.

Un confronto fra i metodi esposti viene ora fatto per il caso particolare $x_i = \frac{1}{i},\ i = 1, \ldots, m,\ f(x) = \sqrt{x}$, per $m = 10$ in cui al variare di $n-1$ il numero di condizionamento $\mu_2(A)$ e il minimo γ del problema sono approssimativamente i seguenti:

$n-1$	$\mu_2(A)$	γ
2	26	0.5366
3	153	1.112
4	969	2.323
5	6561	5.334
6	48520	6.500
7	397678	14.41

I metodi sperimentati sono:

NR risoluzione del sistema normale con il metodo di Cholesky (par. 1)

QR metodo QR (par. 2)

VS calcolo per mezzo dei valori singolari (par. 5)

GC calcolo per mezzo del gradiente coniugato (par. 10).

I risultati ottenuti sono riportati nella seguente tabella, in cui t è il tempo di calcolo impiegato, misurato in millesimi di secondo, ed $\|\mathbf{e}\|_2$ è la norma 2 dell'errore assoluto dei coefficienti determinati.

metodo	$n-1=2$		$n-1=3$		$n-1=4$	
	t	$\|\mathbf{e}\|_2$	t	$\|\mathbf{e}\|_2$	t	$\|\mathbf{e}\|_2$
NR	0.22	$0.16\ 10^{-4}$	0.24	$0.13\ 10^{-2}$	0.37	$0.11\ 10^{0}$
QR	0.18	$0.27\ 10^{-5}$	0.25	$0.11\ 10^{-4}$	0.37	$0.18\ 10^{-3}$
VS	2.34	$0.84\ 10^{-5}$	3.67	$0.62\ 10^{-4}$	5.28	$0.18\ 10^{-3}$
GC	0.59	$0.76\ 10^{-5}$	1.10	$0.98\ 10^{-4}$	2.34	$0.29\ 10^{-3}$

metodo	$n-1=5$		$n-1=6$		$n-1=7$	
	t	$\|\mathbf{e}\|_2$	t	$\|\mathbf{e}\|_2$	t	$\|\mathbf{e}\|_2$
NR	–	–	–	–	–	–
QR	0.46	$0.43\ 10^{-2}$	0.57	$0.36\ 10^{-1}$	0.80	$0.10\ 10^{0}$
VS	7.02	$0.15\ 10^{-2}$	9.13	$0.65\ 10^{-2}$	10.7	$0.58\ 10^{1}$
GC	3.87	$0.92\ 10^{-3}$	8.00	$0.11\ 10^{-1}$	7.18	$0.37\ 10^{2}$

Dai risultati riportati in questa tabella, risulta che, in accordo con quanto esposto nel paragrafo 7, il metodo **NR** è il meno stabile e addirittura non consente di calcolare la soluzione per $n \geq 6$, perché la matrice A^TA è così malcondizionata che, a causa della limitata precisione di calcolo, non viene riconosciuta come definita positiva dal metodo di Cholesky. Il metodo **QR** consente di calcolare la soluzione $\mathbf{x}^*$ in un tempo minore del metodo

NR e nettamente minore dei metodi **VS** e **GC**. Il tempo di esecuzione del metodo **GC** dipende dal valore della tolleranza ϵ fissata per la condizione di arresto: è possibile ridurre il tempo di esecuzione utilizzando una tecnica di precondizionamento. Per quanto riguarda l'errore della soluzione, i metodi **QR**, **VS** e **GC** hanno un comportamento confrontabile, con un errore che è legato a $\mu_2(A)$.

Comunque queste tecniche non sono adatte a risolvere il problema dell'approssimazione polinomiale quando n è grande per l'elevato malcondizionamento della matrice A. Altre tecniche, basate sull'uso di opportuni polinomi ortogonali, consentono una migliore risoluzione di questo problema. ∎

11. Il metodo di Lanczos per il calcolo dei valori e dei vettori singolari

In molte applicazioni la matrice A del problema (2) è di grosse dimensioni e sparsa: in questi casi le tecniche esposte nei paragrafi precedenti possono non essere praticamente applicabili se le matrici intermedie generate non sono sparse. Inoltre talvolta possono essere richiesti solo pochi valori e vettori singolari. Se la matrice A è reale il metodo di Lanczos può essere convenientemente applicato anche in questo caso.

Nel paragrafo 8 si è visto che gli autovalori della matrice B definita nella (38) e i corrispondenti autovettori, consentono di calcolare i valori e i vettori singolari della matrice A. Se $m = n$ risulta $U_1 = U$ e

$$Z = \frac{1}{\sqrt{2}} \begin{bmatrix} U_1 & U_1 \\ V & -V \end{bmatrix},$$

e quindi gli autovettori $\mathbf{z}$ di B sono tutti della forma

$$\mathbf{z} = \frac{1}{\sqrt{2}} \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix}, \quad \|\mathbf{u}\|_2 = \|\mathbf{v}\|_2 = 1.$$

Se $m > n$, allora vi sono degli autovettori $\mathbf{z}$ di B della forma

$$\mathbf{z} = \begin{bmatrix} \mathbf{u} \\ \mathbf{0} \end{bmatrix}, \quad \mathbf{u} \in \mathbf{C}^m, \ \|\mathbf{u}\|_2 = 1, \tag{43}$$

corrispondenti all'autovalore nullo, che, essendo dovuti al fatto che la matrice A è rettangolare, non corrispondono ad alcun valore singolare di A.

Perciò quando si determinano gli autovettori di B, si devono eliminare quelli della forma (43).

L'ordine della matrice B è dato dalla somma delle dimensioni di A, ma se A è sparsa, anche B risulta essere sparsa. Il seguente teorema illustra come una scelta opportuna, suggerita in [7], del vettore iniziale $\mathbf{q}_1$ consenta di ridurre sia il tempo di esecuzione che lo spazio di memoria necessario a contenere i vettori generati dal metodo di Lanczos.

7.34 Teorema. *Sia $A \in \mathbf{R}^{m\times n}$, $m \geq n$, e sia B la matrice definita in (38). Si applichi a B il metodo di Lanczos descritto nel paragrafo 5 del capitolo 6, scegliendo come vettore iniziale $\mathbf{q}_1$ un vettore di una delle due forme*

$$\begin{aligned} &[\mathbf{u}\ ,\ \mathbf{0}]^T, \quad \mathbf{u} \in \mathbf{R}^m,\ \|\mathbf{u}\|_2 = 1,\ \mathbf{0} \in \mathbf{R}^n, \\ &[\mathbf{0}\ ,\ \mathbf{v}]^T, \quad \mathbf{v} \in \mathbf{R}^n,\ \|\mathbf{v}\|_2 = 1,\ \mathbf{0} \in \mathbf{R}^m. \end{aligned} \tag{44}$$

La matrice tridiagonale T ottenuta ha gli elementi principali nulli e i vettori $\mathbf{q}_i$ generati assumono, alternativamente, una delle due forme (44).

Dim. Se $\mathbf{q}_1 = [\mathbf{u}_1\ ,\ \mathbf{0}]^T$, allora per le (20) del capitolo 6 si ha

$$\alpha_1 = \mathbf{q}_1^H B\mathbf{q}_1 = 0, \quad \mathbf{q}_2 = \frac{1}{\beta_1}\, B\mathbf{q}_1 = \frac{1}{\beta_1} \begin{bmatrix} \mathbf{0} \\ \\ A^H\mathbf{u}_1 \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \\ \mathbf{v}_2 \end{bmatrix},$$

dove $\beta_1 = \|A^H\mathbf{u}_1\|_2$, e quindi $\|\mathbf{v}_2\|_2 = 1$,

$$\alpha_2 = \mathbf{q}_2^H B\mathbf{q}_2 = 0,$$

$$\mathbf{q}_2 = \frac{1}{\beta_2}\,(B\mathbf{q}_2 - \beta_1\mathbf{q}_1) = \frac{1}{\beta_2} \begin{bmatrix} A\mathbf{v}_2 - \beta_1\mathbf{u}_1 \\ \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{u}_3 \\ \\ \mathbf{0} \end{bmatrix},$$

dove $\beta_2 = \|A\mathbf{v}_2 - \beta_1\mathbf{u}_1\|_2$, e quindi $\|\mathbf{u}_3\|_2 = 1$, e così via, fino a $\mathbf{q}_n$.

Se $\mathbf{q}_1 = [\mathbf{0}\ ,\ \mathbf{v}_1]^T$, si procede in modo analogo. ∎

Scegliendo come vettore iniziale un vettore che abbia una delle forme (44), basta memorizzare ogni due passi un vettore di lunghezza $m+n$; anche il costo computazionale del calcolo di $B\mathbf{q}_i$, $i = 1, \ldots, n-1$, diminuisce proporzionalmente. Un'ulteriore semplificazione si può ottenere tenendo conto del fatto che la matrice tridiagonale T ha gli elementi principali nulli.

Il calcolo con il metodo di Lanczos dei valori singolari più grandi di A non presenta in generale grosse difficoltà, mentre più delicato può risultare il calcolo dei valori singolari più piccoli, in particolare se sono molto vicini allo zero, perché la matrice B ha come autovalori anche i valori singolari

di A cambiati di segno, ed inoltre ha gli autovalori nulli dovuti al fatto che la matrice A è rettangolare. Per cui, se il minimo valore singolare non nullo σ_r di A, è piccolo, nell'intervallo $[-\sigma_r, \sigma_r]$ si trovano $m + n - 2r + 2$ autovalori molto vicini fra loro e quindi di difficile approssimazione. In [4] Cullum e Willoughby espongono un algoritmo che consente un'agevole approssimazione anche degli autovalori singolari più piccoli.

7.35 Esempio. I valori singolari della matrice $A \in \mathbf{R}^{m\times n}$, $m = n + 3$, di elementi

$$a_{ij} = \begin{cases} 1 & \text{se } i-3 \le j \le i, \\ \\ 0 & \text{altrimenti,} \end{cases}$$

sono tutti minori di 4 e $\lim_{n\to\infty} \sigma_1 = 4$. Per $n = 1000$ si calcola σ_1 con il metodo di Lanczos, scegliendo come vettore iniziale il vettore

$$\mathbf{q}_1 = [\mathbf{0},\ \mathbf{v}]^T, \quad \mathbf{0} \in \mathbf{R}^m,\ v_i = \frac{1}{\sqrt{n}}\ ,\ \text{per } i = 1, \ldots, n.$$

Operando in precisione semplice (con sei cifre esadecimali), σ_1 viene approssimato con un errore minore di 10^{-4} in 5 passi; non si può ottenere un risultato migliore a causa degli errori di arrotondamento. Per ottenere un risultato migliore si ricorre alla doppia precisione (con 14 cifre esadecimali), che fornisce il valore approssimato $\sigma_1 = 3.999889$, affetto da un errore minore di 10^{-6} in 40 passi. ∎

Esercizi proposti

7.1 Sia $A \in \mathbf{C}^{m\times n}$. Si dimostri la seguente proprietà dell'*alternativa di Fredholm*: o il sistema $A\mathbf{x} = \mathbf{b}$ è risolubile per ogni $\mathbf{b} \in \mathbf{C}^m$, oppure il sistema omogeneo $A^H\mathbf{y} = \mathbf{0}$ ha soluzioni non nulle.

(Traccia: se $A\mathbf{x} = \mathbf{b}$ è risolubile per ogni $\mathbf{b} \in \mathbf{C}^m$, allora $m \le n$ e rango di $A = m$.)

7.2 Siano

$$A = \begin{bmatrix} 1 & -1 \\ -1 & 1 \\ 2 & \alpha \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \quad \alpha = 1 \text{ oppure } \alpha = -2.$$

Si risolva il sistema $A\mathbf{x} = \mathbf{b}$ nel senso dei minimi quadrati, cioè si risolva il problema dei minimi quadrati

$$\min_{\mathbf{x}\in\mathbf{R}^2} \|A\mathbf{x} - \mathbf{b}\|_2,$$

con i diversi metodi descritti nel capitolo.

7.3 Si risolvano il sistema lineare nel senso dei minimi quadrati

$$(1)\ \begin{cases} x_1 + x_2 = 1 \\ 2x_1 + 2x_2 = 0 \\ -x_1 - x_2 = 2 \end{cases} \quad (2)\ \begin{cases} x_1 + x_2 = 1 \\ 2x_1 = 0 \\ -x_1 + 3x_2 = 2 \end{cases} \quad (3)\ \begin{cases} x_1 + x_2 = 2 \\ x_1 - x_2 = 0 \\ 2x_1 + x_2 = 2 \end{cases}$$

(Risposta: (1) $\mathbf{x} = [x_1, x_2]^T$, tali che $x_1 + x_2 = -\frac{1}{6}$, $\mathbf{x}^* = -\frac{1}{12}\,[1\ ,\ 1]^T$;
(2) $\mathbf{x} = \mathbf{x}^* = \frac{1}{14}\,[1\ ,\ 10]^T$; (3) $\mathbf{x} = \mathbf{x}^* = \frac{1}{7}\,[5\ ,\ 6]^T$.)

