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

$$\Pi' A = \begin{bmatrix} 1 & \mathbf{0}^H \\ \dfrac{1}{\alpha}\,\mathbf{d} & I \end{bmatrix} \begin{bmatrix} \alpha & \mathbf{c}^H \\ \mathbf{0} & B_{k-1} \end{bmatrix}$$

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

$$\Pi' A = \begin{bmatrix} 1 & \mathbf{0}^H \\ \frac{1}{\alpha}\,\mathbf{d} & I \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0}^H \\ \mathbf{0} & \Pi_{k-1}^T L_{k-1} \end{bmatrix} \begin{bmatrix} \alpha & \mathbf{c}^H \\ \mathbf{0} & U_{k-1} \end{bmatrix}$$

$$= \begin{bmatrix} 1 & \mathbf{0}^H \\ \frac{1}{\alpha}\mathbf{d} & \Pi_{k-1}^T L_{k-1} \end{bmatrix} \begin{bmatrix} \alpha & \mathbf{c}^H \\ \mathbf{0} & U_{k-1} \end{bmatrix}$$

$$= \begin{bmatrix} 1 & \mathbf{0}^H \\ \mathbf{0} & \Pi_{k-1}^T \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0}^H \\ \frac{1}{\alpha}\Pi_{k-1}\mathbf{d} & L_{k-1} \end{bmatrix} \begin{bmatrix} \alpha & \mathbf{c}^H \\ \mathbf{0} & U_{k-1} \end{bmatrix},$$

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.

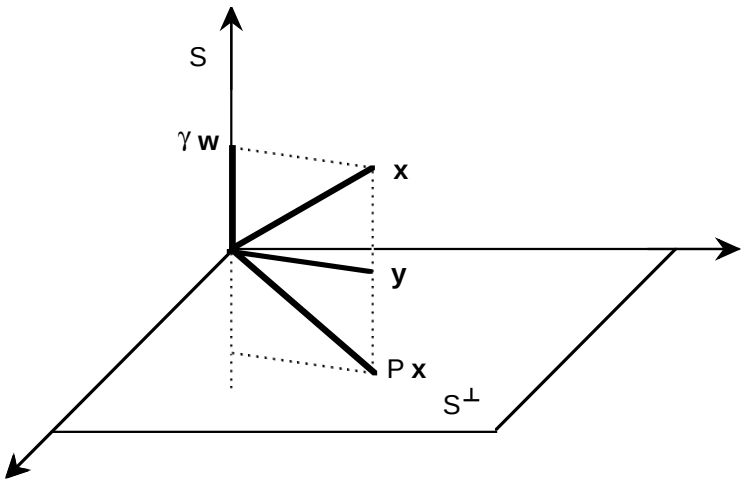

Fig. 4.1 - Riflessione di un vettore $\mathbf{x}$.

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

$$A^{(2)} = \begin{bmatrix} \alpha & \mathbf{c}^H \\ \mathbf{0} & B^{(2)} \end{bmatrix} \begin{matrix} \} \\ \\ \} \end{matrix} \begin{matrix} 1 \text{ riga} \\ \\ n-1 \text{ righe} \end{matrix}$$

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

$$\alpha_1 = a_1$$

$$\left.\begin{array}{l} \beta_i = b_i/\alpha_i \\ \alpha_{i+1} = a_{i+1} - \beta_i c_i \end{array}\right\} \quad \text{per } i = 1, \ldots, n-1,$$

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, $\ldots$, $(n-1)$-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.

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.