7.4 Si determini la migliore approssimazione lineare nel senso dei minimi quadrati ai dati

x	1	2	3
y	1	0	1

(Risposta: indicata con $y = x_1 x + x_2$ l'equazione della retta cercata, i coefficienti x_1 e x_2 devono soddisfare, nel senso dei minimi quadrati, il sistema lineare

$$\begin{cases} x_1 + x_2 = 1 \\ 2x_1 + x_2 = 0 \\ 3x_1 + x_2 = 1 \end{cases}$$

Risulta $[x_1, x_2] = [0, \frac{2}{3}]$.)

7.5 Si calcoli la decomposizione ai valori singolari e la pseudoinversa di Moore-Penrose delle seguenti matrici A

$$(1)\quad \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}, \qquad (2)\quad \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \qquad (3)\quad \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix},$$

$$(4)\quad \frac{1}{15}\begin{bmatrix} \text{-}2 & \text{-}14 \\ \text{-}8 & 19 \\ \text{-}20 & 10 \end{bmatrix}, \quad (5)\quad \frac{1}{11}\begin{bmatrix} 36 & 27 \\ 24 & 18 \\ 8 & 6 \end{bmatrix}, \quad (6)\quad \frac{1}{6}\begin{bmatrix} \text{-}3 & \text{-}1 & \text{-}5 & 1 \\ \text{-}3 & \text{-}1 & 1 & \text{-}5 \\ \text{-}3 & 5 & 1 & 1 \end{bmatrix},$$

(Risposta: (1) $U = V = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ 1 & \text{-}1 \end{bmatrix}$, $\Sigma = \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix}$, $A^+ = A^{-1}$;
(2) $U = A$, $\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, $V = I$, $A^+ = A^{-1}$;

(3) $U = \dfrac{1}{\sqrt{2}} \begin{bmatrix} 1 & 0 & 1 \\ 0 & \sqrt{2} & 0 \\ 1 & 0 & -1 \end{bmatrix}$, $\Sigma = \begin{bmatrix} \sqrt{2} & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}$, $V = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$,

$A^+ = \dfrac{1}{2} \begin{bmatrix} 1 & 0 & 1 \\ 0 & 2 & 0 \end{bmatrix}$;

(4) $U = \dfrac{1}{3} \begin{bmatrix} 1 & -2 & -2 \\ -2 & 1 & -2 \\ -2 & -2 & 1 \end{bmatrix}$, $\Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}$, $V = \dfrac{1}{5} \begin{bmatrix} 3 & 4 \\ -4 & 3 \end{bmatrix}$,

$A^+ = \dfrac{1}{30} \begin{bmatrix} -13 & 2 & -22 \\ -16 & 14 & -4 \end{bmatrix}$;

(5) $U = \dfrac{1}{11} \begin{bmatrix} 9 & 2 & -6 \\ 6 & -6 & 7 \\ 2 & 9 & 6 \end{bmatrix}$, $\Sigma = \begin{bmatrix} 5 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$, $V = \dfrac{1}{5} \begin{bmatrix} 4 & -3 \\ 3 & 4 \end{bmatrix}$, $A^+ = \dfrac{1}{25} A^T$;

(6) $U = \dfrac{1}{3} \begin{bmatrix} -2 & -2 & 1 \\ -2 & 1 & -2 \\ 1 & -2 & -2 \end{bmatrix}$, $\Sigma = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$, $V = \dfrac{1}{2} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \end{bmatrix}$,

$A^+ = A^T$.)

7.6 Si dimostri che i valori singolari di A e di A^H sono uguali.

(Traccia: è $A^H = V \Sigma U^H$.)

7.7 Si dimostri che se $A \in \mathbf{C}^{n \times n}$, allora

$$|\det A| = \prod_{i=1}^{n} \sigma_i.$$

(Traccia: $\det A^H A = |\det A|^2$.)

7.8 Sia $A \in \mathbf{C}^{m \times n}$, $m \geq n$.

a) Si dimostri che $A^H A$ è definita positiva se e solo se le colonne di A sono linearmente indipendenti, cioè se A ha rango massimo;

b) sia

$$A = \begin{bmatrix} 1000 & 1020 \\ 1000 & 1000 \\ 1000 & 1000 \\ 1000 & 1000 \end{bmatrix},$$

si calcoli $A^T A$ utilizzando un'aritmetica con base 10 e 4 cifre significative, e si verifichi che la matrice $A^T A$ effettivamente ottenuta non è definita positiva.

7.9 Sia $A \in \mathbf{C}^{m\times n}$ e sia $A = QR$ una fattorizzazione QR di A, con $Q \in \mathbf{C}^{m\times m}$ unitaria e $R \in \mathbf{C}^{m\times n}$ triangolare superiore. Si dimostri che i valori singolari di A e di R sono uguali.

(Traccia: segue dal fatto che $A^H A = R^H R$.)

7.10 Sia $A \in \mathbf{C}^{n\times n}$. Si definisce *radice quadrata* di A una matrice $B \in \mathbf{C}^{n\times n}$ tale che $B^2 = A$. Si dimostri che

a) A è definita (semidefinita) positiva se e solo se ha una radice quadrata B definita (semidefinita) positiva;

b) se B è una radice quadrata di A, allora rango di A = rango di B;

c) la radice quadrata definita (semidefinita) positiva è unica e si indica con $A^{1/2}$;

d) gli autovalori di $(A^H A)^{1/2}$ e $(AA^H)^{1/2}$ coincidono;

e) gli autovalori di $(A^H A)^{1/2}$ sono i valori singolari di A.

(Traccia: a) se A è definita (semidefinita) positiva, sia $A = UDU^H$ la sua forma normale di Schur. Si definisca $B = UD_0U^H$, D_0 matrice diagonale i cui elementi principali sono $\sqrt{\lambda_i}$, dove λ_i è autovalore di A; viceversa se $B^2 = A$, gli autovalori di A sono i quadrati degli autovalori di B; d) segue dall'esercizio 2.13; e) segue dal teorema 7.10.)

7.11 Sia $A \in \mathbf{C}^{m\times n}$. Si dimostri che

a) esiste un'unica matrice X che soddisfa alle seguenti *equazioni di Moore-Penrose*:
   1) $AXA = A$,
   2) $XAX = X$,
   3) $(AX)^H = AX$,
   4) $(XA)^H = XA$;

b) se A ha rango massimo tale matrice è

$$X = \begin{cases} (A^H A)^{-1} A^H & \text{se } m \geq n, \\ A^H (AA^H)^{-1} & \text{se } m \leq n; \end{cases}$$

c) è $X = A^+$, dove A^+ è la pseudoinversa di Moore-Penrose, definita in 7.17.

(Traccia: a) per assurdo siano X_1 e X_2 due matrici che soddisfano alle equazioni; da 2) e 3) si ottiene

$$X_1^H - X_2^H = A(X_1 X_1^H - X_2 X_2^H),$$

per cui $X_1^H - X_2^H$ ha colonne appartenenti a $S(A)$, dove $S(A)$ è l'immagine di A; da 1) e 4) si ottiene $AA^H(X_1^H - X_2^H) = O$, per cui $X_1^H - X_2^H$ ha colonne appartenenti a $N(AA^H) = N(A^H) = S(A)^\perp$. Ne segue che $X_1^H = X_2^H$;
b) per $m \geq n$ da 1) e 3) si ha $A^H = A^H X^H A^H = A^H AX$ e poiché $A^H A$ è non singolare, $X = (A^H A)^{-1} A^H$; per $m \leq n$ si proceda in modo analogo;
c) basta verificare che A^+ soddisfa alle equazioni di Moore-Penrose.)

7.12 Si descriva la classe delle matrici $A \in \mathbf{R}^{3\times 2}$ tali che

$$A^+ = A^T.$$

(Traccia: si dica come devono essere i valori singolari di A.)

7.13 Si dica quali sono le decomposizioni ai valori singolari e le pseudoinverse di Moore-Penrose delle seguenti matrici

a) un vettore $\mathbf{v} \in \mathbf{C}^n$,

b) una matrice nulla $O \in \mathbf{C}^{m\times n}$,

c) una diade $A = \mathbf{x}\mathbf{y}^H,\ \mathbf{x} \in \mathbf{C}^m,\ \mathbf{y} \in \mathbf{C}^n$.

(Traccia: a) sia Q la matrice elementare di Householder tale che

$$Q\mathbf{v} = -\|\mathbf{v}\|_2 \theta \mathbf{e}_1, \quad \text{dove } \theta = \frac{v_1}{|v_1|},\ \text{se } v_1 \neq 0,\ \theta = 1,\ \text{se } v_1 = 0,$$

allora

$$\mathbf{v} = U\Sigma V^T, \quad \text{dove } U = -\theta Q^H,\ \Sigma = \|\mathbf{v}\|_2 \mathbf{e}_1,\ V = [1],$$

e

$$\mathbf{v}^+ = \frac{1}{\|\mathbf{v}\|_2^2}\ \mathbf{v}^H;$$

c) $A^H A = \mathbf{y}\mathbf{x}^H \mathbf{x}\mathbf{y}^H = \|\mathbf{x}\|_2^2 \mathbf{y}\mathbf{y}^H = QDQ^H$ (si veda l'esercizio 2.23), dove $D \in \mathbf{C}^{n\times n}$ è la matrice nulla a parte $d_{11} = \|\mathbf{x}\|_2^2 \|\mathbf{y}\|_2^2$, e $Q \in \mathbf{C}^{n\times n}$ è una qualunque matrice unitaria la cui prima colonna è $\mathbf{y}/\|\mathbf{y}\|_2$. La matrice $C = AQ = \mathbf{x}\mathbf{y}^H Q \in \mathbf{C}^{m\times n}$ ha la prima colonna uguale a $\|\mathbf{y}\|_2 \mathbf{x}$ e le altre nulle. Si costruisce la matrice U come al punto a), tale che $\|\mathbf{y}\|_2 \mathbf{x} = U\|\mathbf{y}\|_2 \|\mathbf{x}\|_2 \mathbf{e}_1$. Risulta quindi

$$\mathbf{x}\mathbf{y}^H = U\Sigma V^H, \quad \Sigma = \begin{bmatrix} \|\mathbf{x}\|_2\|\mathbf{y}\|_2 & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix}, \quad V = I_n,$$

$$A^+ = (\mathbf{x}\mathbf{y}^H)^+ = V\Sigma^+U^H = \frac{1}{\|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2}(U\Sigma V^H)^H = \frac{1}{\|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2}(\mathbf{x}\mathbf{y}^H)^H$$
$$= \frac{1}{\|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2}\mathbf{y}\mathbf{x}^H.)$$

7.14 Sia $A \in \mathbf{C}^{m\times n}$. Si dimostri che la matrice A^+, pseudoinversa di Moore-Penrose, verifica le seguenti proprietà

1) $(A^H)^+ = (A^+)^H$,
2) $(\alpha A)^+ = \frac{1}{\alpha}\, A^+$, per ogni $\alpha \in \mathbf{C},\ \alpha \neq 0$,
3) $(A^+)^+ = A$,
4) $(AA^H)^+ = (A^+)^H A^+$,
5) $(A^H A)^+ = A^+(A^+)^H$.

(Traccia: ci si riferisca alla definizione 7.17)

7.15 Si dimostri che sono chiuse rispetto all'operazione di pseudoinversa di Moore-Penrose le seguenti classi di matrici:

matrici normali,

matrici hermitiane,

matrici definite (semidefinite) positive.

7.16 Sia $A \in \mathbf{C}^{n\times n}$ normale. Si dimostri che

$$(1)\quad A^+A = AA^+, \qquad (2)\quad (A^n)^+ = (A^+)^n.$$

(Traccia: si dimostri che, poiché A è normale, $U = VS$, dove S è una matrice di fase.)

7.17 Siano $A \in \mathbf{C}^{m\times n}$, $B \in \mathbf{C}^{n\times r}$ due matrici di rango massimo. Si dimostri che

$$(AB)^+ = B^+A^+.$$

Si trovi un controesempio che dimostri come questa relazione può non valere se una delle due matrici non è di rango massimo.

(Traccia: si verifichi che $X = B^+A^+$ soddisfa alle equazioni di Moore-Penrose; si consideri ad esempio

$$A = \frac{1}{3}\begin{bmatrix} 2 & 2 \\ -2 & -2 \\ 1 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 1 \\ -2 \end{bmatrix},$$

per cui è

$$(AB)^+ = \frac{1}{3}\,[\text{-}2\,,\ 2\,,\ \text{-}1], \quad B^+A^+ = \frac{1}{30}\,[\text{-}2\,,\ 2\,,\ \text{-}1].)$$

7.18 Sia $A \in \mathbf{C}^{m\times n}$. Si dimostri che X è la pseudoinversa di Moore-Penrose di A se e solo se $XAA^H = A^H$ ed esiste una matrice hermitiana $B \in \mathbf{C}^{n\times n}$ per cui $X = BA^H$.

(Traccia: si sfruttino le equazioni di Moore-Penrose e si ponga $B = XX^H$. Per il viceversa, si ha $AX = ABA^H = AB^HA^H = X^HA^H$, da cui segue la terza equazione di Moore-Penrose; per le altre tre equazioni si dimostri prima che $A = X^HA^HA$, sfruttando il fatto che la matrice $AA^H = AA^HAX$ è hermitiana.)

7.19 Sia $A \in \mathbf{C}^{m\times n}$ e sia $A = U\Sigma V^H$ la sua decomposizione ai valori singolari.

a) Si determinino le decomposizioni ai valori singolari delle matrici

$$P_1 = A^+A, \quad P_2 = I - A^+A, \quad P_3 = AA^+, \quad P_4 = I - AA^+;$$

b) si verifichi che le matrici P_i, $i = 1, \ldots, 4$, sono idempotenti (per la definizione e le proprietà delle matrici idempotenti si vedano gli esercizi 1.9 e 2.31);

c) si dimostri che l'insieme X delle soluzioni del problema dei minimi quadrati (2) è costituito dai vettori della forma

$$\mathbf{x} = A^+\mathbf{b} + (I - A^+A)\mathbf{v},$$

dove $\mathbf{v} \in \mathbf{C}^n$ è arbitrario;

d) si dimostri che la soluzione di minima norma è proprio

$$\mathbf{x}^* = A^+\mathbf{b}.$$

(Traccia: a) se k è il rango di A, si ha

$$P_1 = V \begin{bmatrix} I_k & O \\ O & O \end{bmatrix} V^H, \quad P_2 = V\Pi \begin{bmatrix} I_{n-k} & O \\ O & O \end{bmatrix} \Pi^T V^H,$$

$$P_3 = U \begin{bmatrix} I_k & O \\ O & O \end{bmatrix} U^H, \quad P_4 = U\Pi' \begin{bmatrix} I_{m-k} & O \\ O & O \end{bmatrix} \Pi'^T U^H,$$

dove Π e Π' sono opportune matrici di permutazione; b) si noti che le quattro matrici sono diagonalizzabili e hanno autovalori uguali a zero o a

uno; c) si dimostri che i vettori $\mathbf{x}$ della forma data soddisfano il sistema normale (3): per questo si osservi che $A^H A A^+ = A^H$, e viceversa che tutti i vettori del nucleo di $A^H A$ sono della forma $(I - A^+A)\mathbf{v}$, $\mathbf{v} \in \mathbf{C}^n$; d) si dimostri che i vettori $A^+\mathbf{b}$ e $(I - A^+A)\mathbf{v}$ sono ortogonali e quindi $\|\mathbf{x}\|_2^2 = \|A^+\mathbf{b}\|_2^2 + \|(I - A^+A)\mathbf{v}\|_2^2$; per questo basta dimostrare che $(A^+)^H(I - A^+A) = O$.)

7.20 Sia $A \in \mathbf{C}^{m\times n}$, $m \geq n$. Si dimostri che

$$\sigma_n \|\mathbf{x}\|_2 \leq \|A\mathbf{x}\|_2 \leq \sigma_1 \|\mathbf{x}\|_2,$$

per ogni $\mathbf{x} \in \mathbf{C}^n$.

(Traccia: segue dal punto d) del teorema 7.10.)

7.21 Sia $A \in \mathbf{C}^{m\times n}$, $m \geq n$, di rango massimo e sia B una sottomatrice di A ottenuta cancellando una o più colonne. Si dimostri che

$$\mu_2(B) \leq \mu_2(A).$$

(Traccia: si sfrutti il teorema 7.23 e la (34).)

7.22 Sia $A \in \mathbf{C}^{n\times n}$ non singolare.

a) Si dimostri che

$$\|A^{-1}\|_2 = \frac{1}{\sigma_n};$$

b) Se A è triangolare, si dimostri che

$$\mu_2(A) \geq \frac{\max\limits_{i,j=1,\ldots,n} |a_{ij}|}{\min\limits_{i=1,\ldots,n} |a_{ii}|}.$$

(Traccia: a) si noti che la matrice $(A^H A)^{-1}$ ha gli stessi autovalori della matrice $A^{-H}A^{-1}$ (si veda l'esercizio 2.13); b)

$$\sigma_1 = \|A\|_2 \geq \max_{i,j=1,\ldots,n} |a_{ij}|, \quad \frac{1}{\sigma_n} = \|A^{-1}\|_2 \geq \frac{1}{\min\limits_{i=1,\ldots,n} |a_{ii}|},$$

perché gli elementi principali di A^{-1} sono i reciproci di quelli di A.)

7.23 Sia $A \in \mathbf{C}^{m\times n}$, $m \geq n$, $\mathbf{y} \in \mathbf{C}^n$ e B la matrice

$$B = \begin{bmatrix} A \\ \mathbf{y}^H \end{bmatrix}.$$

Indicati con $\sigma_1, \ldots, \sigma_n$ i valori singolari di A e con $\tau_1, \ldots, \tau_n$ quelli di B, si dimostri che

$$\tau_1 \leq \sqrt{\|A\|_2^2 + \|\mathbf{y}\|_2^2}, \quad \tau_n \geq \sigma_n.$$

(Traccia: si sfruttino i teoremi 7.10 e 6.12.)

7.24 Sia $A \in \mathbf{C}^{m\times n}$, $m \geq n$. Si dica qual è la decomposizione ai valori singolari della matrice

$$B = \begin{bmatrix} O & A \\ A^H & O \end{bmatrix}.$$

(Traccia: sia $A = U\Sigma V^H$ la decomposizione ai valori singolari di A, e sia

$$W = \begin{bmatrix} U & O \\ O & V \end{bmatrix};$$

si esamini la struttura della matrice $W^H B W$ e si determinino le opportune matrici di permutazione.)

7.25 Siano $A,\ B \in \mathbf{C}^{n\times n}$. Si dica quali sono le decomposizioni ai valori singolari delle matrici

$$C_1 = \begin{bmatrix} A & B \\ B & A \end{bmatrix}, \quad C_2 = \begin{bmatrix} A & -B \\ B & A \end{bmatrix}.$$

(Traccia: seguendo la traccia dell'esercizio 2.26, si determini una matrice unitaria Z tale che

$$ZC_1Z^H = \begin{bmatrix} \Sigma' & O \\ O & \Sigma'' \end{bmatrix},$$

dove $\Sigma',\ \Sigma'' \in \mathbf{R}^{n\times n}$ sono le matrici diagonali i cui elementi principali sono i valori singolari di $A+B$ e di $A-B$. Si determinino poi delle opportune matrici di permutazione. Si proceda in modo analogo con la seconda matrice.)

7.26 Sia $A \in \mathbf{C}^{m\times n}$ della forma

$$A = \begin{bmatrix} A_{11} & A_{12} \\ O & A_{22} \end{bmatrix},$$

con $A_{11} \in \mathbf{C}^{k\times k},\ A_{12} \in \mathbf{C}^{k\times(n-k)},\ A_{22} \in \mathbf{C}^{(m-k)\times(n-k)}$, e siano $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n \geq 0$ i valori singolari di A. Si dimostri che $\sigma_{k+1} \leq \|A_{22}\|_2$.

(Traccia: sia

$$B = \begin{bmatrix} A_{11} & A_{12} \\ O & O \end{bmatrix}.$$

Per il teorema 7.13 è $\|A - B\|_2 \geq \sigma_{r+1}$, dove r è il rango di B.)

7.27 Sia $A \in \mathbf{C}^{n\times n}$, e sia $A = U\Sigma V^H$ la sua decomposizione ai valori singolari. Si dimostri che

$$\|A - UV^H\|_F = \min_{\substack{W\in\mathbf{C}^{n\times n}\\ W^HW=I}} \|A - W\|_F,$$

cioè la matrice UV^H è la matrice unitaria più vicina alla matrice A nel senso della norma di Frobenius.

(Traccia: basta dimostrare che $\|A - UV^H\|_F \leq \|A - W\|_F$ per ogni $W \in \mathbf{C}^{n\times n}$ unitaria; moltiplicando per U^H a sinistra e per V a destra, si vede che basta dimostrare che $\|\Sigma - I\|_F \leq \|\Sigma - Z\|_F$ per ogni $Z \in \mathbf{C}^{n\times n}$ unitaria. Poiché Z è unitaria è $|\mathrm{Re}(z_{ii})| \leq 1$, e quindi

$$\begin{aligned}\|\Sigma - Z\|_F &= \mathrm{tr}(\Sigma^2 - \Sigma Z - Z^H\Sigma + I) = \sum_{i=1}^{n}(\sigma_i^2 - \sigma_i z_{ii} - \sigma_i\overline{z}_{ii} + 1)\\ &= \sum_{i=1}^{n}(\sigma_i^2 - 2\sigma_i\mathrm{Re}(z_{ii}) + 1) \geq \sum_{i=1}^{n}(\sigma_i^2 - 2\sigma_i + 1) = \|\Sigma - I\|_F.\ )\end{aligned}$$

7.28 Siano $A \in \mathbf{C}^{m\times n}$ e $\mathbf{b} \in \mathbf{C}^n$ della forma

$$A = \begin{bmatrix} B & \mathbf{v} \\ O & \mathbf{w} \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix},$$

dove $B \in \mathbf{C}^{k\times(n-1)}$, $\mathbf{v},\ \mathbf{c} \in \mathbf{C}^k$, $\mathbf{w},\ \mathbf{d} \in \mathbf{C}^{m-k}$ e $O \in \mathbf{C}^{(m-k)\times(n-1)}$ è una matrice nulla. Si dimostri che se B ha rango massimo, allora

$$\min_{\mathbf{x}\in\mathbf{C}^n} \|A\mathbf{x} - \mathbf{b}\|_2^2 = \|\mathbf{d}\|_2^2 - \frac{|\mathbf{d}^H\mathbf{w}|^2}{\|\mathbf{w}\|_2^2}\ .$$

(Traccia: posto $\mathbf{x} = \begin{bmatrix} \mathbf{y} \\ z \end{bmatrix},\ \mathbf{y} \in \mathbf{C}^{(n-1)},\ z \in \mathbf{C}$, si ha

$$\|A\mathbf{x} - \mathbf{b}\|_2^2 = \|B\mathbf{y} + \mathbf{v}z - \mathbf{c}\|_2^2 + \|\mathbf{w}z - \mathbf{d}\|_2^2;$$

poiché per ogni z si può determinare un $\mathbf{y}$ tale che $B\mathbf{y} + \mathbf{v}z - \mathbf{c} = \mathbf{0}$, è

$$\min_{\mathbf{x}\in\mathbf{C}^n} \|A\mathbf{x} - \mathbf{b}\|_2^2 = \min_{z\in\mathbf{C}} \|\mathbf{w}z - \mathbf{d}\|_2^2.)$$

7.29 Sia $A \in \mathbf{C}^{m\times n}$. Si dimostri che A^+ è la soluzione dei seguenti problemi

$$\min_{X\in\mathbf{C}^{n\times m}} \|AX - I_m\|_2,$$

$$\min_{X\in\mathbf{C}^{n\times m}} \|AX - I_m\|_F,$$

e si dica quanto vale il minimo nei due casi.

(Traccia: per la norma 2 si considerino gli n problemi dei minimi quadrati

$$\min_{\mathbf{y}\in\mathbf{C}^n} \|A\mathbf{y} - \mathbf{e}_i\|_2,$$

dove $\mathbf{e}_i$ per $i = 1, \ldots, m$ è l'i-esima colonna della matrice I_m. Il minimo risulta uguale a zero se $m \leq n$ e A ha rango massimo, 1 altrimenti. Per la norma di Frobenius, sia $AX = U\Sigma V^H$, $U,\ V \in \mathbf{C}^{m\times m}$, la decomposizione ai valori singolari di AX; si ha $\|AX - I_m\|_F = \|\Sigma - U^H V\|_F$, e tale quantità è minima quando $U^H V = I$ (si veda l'esercizio 7.27), quindi $V = U$; ne segue che $\|\Sigma - I\|_F$ è minima se gli elementi non nulli di Σ sono uguali a 1 e risulta

$$\min_{X\in\mathbf{C}^{n\times m}} \|AX - I_m\|_F = \sqrt{m-k},$$

dove k è il rango di AX. Perciò $AX = U\Sigma U^H$, dove $\sigma_1 = \ldots = \sigma_k = 1$, $\sigma_{k+1} = \ldots = \sigma_m = 0$. Posto $A = U\Sigma' V^H$, deve essere $X = V\Sigma'' U^H$ tale che $\Sigma'\Sigma'' = \Sigma$.)