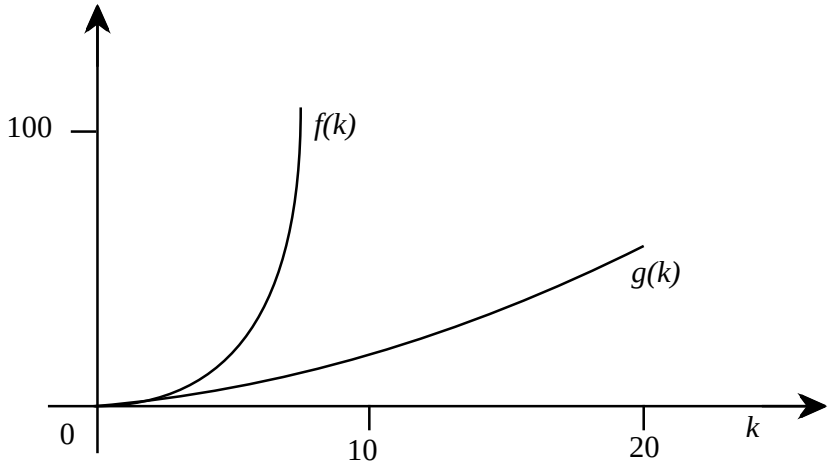

Fig. 4.2 - Grafici delle funzioni delle maggiorazioni (39) e (40).

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).

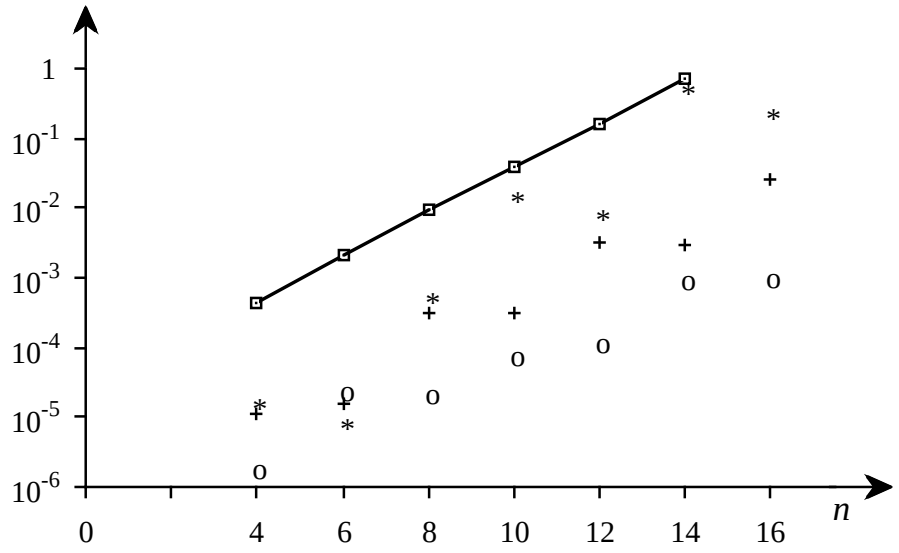

Fig. 4.3 - Errori relativi del metodo di Gauss e delle sue varianti.

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

$$Q = P^{(1)} P^{(2)} = \frac{1}{6} \begin{bmatrix} \text{-}2 & 4 & 4 \\ 4 & \text{-}2 & 4 \\ 4 & 4 & \text{-}2 \end{bmatrix}.$$

∎

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

$$\mathbf{w} = \widehat{Q}(\mathbf{z} + \mathbf{e}),$$

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.

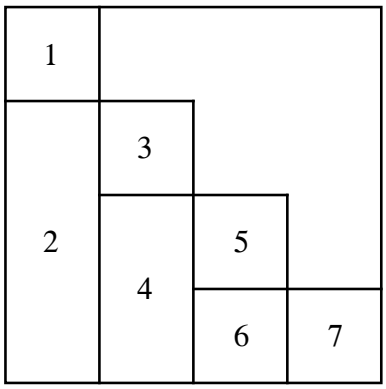

Fig. 4.7 - Ordinamento nel metodo di Cholesky.

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

$$\begin{aligned} &s_1 = (a_{11} + a_{22})(b_{11} + b_{22}) && s_2 = (a_{21} + a_{22})b_{11} \\ &s_3 = a_{11}(b_{12} - b_{22}) && s_4 = a_{22}(b_{21} - b_{11}) \\ &s_5 = (a_{11} + a_{12})b_{22} && s_6 = (a_{21} - a_{11})(b_{11} + b_{12}) \\ &s_7 = (a_{12} - a_{22})(b_{21} + b_{22}) \\ &c_{11} = s_1 + s_4 - s_5 + s_7 && c_{12} = s_3 + s_5 \\ &c_{21} = s_2 + s_4 && c_{22} = s_1 - s_2 + s_3 + s_6. \end{aligned}$$

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.)