7.30 Sia $A \in \mathbf{C}^{m\times n}$, $m \geq n$, i cui valori singolari sono $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n \geq 0$. Si dimostri che per $k = 1, \ldots, n$ è

$$\sigma_{n-k+1} = \min_{V_k} \max_{\substack{\|\mathbf{x}\|_2=1 \\ \mathbf{x}\in V_k}} \|A\mathbf{x}\|_2,$$

$$\sigma_k = \max_{V_k} \min_{\substack{\|\mathbf{x}\|_2=1 \\ \mathbf{x}\in V_k}} \|A\mathbf{x}\|_2,$$

dove V_k è un qualunque sottospazio di $\mathbf{C}^n$ di dimensione k.

(Traccia: si applichi il teorema 6.7 alla matrice $A^H A$.)

7.31 Siano $A,\ B,\ C,\ D \in \mathbf{C}^{n\times n}$, $C = AB$, $D = A + B$, e siano α_i, β_i, γ_i, δ_i, $i = 1, \ldots, n$, i loro valori singolari. Si dimostri che

$$\left.\begin{array}{c} \gamma_{i+j+1} \le \alpha_{i+1}\beta_{j+1} \\ \delta_{i+j+1} \le \alpha_{i+1} + \beta_{j+1} \end{array}\right\} \quad i,\ j = 0, \ldots, n-1,\ i+j+1 \le n.$$

(Traccia: siano $\mathbf{u}_r$, $r = 1, \ldots, n$ e $\mathbf{v}_s$, $s = 1, \ldots, n$, tali che

$$AA^H\mathbf{u}_r = \alpha_r^2\mathbf{u}_r, \quad B^HB\mathbf{v}_s = \beta_s^2\mathbf{v}_s,$$

e siano $\mathcal{U}_i$ e $\mathcal{V}_j$ i sottospazi di $\mathbf{C}^n$ generati dai vettori $\mathbf{u}_r$, $r = 1, \ldots, i$ e $\mathbf{v}_s$, $s = 1, \ldots, j$. Sia $AB = U\Sigma V^H$ la decomposizione ai valori singolari di AB. Per la disuguaglianza di Cauchy-Schwartz si ha, posto $\mathbf{x} = U\mathbf{z}$ e $\mathbf{y} = V\mathbf{z}$,

$$0 \le \mathbf{z}^H\Sigma\mathbf{z} = \mathbf{z}^HU^HABV\mathbf{z} = \mathbf{x}^HAB\mathbf{y} \le \|A^H\mathbf{x}\|_2\|B\mathbf{y}\|_2.$$

Se $\mathbf{x}$ e $\mathbf{y}$ appartengono rispettivamente ai sottospazi $\mathcal{U}_i^\perp$ e $\mathcal{V}_j^\perp$, si ha

$$\|A^H\mathbf{x}\|_2 \le \alpha_{i+1}, \quad \text{e} \quad \|B\mathbf{y}\|_2 \le \beta_{j+1}.$$

Sia $\mathcal{Z}$ l'insieme dei vettori $\mathbf{z}$ tali che $U\mathbf{z} \in \mathcal{U}_i^\perp$ e $V\mathbf{z} \in \mathcal{V}_j^\perp$, allora $k = \dim \mathcal{Z} \ge n - (i+j)$ e per il teorema del minimax è

$$\alpha_{i+1}\beta_{j+1} \ge \min_{V_k} \max_{\substack{\|\mathbf{x}\|_2=1 \\ \mathbf{x}\in V_k}} \mathbf{z}^H\Sigma\mathbf{z} = \gamma_{n-k+1} \ge \gamma_{i+j+1}.$$

La corrispondente relazione per la somma si dimostra in modo analogo.)

7.32 Siano A e $B \in \mathbf{C}^{n\times n}$ e siano $\sigma_i \ge \sigma_2 \ge \ldots \ge \sigma_n$ e $\tau_i \ge \tau_2 \ge \ldots \ge \tau_n$ i loro valori singolari. Si dimostri che

$$|\sigma_i - \tau_i| \le \|A - B\|_2, \quad \text{per } i = 1, \ldots, n.$$

(Traccia: si veda il teorema 7.24.)

7.33 Sia $A \in \mathbf{C}^{m\times n}$. Si dimostri che

$$\sigma_1 = \max_{\substack{\mathbf{x}\in\mathbf{C}^n \\ \mathbf{y}\in\mathbf{C}^m \\ \mathbf{x}\neq\mathbf{0} \\ \mathbf{y}\neq\mathbf{0}}} \frac{|\mathbf{y}^HA\mathbf{x}|}{\|\mathbf{x}\|_2\|\mathbf{y}\|_2}.$$

(Traccia: posto $A = U\Sigma V^H$, si considerino i vettori $\mathbf{u} = U^H\mathbf{y}$ e $\mathbf{v} = V^H\mathbf{x}$, per cui è

$$\frac{|\mathbf{y}^HA\mathbf{x}|}{\|\mathbf{x}\|_2\|\mathbf{y}\|_2} = \frac{|\mathbf{u}^H\Sigma\mathbf{v}|}{\|\mathbf{u}\|_2\|\mathbf{v}\|_2} \le \sigma_1\frac{|\mathbf{u}|^T|\mathbf{v}|}{\|\mathbf{u}\|_2\|\mathbf{v}\|_2} \le \sigma_1.$$

Si determinino poi due vettori $\mathbf{u}$ e $\mathbf{v}$ per cui il valore σ_1 viene effettivamente raggiunto.)

7.34 Sia $A \in \mathbf{R}^{n\times n}$ bidiagonale superiore. Si dimostri che se A ha un valore singolare di molteplicità maggiore di 1, allora almeno uno degli elementi della diagonale o della sopradiagonale sono nulli.

(Traccia: la matrice $A^H A$ è tridiagonale con elementi sopradiagonali della forma $\alpha_i\beta_i$, $i = 1, \ldots, n-1$, dove α_i sono gli elementi principali di A e β_i sono gli elementi sopradiagonali. Se $\alpha_i\beta_i \neq 0$, allora gli autovalori di $A^H A$ sono tutti distinti (si veda l'esercizio 6.26).)

7.35 Sia $A \in \mathbf{C}^{n\times n}$ la seguente matrice bidiagonale superiore

$$A = \begin{bmatrix} \alpha_1 & \beta_1 & & & \\ & \alpha_2 & \beta_2 & & \\ & & \ddots & \ddots & \\ & & & \ddots & \beta_{n-1} \\ & & & & \alpha_n \end{bmatrix}.$$

Si costruiscano due matrici di fase S, $T \in \mathbf{C}^{n\times n}$ tali che la matrice $B = SAT$ abbia elementi reali.

(Risposta: $t_{11} = 1, \quad s_{ii} = \dfrac{|\alpha_1|}{\alpha_1},$

$$t_{ii} = \frac{|\beta_{i-1}|}{\beta_{i-1}s_{i-1,i-1}}\,, \quad s_{ii} = \frac{|\alpha_i|}{\alpha_i t_{ii}}\,, \quad \text{per } i = 2, \ldots, n.\ )$$

7.36 a) Si dimostri che una matrice $A \in \mathbf{C}^{n\times n}$ può essere scritta in uno e un sol modo nella forma

$$A = HQ,$$

dove H è semidefinita positiva e Q è unitaria. Tale decomposizione viene detta *decomposizione polare*.

b) Si calcoli la decomposizione polare della matrice

$$A = \begin{bmatrix} 7 & -1 \\ 4 & 8 \end{bmatrix};$$

c) si dimostri che $A^+ = Q^H H^+$.

(Traccia: a) sia $A = U\Sigma V^H$ la decomposizione ai valori singolari di A, si ponga $H = U\Sigma U^H$, $Q = UV^H$; b)

$$H = \sqrt{\frac{2}{5}}\begin{bmatrix} 11 & 2 \\ 2 & 14 \end{bmatrix}, \quad Q = \frac{1}{\sqrt{10}}\begin{bmatrix} 3 & -1 \\ 1 & 3 \end{bmatrix}.)$$

7.37 Siano $A \in \mathbf{C}^{m\times n},\ C \in \mathbf{C}^{r\times n},\ \mathbf{b} \in \mathbf{C}^m,\ \mathbf{d} \in \mathbf{C}^r$ tali che il sistema lineare

$$C\mathbf{y} = \mathbf{d} \tag{45}$$

sia consistente. Si consideri il seguente problema lineare dei minimi quadrati con vincoli di uguaglianza

$$\min_{C\mathbf{y}=\mathbf{d}} \|A\mathbf{y} - \mathbf{b}\|_2. \tag{46}$$

La soluzione di questo problema è quindi un vettore $\mathbf{x}$ che soddisfa il sistema (45) e che minimizza la norma 2 del residuo. Indicati con $s \leq \min\{r, n\}$ il rango di C e con $C = U\Sigma V^H$ la decomposizione ai valori singolari di C, si dimostri che

a) l'insieme delle soluzioni del sistema (45) è dato da

$$Y = \{\ \mathbf{y} \in \mathbf{C}^n\ :\ \mathbf{y} = C^+\mathbf{d} + V_2\mathbf{z},\ \mathbf{z} \in \mathbf{C}^{n-s}\ \},$$

dove $V_2 \in \mathbf{C}^{n\times(n-s)}$ è la matrice formata dalle ultime $n - s$ colonne di V;

b) $V_2(AV_2)^+ = (AZ)^+$, dove $Z = I - C^+C$;

c) la soluzione di minima norma del problema (46) è data da

$$\mathbf{x}^* = C^+\mathbf{d} + (AZ)^+(\mathbf{b} - AC^+\mathbf{d});$$

d) la soluzione $\mathbf{x}^*$ è l'unica soluzione di (46) se e solo se la matrice

$$\begin{bmatrix} C \\ A \end{bmatrix}$$

ha rango n;

e) posto $\mathbf{g} = \begin{bmatrix} \mathbf{d} \\ \mathbf{b} \end{bmatrix}$, si determini la matrice $F \in \mathbf{C}^{(m+r)\times n}$ tale che la soluzione di minima norma $\mathbf{x}^*$ del problema (46) possa essere espressa come

$$\mathbf{x}^* = F^+\mathbf{g};$$

f) si applichi il metodo QR per il calcolo di una soluzione del problema (46) nell'ipotesi che $s = r < n$.

(Traccia: a) basta dimostrare che $CV_2 = O$, infatti

$$\begin{aligned} CV_2 &= U\Sigma \begin{bmatrix} V_1^H \\ V_2^H \end{bmatrix} V_2 = U\Sigma \begin{bmatrix} V_1^H V_2 \\ V_2^H V_2 \end{bmatrix} = U\Sigma \begin{bmatrix} O \\ I_{n-s} \end{bmatrix} \\ &= U \begin{bmatrix} \Sigma_1 \\ O \end{bmatrix} \begin{bmatrix} O \\ I_{n-s} \end{bmatrix} = O; \end{aligned}$$

b) si dimostri che

$$Z = I - C^+C = I - V_1V_1^H = V_2V_2^H,$$

e si verifichi che

$$(AV_2V_2^H)^+ = V_2(AV_2)^+$$

per mezzo delle equazioni di Moore-Penrose; c) la soluzione $\mathbf{x}^*$ di minima norma di (46) può essere espressa come

$$\mathbf{x}^* = C^+\mathbf{d} + V_2\mathbf{z}^*,$$

dove $\mathbf{z}^*$ è la soluzione di minima norma di

$$\min_{\mathbf{z}\in\mathbf{C}^{n-s}} \|A(C^+\mathbf{d} + V_2\mathbf{z}) - \mathbf{b}\|_2 = \min_{\mathbf{z}\in\mathbf{C}^{n-s}} \|AV_2\mathbf{z} - (\mathbf{b} - AC^+\mathbf{d})\|_2, \qquad (47)$$

si dimostri infatti che i vettori $C^+\mathbf{d}$ e $V_2\mathbf{z}^*$ sono ortogonali. Inoltre risulta

$$\mathbf{z}^* = (AV_2)^+(\mathbf{b} - AC^+\mathbf{d});$$

d) si consideri la matrice

$$\begin{bmatrix} C \\ A \end{bmatrix} V = \begin{bmatrix} C \\ A \end{bmatrix} [V_1 \mid V_2] = \begin{bmatrix} CV_1 & CV_2 \\ AV_1 & AV_2 \end{bmatrix} = \begin{bmatrix} CV_1 & O \\ AV_1 & AV_2 \end{bmatrix}.$$

Poiché la matrice $CV_1 \in \mathbf{C}^{r\times s}$ ha rango s, la matrice $\begin{bmatrix} C \\ A \end{bmatrix}$ ha rango n se e solo se $r + m \geq n$ e la matrice AV_2 ha rango $n - s$, cioè se e solo se il problema (47) ha la sola soluzione $\mathbf{z}^*$; e)

$$\mathbf{x}^* = [C^+ - (AZ)^+AC^+ \mid (AZ)^+] \begin{bmatrix} \mathbf{d} \\ \mathbf{b} \end{bmatrix},$$

per cui

$$F^+ = [C^+ - (AZ)^+AC^+ \mid (AZ)^+];$$

si dimostri che la matrice

$$F = \begin{bmatrix} C \\ (AZ)(AZ)^+A \end{bmatrix}$$

verifica le equazioni di Moore-Penrose, sfruttando le relazioni

$$CV_2 = V_2^HC^+ = O, \quad (AZ)^+ = V_2(AV_2)^+, \quad AZC^+ = C(AZ)^+ = O,$$

$$A(AZ)^+ = AV_2(AV_2)^+ = AV_2V_2^HV_2(AV_2)^+ = AZ(AZ)^+,$$
$$(C^+C)^H = C^+C, \quad (CC^+)^H = CC^+,$$
$$[AZ(AZ)^+]^H = AZ(AZ)^+, \quad [(AZ)^+AZ]^H = (AZ)^+AZ;$$

f) sia $Q \in \mathbf{C}^{n\times n}$ unitaria tale che $C^H = QR$, $R \in \mathbf{C}^{n\times r}$ triangolare superiore, dove

$$R = \begin{bmatrix} R_1 \\ \\ O \end{bmatrix} \begin{matrix} \} & r & \text{righe,} \\ \\ \} & n-r & \text{righe,} \end{matrix}$$

e R_1 è non singolare; allora

$$\begin{bmatrix} C \\ \\ A \end{bmatrix} Q = \begin{bmatrix} R_1^H & O \\ \\ A_1 & A_2 \end{bmatrix}.$$

Si calcola la generica soluzione del sistema $C\mathbf{y} = \mathbf{d}$ per mezzo del sistema equivalente $[R_1^H \mid O]Q^H\mathbf{y} = \mathbf{d}$, ottenendo

$$Q^H\mathbf{y} = \begin{bmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \end{bmatrix} = \begin{bmatrix} R_1^{-H}\mathbf{d} \\ \mathbf{z}_2 \end{bmatrix}, \quad \mathbf{z}_2 \in \mathbf{C}^{n-r}.$$

Sostituendo si ha

$$A\mathbf{y} - \mathbf{b} = AQQ^H\mathbf{y} - \mathbf{b} = A_2\mathbf{z}_2 - (\mathbf{b} - A_1\mathbf{z}_1).$$

Si calcola $\mathbf{z}_2$ come soluzione del problema dei minimi quadrati

$$\min_{\mathbf{z}\in\mathbf{C}^{n-r}} \|A_2\mathbf{z} - (\mathbf{b} - A_1\mathbf{z}_1)\|_2$$

e si ricava

$$\mathbf{y} = Q \begin{bmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \end{bmatrix}.)$$

Commento bibliografico

Il metodo dei minimi quadrati, che si basa su alcuni metodi più o meno empirici usati dagli astronomi nel 18^o secolo, è stato pubblicato per la prima volta da Legendre nel 1805, ma già dal 1795 era stato usato da Gauss per determinare le orbite dei corpi celesti. Comunque è stato Gauss il primo che nel 1809 ha individuato i rapporti fra il metodo dei minimi quadrati e la teoria della probabilità. Questo metodo riveste una grande importanza

storica, perché è proprio per risolvere i sistemi lineari generati da esso che sono state sollevate molte delle più importanti problematiche dell'algebra lineare numerica. Il metodo che Gauss suggerisce per risolvere il problema dei minimi quadrati è quello della risoluzione del sistema normale.

Problemi di minimi quadrati si incontrano in tutte le aree della scienza e della tecnica: si veda ad esempio il libro [1], in cui è riportato un problema di aggiustamento di dati geodetici, un problema di ricostruzione di immagini e un problema di trasporto. Per questo tipo di problemi, in cui si possono presentare moltissime equazioni, anche dell'ordine di milioni, vengono utilizzate particolari tecniche di partizionamento e sfruttate tutte le specifiche proprietà di struttura della matrice, in particolare la sparsità, che deve essere mantenuta per quanto possibile, in ogni fase della risoluzione (si veda ad esempio [2]).

La trattazione più esauriente sul problema dei minimi quadrati lineari e dei metodi numerici per risolverlo è quella di Lawson e Hanson che nel loro libro [12] del 1974 hanno raccolto in maniera sistematica i contributi dati alla risoluzione del problema nei venti anni precedenti. In questo periodo si è sviluppata una intensa attività sia dal punto di vista teorico che da quello computazionale, a partire dal 1954, anno in cui Givens [6] suggerisce di usare trasformazioni ortogonali operando su matrici non quadrate per risolvere il problema dei minimi quadrati, e dal 1958, anno in cui Householder [11] propone di usare le matrici che portano il suo nome. Ma il nome che ricorre più frequentemente in questo campo è quello di Golub, a cui si deve l'uso dei metodi basati sui valori singolari [7] e le tecniche numeriche fondamentali [8]. In [12] si trova anche l'indicazione che il metodo QR consente di risolvere in pratica una classe più ampia di problemi di quelli risolti con il sistema normale, cioè tutti quelli per cui il condizionamento della matrice A è dell'ordine della radice della precisione di macchina.

Tentativi per definire un'estensione del concetto di inversa di una matrice, valido anche nel caso di matrici singolari, sono stati fatti varie volte nell'ultimo secolo. Moore [13] fu il primo nel 1920 ad affrontare uno studio sistematico delle inverse generalizzate, ma il suo lavoro non ricevette alcuna attenzione fino al 1950. Nel 1955 Penrose [15], che però non era a conoscenza del lavoro di Moore, riscoprì le inverse generalizzate e dette un tale impulso a questo argomento, che nel 1976 Nashed nel suo libro [14] potè elencare più di 1700 titoli di bibliografia sulle inverse generalizzate.

La decomposizione ai valori singolari, che è una delle decomposizioni più importanti del calcolo matriciale e che consente di trattare in modo profondo il problema della determinazione pratica del rango di una matrice, è stata introdotta e formalizzata negli anni '30 da Eckart e Young, che nel 1936 e nel 1939 [5] hanno formulato e dimostrato i teoremi fondamentali della decomposizione ai valori singolari per matrici qualsiasi. Il

caso particolare della decomposizione di una matrice quadrata ad elementi reali era stato dimostrato da Sylvester nel 1889. La prima utilizzazione dei valori singolari si è avuta nell'ambito della localizzazione degli autovalori. La limitazione $\sigma_n \leq |\lambda| \leq \sigma_1$ è stata dimostrata da Browne nel 1928, ma si riallaccia ad un risultato di Bendixson che nel 1900 aveva dimostrato che gli autovalori di una matrice reale A sono compresi in modulo fra il massimo e il minimo autovalore di $\frac{1}{2}(A + A^T)$. Secondo O. Taussky, il termine *valore singolare* è dovuto a Weyl, che nel 1949 [16] ha dimostrato alcune importanti relazioni fra gli autovalori e i valori singolari di una matrice. Negli anni '50 una notevole mole di risultati riguardanti relazioni fra autovalori e valori singolari delle matrici A, $\frac{1}{2}(A + A^H)$ e $\frac{1}{2\mathrm{i}}(A - A^H)$ e fra prodotti e somme di matrici, sono dovuti a Horn e Amir-Moez.

I metodi per il calcolo della decomposizione ai valori singolari sono descritti in [9]: il metodo di Golub e Reinsch [8] per la riduzione della matrice A a forma bidiagonale, la variante di Chan [3], utile quando m è maggiore di n, e la tecnica per calcolare gli autovalori della matrice tridiagonale $B^T B$ operando con il metodo QR direttamente sulla matrice B.

Per l'applicazione del metodo del gradiente coniugato alla risoluzione del problema dei minimi quadrati si veda [10], per l'applicazione del metodo di Lanczos al calcolo dei valori singolari di una matrice si veda [4].

Bibliografia

[1] Å. Björk, R. J. Plemmons, H. Schneider, *Large Scale Matrix Problems*, North Holland, New York, 1981.

[2] Å. Björk, I. S. Duff, "Sparse Linear Least Squares Problems", in *Large Scale Matrix Problems*, North Holland, New York, 1981.

[3] T. F. Chan, "An Improved Algorithm for Computing the Singular Value Decomposition", *ACM Trans. Math. Soft.*, 8, 1982, pp. 72-83.

[4] J. Cullum, R. A. Willoughby, *Lanczos Algorithms for Large Symmetric Eigenvalue Computations*, vol. I, Theory, Progress in Scientific Computing, 3., Birkhäuser, Boston, 1985.

[5] C. Eckart, G. Young, "A Principal Axis Transformation for Non-Hermitian Matrices", *Bull. Amer. Math. Soc.*, 45, 1939, pp. 118-121.

[6] W. Givens, "Numerical Computation of Characteristic Values of a Real Symmetric Matrix", Oak Ridge National Laboratory, *ORN* L-1574, 1954.

[7] G. H. Golub, W. Kahan, "Calculating the Singular Values and Pseudo-Inverse of a Matrix", *SIAM J. Num. Anal. Ser. B 2*, 1965, pp. 205-224.

[8] G. H. Golub, C. Reinsch, "Singular Value Decomposition and Least Squares Solutions", *Numer. Math.*, 14, 1970, pp. 403-420.

[9] G. H. Golub, C. F. Van Loan, *Matrix Computations*, 2nd Edition, The Johns Hopkins University Press, Baltimore, Maryland, 1989.

[10] M. R. Hestenes, *Conjugate Direction Methods in Optimization*, Springer-Verlag, New York, 1980.

[11] A. S. Householder, "Unitary Triangularization of a Nonsymmetric Matrix", *J. Assoc. Comp. Mach.*, 5, 1958, pp. 339-342.

[12] C. L. Lawson, R. J. Hanson, *Solving Least Squares Problems*, Prentice-Hall, Englewood Cliffs, 1980.

[13] E. H. Moore, *General Analysis. Part I*, American Philosophical Society, Philadelphia, 1935.

[14] M. Z. Nashed, *Generalized Inverses and Applications*, Academic Press, New York, 1976.

[15] R. Penrose, "A Generalized Inverse for Matrices", *Proc. Cambridge Philos. Soc.*, 51, 1955, pp. 406-413.

[16] H. Weyl, "Inequalities Between the Two Kinds of Eigenvalues of a Linear Transformation", *Proc. Nat. Acad. Sci. USA*, 35, 1949, pp. 408-411.

Bibliografia generale

A. C. Aitken, *Determinants and Matrices*, Interscience, New York, 1956.

K. E. Atkinson, *An Introduction to Numerical Analysis*, John Wiley and Sons, New York, 1978.

D. Bini, M. Capovani, G. Lotti, F. Romani, *Complessità numerica*, Boringhieri, Torino, 1981.

J. Cullum, R. A. Willoughby, *Lanczos Algorithms for Large Symmetric Eigenvalue Computations*, vol. I, Theory, Progress in Scientific Computing, 3., Birkhäuser, Boston, 1985.

D. K. Faddeev, V. N. Faddeeva, *Computational Methods of Linear Algebra*, Freeman and Co., San Francisco, 1963

G. E. Forsythe, M. A. Malcom, C. B. Moler, *Computer Methods for Mathematical Computations*, Prentice Hall, Englewood Cliffs, N. J., 1977.

G. E. Forsythe, C. B. Moler, *Computer Solution of Linear Algebraic Systems*, Prentice Hall, Englewood Cliffs, N. J., 1967.

F. R. Gantmacher, *The Theory of Matrices*, vol. I e II. Chelsea, New York, 1959.

G. H. Golub, C. F. Van Loan, *Matrix Computations*, 2nd Edition, The Johns Hopkins University Press, Baltimore, Maryland, 1989.

L. A. Hageman, D. M. Young, *Applied Iterative Methods*, Academic Press, New York, 1981.

P. R. Halmos, *Finite-Dimentional Vector Spaces*, Van Nostrand-Reinhold, Princeton, 1958.

M. R. Hestenes, *Conjugate Direction Methods in Optimization*, Springer-Verlag, New York, 1980.

A. S. Householder, *The Theory of Matrices in Numerical Analysis*, Blaisdell, Boston, 1964.

E. Isaacson, H. B. Keller, *Analysis of Numerical Methods*, John Wiley and Sons, New York, 1966.

L. I. Kronsjö, *Algorithms, their Complexity and Efficiency*, J. Wiley and Sons, New York, 1979.

P. Lancaster, M. Tismenetsky, *The Theory of Matrices*, Academic Press New York, 1985.

C. L. Lawson, R. J. Hanson, *Solving Least Squares Problems*, Prentice Hall, Englewood Cliffs, N. J., 1974.

C. C. MacDuffee, *The Theory of Matrices*, Chelsea, New York, 1946.

T. Muir, *Theory of Determinants in the Historical Order of Development*, Dover, London, 1906.

T. Muir, *A Treatise on the Theory of Determinants*, Dover, London, 1933.

M. Z. Nashed, *Generalized Inverses and Applications*, Academic Press, New York, 1976.

A. M. Ostrowski, *Solution of Equations and Systems of Equations*, Academic Press, 1960.

B. N. Parlett, *The Symmetric Eigenvalue Problem*, Prentice-Hall, Englewood Cliffs, 1980.

M. C. Pease, *Methods of Matrix Algebra*, Academic Press, New York, 1965.

H. R. Schwarz, H. Rutishauser, E. Stiefel, *Numerical Analysis of Symmetric Matrices*, Prentice-Hall, Englewood Cliffs, 1973.

G. W. Stewart, *Introduction to Matrix Computation*, Academic Press, New York, 1973.

J. Stoer, R. Bulirsch, *Introduction to Numerical Analysis*, Springer-Verlag, New York, 1980.

R. S. Varga, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, New Jersey, 1962.

J. H. Wilkinson, *Rounding Errors in Algebraic Processes*, Prentice Hall, Englewood Cliffs, N. J., 1963.

J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.

J. H. Wilkinson, C. Reinsch, *Handbook for Automatic Computation, vol. 2, Linear Algebra*, Springer-Verlag, New York, 1971.

D. M. Young, *Iterative Solution of Large Linear Systems*, Academic Press, New York, 1971.

A-coniugati, vettori, 275, 277, 308-312
a-posteriori, stima dell'errore, 207
accelerazione di Ritz, 389
aggiunta, matrice, 12, 34, 50
albero, matrice ad, 40, 100, 299, 409
algebrica, molteplicità, 53, 54, 57, 59, 321
algoritmico, errore, 136, 177
algoritmo di Golub e Reinsch, 466
alternativa di Fredholm, 481
ampiezza di banda, 27, 210
analisi dell'errore, 136
    all'indietro, 140, 165
    del metodo di Gauss per la fattorizzazione LU, 167
    del metodo di Gauss per la risoluzione di un sistema lineare, 170
    del metodo di Householder, 187
    nella risoluzione del sistema triangolare, 165
analitico, errore, 136
angolo di due vettori, 5, 410
antihermitiana, matrice, 26, 34
antisimmetrica, matrice, 26, 30, 33
approssimazione ai minimi quadrati, polinomio di, 477
arco orientato, 18
aritmetica intera modulo p, 224
arrotondamento, errore di, 164
asintotica media, riduzione, 239
asintotico, tasso di convergenza, 241
assoluta, norma, 131, 319
autovalore, 45-107
    numero di condizionamento di un, 403
autovalori
    condizionamento del problema del calcolo degli, 320
    del prodotto di Kronecker, 102, 270
    di una matrice definita positiva, 73
    di una matrice hermitiana, 68
    di una matrice normale, 70, 448
    di una matrice tridiagonale, 100, 270
    di una matrice unitaria, 48, 87
    dominanti, 387
    generalizzati, 425
    metodi iterativi per, 331, 353, 367
    metodo di Jacobi per, 333, 367-370, 428
    metodo di Lanczos per, 333, 391-397, 429
    problema generalizzato agli, 425-427
    teoremi di localizzazione degli, 76-80, 313, 316-319, 401
    teoremi di perturbazione, 319
    teoremi di separazione, 324-330
    teoria della perturbazione, 136, 137, 428
autovettore, 45-107
    numero di condizionamento di un, 403
autovettori
    dominanti, 387
    linearmente indipendenti, 52-53, 57, 371
    ortonormali, 69, 70

banda
    ampiezza di, 27, 210
    matrice a, 27, 210, 213, 220
base
    cambiamento di, 55
    canonica, 6, 9, 110
    della rappresentazione, 164
    di un sottospazio, 6
    ortonormale, 9, 56, 63
Bauer-Fike, teorema di, 319
ben condizionata, matrice, 138
ben condizionato, problema, 136
ben posto, problema, 136
Binet, regola di, 12, 33
bisezione, procedimento di, 347
blocchi
    matrice a, 16, 17, 30, 31, 34, 41, 59, 66, 89-93, 129, 133, 257, 300, 489-490
    matrice tridiagonale a, 221, 259, 269
    predominanza diagonale a, 222
Bodewig, matrice di, 415
Brauer, 107
Brioschi, 106
Browne, 498

calcolo
    degli autovalori, 334-397
    dei valori singolari, 448, 479
    del determinante, 178
    del rango, 179, 190, 454, 473

della forma normale di Schur di $A^H A$, 450, 466
della matrice inversa con i vettori A-coniugati, 309
della matrice inversa con un metodo diretto, 163, 181, 186
della matrice inversa con un metodo iterativo, 294
cambiamento di base, 55
cammino orientato, 19
canonica, base, 6, 9, 110
caratteristica, equazione, 45, 54, 105
caratteristico, polinomio, 45, 51, 83, 100, 101
Cassini, ovale di, 104, 107
Cauchy, 42, 106, 314, 428
Cauchy-Schwartz, disuguaglianza di, 4, 28, 109, 112, 117, 123, 492
Cayley, 42
Cayley-Hamilton, teorema di, 50, 106
centroantisimmetrici, vettori, 96
centrosimmetrica, matrice, 96, 97, 219
centrosimmetrici, vettori, 96
cerchi di Gerschgorin, 76, 134, 402
unione disgiunta di, 79
Chebyshev, 135
polinomio di, 395
Cholesky
costo computazionale del metodo di, 200, 223
fattorizzazione incompleta di, 283, 286
fattorizzazione LL^H con il metodo di, 198, 435
metodo di, 143, 149, 198, 218, 226, 286, 435, 478
ciclica, strategia per il metodo di Jacobi, 369
cifre della rappresentazione, 164
circolante, matrice, 40, 98
Clasen, 226
classica, strategia per il metodo di Jacobi, 369
colonna, vettore, 4, 17
colonne linearmente indipendenti di una matrice, 13, 29, 57
commutative, matrici, 21, 25, 87, 89
compatte, tecniche, 149, 196
complemento di Schur, 31, 201, 211, 219, 306
completa, riortogonalizzazione, 396
componenti di un, vettore, 4
condizionamento
del problema dei minimi quadrati, 459
del problema del calcolo degli autovalori, 320
di un autovalore, 403
di un autovettore, 403
di una matrice, 137, 140, 176, 203-208, 226, 410, 458, 466
condizione di arresto
del metodo del gradiente coniugato, 278, 475
del metodo delle potenze, 374
del metodo QR, 362
di un metodo iterativo, 241
coniugata, matrice trasposta, 1, 47
coniugato, v. gradiente coniugato
connesso, grafo fortemente, 19
consistente, sistema lineare, 15
consistentemente ordinata, matrice, 299
continuità della norma, 110
controllo della convergenza, 238
convergente
metodo, 236
successione, 231, 289
convergenza
controllo della, 238
del metodo del gradiente coniugato, 279
del metodo dello steepest descent, 274, 307
del metodo di Gauss-Seidel, 247, 250, 296, 303
del metodo di Jacobi per il calcolo degli autovalori, 369
del metodo di Jacobi per il sistema lineare, 247
del metodo di rilassamento, 262, 263
di un metodo iterativo, 236, 237
di una successione di matrici, 232, 234, 289, 290, 412
per il metodo del quoziente di Rayleigh, 420
per il metodo QR, 355, 360, 363, 414
per matrici tridiagonali, 254, 265
per matrici tridiagonali a blocchi, 259, 269
tasso asintotico di, 241
convesso, insieme, 25, 122, 433
correzione post-iterativa, metodo di, 295
costo computazionale, 141
dei metodi diretti, 201

dei metodi iterativi, 242
dei metodi per i minimi quadrati, 435, 439, 461
del metodo di Cholesky, 200, 223
del metodo di Gauss, 159, 162, 163, 200, 223, 352
del metodo di Gauss-Jordan, 180
del metodo di Givens, 194, 220, 337
del metodo di Householder, 185, 220, 335, 439
del metodo di Lanczos, 341
del metodo QR, 358
della risoluzione del sistema triangolare, 143
Courant-Fisher, teorema di, 323
Cramer, 42
regola di, 15
Crout, metodo di, 149, 197, 226

decomposizione
ai valori singolari, 444, 447, 497
polare, 493
definita negativa, matrice, 10
definita positiva, matrice, 10, 13, 25-28, 32, 41, 73, 74, 82, 93, 94, 117, 127, 148, 198, 211, 212, 250, 309, 408, 426
autovalori di una, 73
metodo di Gauss-Seidel, convergenza per una, 250
sottomatrici principali di una, 11
sottomatrici principali di testa di una, 74
teorema di Ostrowsky-Reich per una, 263
teorema di Stein per una, 302
deflazione
implicita, 345
variante della, 385
delta di Kronecker, 8
determinante, 11, 35, 42, 57, 73, 74, 237
calcolo del, 178
diade, 14, 91, 128, 485
diagonale
a blocchi, matrice a predominanza, 222
elemento, 1
in senso stretto, matrice a predominanza, 82, 211, 212, 247, 296
matrice, 1, 13, 47, 131, 180
matrice a predominanza, 82, 247, 258, 303
diagonalizzabile, matrice, 57, 69-71, 88, 94, 120, 130
dimensione di un sottospazio, 6, 90
Dirichlet, problema di, 283, 314
discreta di Fourier, trasformata , 99
distanza fra sottospazi, 387, 422
disuguaglianza
di Cauchy-Schwartz, 4, 28, 109, 112, 117, 123, 492
di Hadamard, 93, 94, 225
di Hölder, 123, 134
di Kantorovich, 307
di Minkowski, 123, 134
triangolare, 108
dominanti, autovalori e autovettori, 387

elementare, matrice, 149, 154, 227, 332
di Gauss, 150, 157, 332, 350
di Householder, 151, 183, 332, 334, 349, 428
fattorizzazione mediante, 154
elemento
diagonale, 1
principale, 1, 63, 82
eliminazione, metodo di, 157, 160
equazione caratteristica, 45, 54, 105
equazioni
di Moore-Penrose, 457, 484
lineari, sistemi di, 15, 136-315
normali, 433
equivalenza delle norme, 110, 111, 121, 131
errore
algoritmico, 136, 177
analisi dello, 136, 140, 165, 167, 170, 187
analitico, 136
di arrotondamento, 164
di troncamento, 136
inerente, 136, 177
riduzione media per passo, 239
stima a-posteriori, 207
esponente della rappresentazione, 164
esponenziale di una matrice, 103, 131
esterno, prodotto, 4

fase, matrice di, 148, 220, 356, 413, 468, 493
fattorizzazione
incompleta di Cholesky, 283, 286
LDL^H, 218
LDR, 216
LL^H, 143, 148, 198, 217, 435

LU, 143, 144, 148, 157, 159, 162, 167, 170, 208, 210, 213, 356, 436
LU a blocchi, 221
mediante le matrici elementari, 154
QR, 144, 148, 182, 187, 191
Fibonacci
matrice di, 35
numeri di, 36
fill-in, 231
forma canonica o normale
di Jordan, 59, 119, 232, 240
di Schur, 63, 69, 106, 289, 449, 450, 466
reale di Jordan, 61
reale di Schur, 66, 95
formula
di Sherman-Morrison, 33, 214
di Woodbury, 33
fortemente connesso, grafo, 19
Fourier, trasformata discreta di, 99
Francis, procedimento di, 365
Fredholm, alternativa di, 481
Frobenius, 43, 106, 135
matrice di, 81, 85
norma di, 117, 118, 226, 443, 448

Gauss, 226, 313, 496
matrice elementare di, 150, 157, 332, 350
Gauss, metodo di, 143, 149, 157-179, 226, 228, 286
analisi dell'errore per la fattorizzazione LU, 167
analisi dell'errore per la risoluzione di un sistema lineare, 170
costo computazionale del, 159, 162, 163, 200, 223, 352
implementazione, 178
per il calcolo del determinante, 178
per il calcolo del rango, 179
per il calcolo dell'inversa, 163
per la fattorizzazione LU, 157
per la riduzione in forma di Hessenberg superiore, 350
per la risoluzione del sistema lineare, 159
Gauss-Jordan, metodo di, 180-181, 226
costo computazionale del, 180
per il calcolo dell'inversa, 181
Gauss-Seidel, metodo di, 242, 252, 286, 313
condizione di convergenza, 247, 296, 303
per matrici a blocchi, 257, 259
per matrici definite positive, 250
per matrici tridiagonali, 254
generalizzata, norma, 132-134
generalizzato
autovalore, 425
problema agli autovalori, 425-427
geometrica, molteplicità, 53, 54, 57, 59
Gerschgorin, 106
cerchi di, 76, 134, 402
teoremi di, 76-80
Givens, 226, 497
matrice di, 191, 336, 367, 428
Givens, metodo di
costo computazionale del, 194, 220, 337
per la fattorizzazione QR, 191-196
per la riduzione in forma di Hessenberg superiore, 349
per la tridiagonalizzazione, 336
Goldstine, 223, 428
Golub, 497
e Reinsch, algoritmo di, 466
Gram-Schmidt, metodo di ortogonalizzazione di, 9
gradiente, 273
gradiente coniugato, metodo del, 272, 275, 286
con precondizionamento, 281
condizione di arresto, 278, 475
convergenza del, 279
per il problema dei minimi quadrati, 474, 478
tecniche di precondizionamento, 281, 286
grado di nilpotenza, 24
grafo orientato, 18
fortemente connesso, 19
Grassman, 43

Hadamard, 107, 226
disuguaglianza di, 93, 94, 225
Hamilton, 42, 106
Hankel, matrice di, 175, 461
Hermite, 43, 106
hermitiana, matrice, 2, 13, 25, 34, 41, 68, 73, 74, 82, 93, 116, 140, 151, 211, 322, 343, 386, 408, 411, 426
Hessenberg superiore, matrice in forma di, 210, 212, 213, 331, 349, 350, 353, 411, 428
Hilbert, matrice di, 138, 304

Hirsch, teorema, 316
Hölder, disuguaglianza di, 123, 134
hölderiana, norma, 123, 134
Householder, 226, 290, 313, 497
  matrice elementare di, 151, 183, 332, 334, 349, 428
Householder, metodo di
  analisi dell'errore, 187
  costo computazionale, 185, 220, 335, 439
  implementazione, 185, 186, 335
  nell'algoritmo di Golub e Reinsch, 467
  per i minimi quadrati, 438, 497
  per la fattorizzazione QR, 144, 149, 182-191, 286
  per la riduzione in forma di Hessenberg superiore, 349
  per la tridiagonalizzazione, 334-336, 411
  per il calcolo del rango, 190
  per il calcolo dell'inversa, 186
Hyman, metodo di, 352

idempotente, matrice, 24, 29, 94
identica, matrice, 2, 116
immagine, 13, 30, 433, 448
implementazione
  del metodo di Gauss, 178
  del metodo di Householder, 185, 186, 335
implicita, deflazione, 345
incompleta di Cholesky, fattorizzazione, 283, 286
indotta, norma, 114, 119, 239
inerente, errore, 136
insieme convesso, 25, 122, 433
intera, aritmetica modulo p, 224
interno, prodotto, 4
invariante
  dominante, sottospazio, 387
  sottospazio, 393
inversa, matrice, 12, 47, 118
  calcolo della, 163, 181, 186
  con i vettori A-coniugati, calcolo della, 309
  con un metodo iterativo, calcolo della, 294
inverse, metodo delle potenze, 333, 378, 419
inversione, operazione di, 12, 13
involutoria, matrice, 22, 24
irriducibile, matrice, 17-20, 80, 81, 247, 258, 303
iterativi, metodi
  condizione di arresto, 241
  convergenza dei, 236, 237
  costo computazionale dei, 242
  per gli autovalori, 331, 353, 367
  per i sistemi lineari, 136, 231, 235-315
iterativo
  metodo per il calcolo dell'inversa, 294
  raffinamento, 294
iterazione, matrice di, 236
iterazioni
  del quoziente di Rayleigh, metodo delle, 381, 420
  di sottospazi, metodo delle, 386, 400, 422
  ortogonali, metodo delle, 386, 400, 422

Jacobi, 42, 106, 313, 428
  metodo per gli autovalori, 333, 367-370, 428
  metodo per la risoluzione dei sistemi lineari, 242, 247, 286, 313
  per matrici a blocchi, 257, 259
Jacobi, matrice di, v. matrice tridiagonale
Jordan C., 106, 226
  forma canonica o normale di, 59, 119, 232, 240
  forma normale reale di, 61
Jordan W., 226

Kahan, teorema di, 262
Kantorovich, disuguaglianza di, 307
Kronecker
  delta di, 8
  prodotto di, 40, 41, 102, 135, 270, 300

Lagrange, 105
Laguerre, 43
Lanczos
  costo computazionale, 341
  metodo per il calcolo degli autovalori, 333, 391-397, 429
  metodo per il calcolo dei valori singolari, 479
  metodo per la tridiagonalizzazione, 339
  vettori, 391, 393
Laplace, 42
  regola di, 11, 12, 34, 36, 101, 344

Leibniz, 42
Legendre, 496
legge del parallelogramma, 28
Leverrier, 429
Levy, 106
L'Hospital, 42
linearmente dipendenti
colonne o righe di una matrice, 12, 14
vettori, 6
linearmente indipendenti
autovettori, 52-53, 57, 371
colonne o righe di una matrice, 13, 29, 57
vettori, 6, 28
Liouville, 105
localizzazione degli autovalori, teoremi di, 76-80, 133, 316-319, 401
LR, metodo, 354, 428
lunghezza euclidea di un vettore, 5

M-matrice, 306
macchina
numero di, 164
operazioni di, 165
precisione di, 164
Maehly, variante di, 345
mal condizionata, matrice, 138, 140
mal condizionato, problema, 136
mal posto, problema, 136
mantissa della rappresentazione, 164
Markov, 106
massimo pivot, 172, 177, 215
parziale, 172, 177
per colonne, 190, 440
totale, 174
massimo, rango, 15, 435, 443, 476, 486
matrice, 1, 43
a banda, 27, 210, 213, 220
a blocchi, 16, 17, 30, 31, 34, 41, 59, 66, 89-93, 129, 133, 257, 33, 489-490
a predominanza diagonale, 82, 247, 258, 303
a predominanza diagonale in senso stretto, 82, 211, 212, 247, 296
ad albero, 40, 100, 214, 299, 409
aggiunta, 12, 34, 50
antihermitiana, 26, 34
antisimmetrica, 26, 30, 33
ben condizionata, 138
centrosimmetrica, 96, 97, 219
circolante, 40, 98
consistentemente ordinata, 299
definita negativa, 10
definita positiva, 10, 13, 25-28, 32, 41, 73, 74, 82, 93, 94, 117, 127, 148, 198, 211, 212, 250, 263, 302, 309, 408, 426
di Bodewig, 415
di fase, 148, 220, 356, 413, 468, 493
di Fibonacci, 35
di Frobenius, 81, 85
di Givens, 191, 336, 367, 428
di Hankel, 175, 461
di Hilbert, 138, 304
di iterazione, 236
di Jacobi, v. matrice tridiagonale
di permutazione, 3, 17, 39, 145, 161, 174, 190, 214, 217, 441, 451
di riflessione, 151
di Toeplitz, 97
di Vandermonde, 39, 223, 461
diagonale, 1, 13, 47, 131, 180
diagonalizzabile, 57, 69-71, 88, 94, 120, 130
elementare, 149, 154, 227, 332
elementare di Gauss, 150, 157, 332, 350
elementare di Householder, 151, 183, 332, 334, 349, 428
esponenziale di una, 103, 131
hermitiana, 2, 13, 25, 34, 41, 68, 73, 74, 82, 93, 116, 140, 151, 211, 322, 343, 386, 408, 411, 426
idempotente, 24, 29, 94
identica, 2, 116
in forma di Hessenberg superiore, 210, 212, 213, 331, 349, 350, 353, 411, 428
inversa, 12, 47, 118, 294, 309
involutoria, 22, 24
irriducibile, 17-20, 80, 81, 247, 258, 303
mal condizionata, 138, 140
nilpotente, 24, 95
non quadrata, 433, 443
normale, 2, 13, 25, 69, 70, 89, 93, 317, 318, 401, 448
ordine di una, 1
ortogonale, 2, 66, 93, 153
persimmetrica, 97, 219
polinomio di una, 21, 48, 72, 89
quadrata, 1
radice quadrata di una, 484
riducibile, 17-20
scalare, 1

semidefinita negativa, 10
semidefinita positiva, 10, 115, 471
simmetrica, 2, 69,391, 409
singolare, 12, 31
spettro di una, 45
trasposta, 1, 47
trasposta coniugata, 1, 47
triangolare, 1, 13, 25, 47, 63, 141, 143, 156, 203, 208
triangolare in senso stretto, 1, 24
tridiagonale, 1, 27, 35, 37, 100, 158, 212, 213, 254, 265, 270, 299, 343, 394, 400, 411, 413, 471
tridiagonale a blocchi, 221, 259, 269
unitaria, 2, 13, 25, 27, 41, 48, 56, 63, 68, 70, 87, 93, 98, 112, 117, 128, 151
matrici
commutative, 21, 25, 87, 89
operazioni fra, 4, 12, 13
simili, 56, 57
successione di, 231-234, 289, 290, 412
matriciale, norma, 113
media
per passo dell'errore, riduzione, 239
riduzione asintotica, 239
metodi diretti
per i sistemi lineari, 136-228 , 231
metodi iterativi
per gli autovalori, 331, 353, 367
per i sistemi lineari, 136 , 231, 235-315
metodo
convergente, 236
degli spostamenti simultanei, 243
degli spostamenti successivi, 243
del gradiente coniugato, v. gradiente coniugato
del quoziente di Rayleigh, 381, 420
delle iterazioni di sottospazi, 386, 400, 422
delle iterazioni ortogonali, 386, 400, 422
delle potenze, 333, 371, 374, 416-419, 429
delle potenze inverse, 333, 378, 419
dello stepeest descent, 274, 307, 314
di Cholesky, v. Cholesky
di correzione post-iterativa, 295
di Crout, 149, 197, 226
di eliminazione, 157, 160
di Gauss, v. Gauss
di Gauss-Jordan, v. Gauss-Jordan
di Gauss-Seidel, v. Gauss-Seidel
di Givens, v. Givens
di Householder, v. Householder
di Hyman, 352
di Jacobi, v. Jacobi
di Lanczos, v. Lanczos
di Newton, 347
di ortogonalizzazione di Gram-Schmidt, 9
di rilassamento, v. rilassamento
di sottorilassamento, 261
di sovrarilassamento, 261
di Strassen, 202, 228
di Wielandt, 378, 429
iterativo per il calcolo dell'inversa, 294
LR, 354, 428
QR, v. QR
SOR, 261
minima norma, soluzione di, 433, 442, 455, 474, 476, 487
minimax, teorema del, 323, 426, 428, 491
minimi quadrati
con vincoli di uguaglianza, 494-496
condizionamento del problema, 459
metodo del gradiente coniugato, 474, 478
metodo di Householder, 438, 497
polinomio di approssimazione ai, 477
problema lineare dei, 433-443, 454, 496
risoluzione con i valori singolari, 454, 478
minimo, polinomio, 51, 81, 83, 86, 87
Minkowski, 107
disuguaglianza di, 123, 134
mobile, sistema in virgola mobile, 164
modulo p, aritmetica intera, 224
molteplicità
algebrica, 53, 54, 57, 59, 321
geometrica, 53, 54, 57, 59
monico, polinomio, 51
monotona, norma, 131
Moore, 497
Moore-Penrose
equazioni di, 457, 484
pseudoinversa di, 457, 486
Muir, 42
Müntz, 429

negativa
matrice definita, 10
matrice semidefinita, 10

Newton, metodo di, 347
nilpotente, matrice, 24, 95
nilpotenza, grado di, 24
nodo di un grafo, 18
non quadrata, matrice, 433, 443
norma, 108-135
    assoluta, 131, 319
    continuità della, 110
    di Frobenius, 117, 118, 226, 443, 448
    di matrici non quadrate, 443
    di Schur, v. norma di Frobenius
    generalizzata, 132-134
    hölderiana, 123, 134
    indotta, 114, 119, 239
    matriciale, 113
    monotona, 131
    soluzione di minima, 433, 442, 455, 474, 476, 487
    vettoriale, 108
norma 1, 108, 114, 121
norma 2, 108, 114, 121, 375, 443, 448
norma ∞, 108, 114, 121, 373
normale
    matrice, 2, 13, 25, 69, 70, 89, 93, 317-318, 401, 448
    sistema, 433, 478, 496, 497
normali, equazioni, 433
normalizzato, vettore, 8
norme
    equivalenza delle, 110, 111, 121, 131
    proprietà delle, 118-121
nucleo, 13, 30, 53, 432, 434, 448, 476
numeri di Fibonacci, 36
numero
    di condizionamento di un autovalore, 403
    di condizionamento di un autovettore, 403
    di condizionamento di una matrice, 137, 140, 176, 203-208, 226, 410, 458, 466
    di macchina, 164

omogeneo, sistema, 15
operazioni
    di macchina, 165
    fra matrici, 2, 12, 13
    fra vettori, 4
ordinata, matrice consistentemente, 299
ordine di una matrice, 1
ortogonale
    matrice, 2, 66, 93, 153
    proiezione, 7
    sottospazio, 7
ortogonali
    metodo delle iterazioni, 386, 400, 422
    vettori, 6
ortogonalizzazione
    di Gram-Schmidt, metodo di, 9
    variante della, 382, 383
ortonormale, base, 9, 56, 63
ortonormali
    autovettori, 69, 70
    vettori, 8
Ostrowski, 135, 313
Ostrowski-Reich, teorema di, 263
ovale di Cassini, 104, 107
overflow, 165, 178

parallelogramma, legge del, 28
Peano, 43, 135
Penrose, 497
permutazione, matrice di, 3, 17, 39, 145, 161, 174, 190, 214, 217, 441, 451
Perron-Frobenius, teorema di, 303-306, 314
persimmetrica, matrice, 97, 219
perturbazione
    teoremi per gli autovalori, 319
    teoremi per i valori singolari, 462
    teoria della, 136, 137, 428
pivot, 161
    massimo, 172, 174, 177, 215
    massimo per colonne, 190, 440
    variante del, 161, 178
polare, decomposizione, 493
polinomio
    caratteristico, 45, 51, 83, 100, 101
    di approssimazione ai minimi quadrati, 477
    di Chebyshev, 395
    di una matrice, 21, 48, 72, 89
    minimo, 51, 81, 83, 86, 87
    monico, 51
positiva
    matrice definita, 10, 13, 25-28, 32, 41, 73, 74, 82, 93, 94, 117, 127, 148, 198, 211, 212, 250, 263, 302, 309, 408, 426
    matrice semidefinita, 10, 115, 471
post-iterativa, metodo di correzione, 295
potenze, metodo delle, 333, 371, 416-419, 429

condizione di arresto, 374
inverse, 333, 378, 419
precisione di macchina, 164
precondizionamento, tecniche di, 281, 286
precondizionatore, 281
predominanza diagonale
a blocchi, 222
in senso stretto, matrice a, 82, 211, 212, 247, 296
matrice a, 82, 247, 258, 303
principale
di testa, sottomatrice, 3, 74, 144, 158, 326
elemento, 1, 63, 82
sottomatrice, 3, 11, 45, 54
problema
ben condizionato, 136
ben posto, 136
di Dirichlet, 283, 314
generalizzato agli autovalori, 425-427
lineare dei minimi quadrati, 433-443, 454, 459, 474, 496
lineare dei minimi quadrati con vincoli di uguaglianza, 494-496
mal condizionato, 136
mal posto, 136
procedimento di
bisezione, 347
Francis, 365
prodotto
diretto, v. prodotto di Kronecker
esterno, 4
interno, 4
scalare, 5, 28
tensoriale, v. prodotto di Kronecker
prodotto di Kronecker, 40, 135, 300
autovalori del, 102, 270
sottoinsiemi chiusi rispetto al, 41
proiezione, 5
ortogonale, 7
proprietà
degli autovalori, 47-51, 86
degli autovettori, 52-54, 86
delle norme, 118-121
pseudoinversa di Moore-Penrose, 457
sottoinsieme chiuso rispetto all'operazione di, 486

QR, metodo per il calcolo degli autovalori, 332, 353, 367, 428, 471, 497
algoritmo di base, 354
con shift, 363, 428
condizione di arresto, 362
convergenza del terzo ordine, 363, 414
convergenza in ipotesi più deboli, 360
costo computazionale, 358
procedimento di Francis per, 365
tecnica di traslazione, 362
teorema di convergenza, 355
quadrata, matrice, 1
radice di una, 484
quadrati, minimi, v. minimi quadrati
quoziente di Rayleigh, 318, 323-330, 392
generalizzato, 389
metodo delle iterazioni del, 381, 420

radice n-esima dell'unità, 98
radice quadrata di una matrice, 484
raffinamento iterativo, 294
raggio spettrale, 45, 114-116, 232, 234, 236, 239
rango, 13, 15, 29-31, 43, 452, 455, 459
calcolo del, 179, 190, 454, 473
massimo, 15, 435, 443, 476, 486
rappresentazione
base della, 164
cifre della, 164
esponente della, 164
mantissa della, 164
Rayleigh, quoziente di, 318, 323-330, 392
generalizzato, 389
metodo delle iterazioni del, 381, 420
regola
di Binet, 12, 33
di Cramer, 15
di Laplace, 11, 12, 34, 36, 101, 344
di Ruffini-Horner, 344
residuo, 207, 273, 277
riducibile, matrice, 17-20
riduzione
asintotica media, 239
in forma di Hessenberg superiore, 349, 350
media per passo dell'errore, 239
riflessione, matrice di, 151
riga, vettore, 4, 17
righe
linearmente indipendenti di una matrice, 13, 29, 57
scalatura per, 221
rilassamento, metodo per la risoluzione dei sistemi lineari, 261-269, 286, 313

condizione di convergenza, 262, 263
per matrici a blocchi, 269
per matrici definite positive, 263
per matrici tridiagonali, 265
riortogonalizzazione
completa, 396
selettiva, 396
risultante, 42
Ritz
accelerazione di, 389
vettore di, 396
Rohrbach, 107
rotazione di un vettore, 192
Ruffini-Horner, regola di, 344

scalare, matrice, 1
prodotto, 28
scalatura per righe, 221
Schur, 106
complemento di, 31, 201, 211, 219, 306
di $A^H A$, calcolo della forma normale di, 450, 466
forma normale o canonica di, 63, 69, 106, 289, 449, 450, 466
forma normale reale di, 66, 95
norma di, v. norma di Frobenius
teorema di, 130
Seidel, 313
selettiva, riortogonalizzazione, 396
semidefinita negativa, matrice, 10
semidefinita positiva, matrice, 10, 115, 471
separazione degli autovalori, teoremi di, 324-330
Sherman-Morrison, formula di, 33, 214
shift, metodo QR con, 363, 428
convergenza del terzo ordine, 363, 414
simili, matrici, 56, 57
similitudine, trasformazione per, 56
simmetrica, matrice, 2, 69, 391, 409
simultanei, metodo degli spostamenti, 243
singolare, matrice, 12, 31
singolari
valori, v. valori singolari
vettori, 444, 448
sistema in virgola mobile, 164
sistema lineare, 15, 136-315
consistente, 15
omogeneo, 15
metodi di risoluzione, 136-228, 242-315
normale, 433, 478, 496, 497
sovradeterminato, 433
triangolare, 141, 165
soluzione di minima norma, 433, 442, 455, 474, 476, 487
somma diretta di sottospazi, 7
SOR, metodo, 261
sostituzione
all'indietro, 142
in avanti, 142
sottoinsiemi chiusi rispetto
al prodotto di Kronecker, 41
alla moltiplicazione, 3
all'operazione di inversione, 13
all'operazione di pseudo-inversa, 486
sottomatrice, 3
principale, 3, 11, 45, 54
principale di testa, 3, 74, 144, 158, 326
sottorilassamento, metodo di, 261
sottospazi
distanza fra, 387, 422
metodo delle iterazioni di, 386, 400, 422
somma diretta di due, 7
sottospazio, 6, 7, 13, 14, 322-324, 452
base di un, 6
dimensione di un, 6, 90
invariante, 393
invariante dominante, 387
ortogonale, 7
sovradeterminato, sistema, 433
sovrarilassamento, metodo di, 261
spettrale, raggio, 45, 114-116, 232, 234, 236, 239
spettro di una matrice, 45
spostamenti
simultanei, metodo degli, 243
successivi, metodo degli, 243
stabilità, 141
steepest descent, metodo dello, 274, 307, 314
Stein, teorema di, 302
Stein-Rosenberg, teorema di, 253, 304
stima a-posteriori dell'errore, 207
Strassen, metodo di, 202, 228
strategia ciclica o classica per il metodo di Jacobi, 369
Sturm, 105
successione di, 345, 346, 428
successione
convergente, 231
di matrici, 231-234, 289, 290, 412
di Sturm, 345, 346, 428

di vettori, 231
successivi, metodo degli spostamenti, 243
Sylvester, 43, 106, 498

tasso asintotico di convergenza, 241
Taussky, 107, 498
tecnica di traslazione per il metodo QR, 362
tecniche
compatte, 149, 196
di precondizionamento, 281, 286
teorema
del minimax, 323, 426, 428, 491
di Bauer-Fike, 319
di Cayley-Hamilton, 50, 106
di Courant-Fisher, 323
di Hirsch, 316
di Kahan, 262
di Ostrowski-Reich, 263
di Perron-Frobenius, 303-306, 314
di Schur, 130
di Stein, 302
di Stein-Rosenberg, 253, 304
di Weinstein, 318
teoremi
di Gerschgorin, 76-80
di localizzazione degli autovalori, 76-80, 133, 316-319, 401
di perturbazione per gli autovalori, 319
di perturbazione per i valori singolari, 462
di separazione per gli autovalori, 324-330
teoria della perturbazione, 136, 137, 428
testa, sottomatrice principale di, 3, 74, 144, 158, 326
Toeplitz, matrice di, 97
traccia, 26, 45, 57, 237, 408
trasformata discreta di Fourier, 99
trasformazione per similitudine, 56
traslazione per il metodo QR, tecnica di, 362
trasposta coniugata, matrice, 1, 47
trasposta, matrice, 1, 47
triangolare
disuguaglianza, 108
in senso stretto, matrice, 1, 24
matrice, 1, 13, 25, 47, 63, 141, 143, 156, 203, 208
sistema lineare, 141-143, 165-167
tridiagonale, matrice 1, 27, 35, 37, 100, 158, 212, 213, 254, 265, 270, 299, 343, 394, 400, 411
a blocchi, 221, 259, 269
autovalori di una matrice, 100, 270
fattorizzazione LU, 210
tridiagonalizzazione
metodo di Givens, 336
metodo di Householder, 334-336, 411
metodo di Lanczos, 339
troncamento, errore di, 136
Turing, 226

underflow, 165, 178
unione disgiunta di cerchi di Gerschgorin, 79
unità, radice n-esima della, 98
unitaria, matrice, 2, 13, 25, 27, 41, 48, 56, 63, 68, 70, 87, 93, 98, 112, 117, 128, 151

valori singolari, 444
calcolo dei, 448
condizionamento del problema dei, 459
decomposizione ai, 444, 447, 497
di una matrice normale, 448
metodo di Lanczos per, 479
risoluzione del problema dei minimi quadrati con, 454, 478
teoremi di perturbazione dei, 462
Vandermonde, 42
matrice di, 39, 223, 461
variante
del pivot, 161, 178
della deflazione, 385
dell'ortogonalizzazione, 382, 383
di Maehly, 345
di Wielandt, 378
vettore
colonna, 4, 17
componenti di un, 4
di Lanczos, 391, 393
di Ritz, 396
lunghezza euclidea di un, 5
riga, 4, 17
rotazione, 192
vettori
A-coniugati, 275, 277, 308-312
angolo di due, 5, 410
centroantisimmetrici, 96
centrosimmetrici, 96
linearmente dipendenti, 6
linearmente indipendenti, 6, 28

normalizzati, 8
operazioni fra, 4
ortogonali, 6, 8
ortonormali, 8
singolari, 444, 448
successioni di, 231
vettoriale, norma, 108
vincoli di uguaglianza, metodo dei minimi quadrati con, 494-496
virgola mobile, sistema in, 164
Von Neumann, 225

Weierstrass, 106
Weinstein, teorema di, 318
Weyl, 498
Wielandt, metodo di, 378, 429
Wilkinson, 226, 227, 428
Woodbury, formula di, 33

Gli autori

Dario Bini, docente di Analisi Numerica presso il Corso di Laurea in Matematica della II Università di Roma, Milvio Capovani e Ornella Menchi, docenti di Calcolo Numerico presso il Corso di Laurea in Scienze dell'Informazione dell'Università di Pisa, hanno svolto e stanno svolgendo ricerche in vari settori della Matematica Computazionale e, in particolare, nel settore dell'analisi e della sintesi di algoritmi numerici sequenziali e paralleli per l'algebra lineare. Sono autori, con Roberto Bevilacqua, di "Introduzione alla matematica computazionale" (Zanichelli, 1987). Dario Bini e Milvio Capovani sono autori, con Grazia Lotti e Francesco Romani, di "Complessità numerica"(Boringhieri, 1981).

L'opera

L'algebra lineare è una parte essenziale del bagaglio culturale di base richiesto in molti campi della matematica e più in generale della scienza. La risoluzione di gran parte dei problemi scientifici comporta la risoluzione di problemi di algebra lineare. Questo fatto ha portato ad un forte interesse per lo sviluppo e l'analisi di *metodi numerici* computazionalmente efficienti per risolvere tali problemi. L'introduzione, e la sempre più vasta diffusione, dei calcolatori nel campo scientifico ha ulteriormente sviluppato questo processo nell'ambito del settore di ricerca noto come *numerical linear algebra*.

In questo testo sono esposti e analizzati i principali metodi per la risoluzione dei problemi fondamentali dell'algebra lineare. Particolare attenzione è rivolta agli aspetti numerici e computazionali. Il materiale presentato è corredato da esempi, direttamente sperimentati al calcolatore, con la presentazione di tabelle e grafici, e da numerosi esercizi. È riportata anche una bibliografia completa e aggiornata e ogni capitolo è chiuso da note bibliografiche e da cenni storici, che possono guidare i lettori interessati in indagini più approfondite.

Il testo è rivolto principalmente agli studenti dei corsi di laurea in Matematica, Fisica, Ingegneria e Scienze dell'Informazione, ed ai ricercatori che operano nel settore del calcolo scientifico.