

Capitolo 1

Sistemi lineari

1.1 Introduzione al Calcolo Numerico

Il Calcolo Numerico è una disciplina che fa parte di un ampio settore della Matematica Applicata che prende il nome di Analisi Numerica. Si tratta di una materia che è al confine tra la Matematica e l'Informatica poiché cerca di risolvere i consueti problemi matematici utilizzando però una via algoritmica. In pratica i problemi vengono risolti indicando un processo che, in un numero finito di passi, fornisca una soluzione numerica e soprattutto che sia implementabile su un elaboratore. I problemi matematici che saranno affrontati nelle pagine seguenti sono problemi di base: risoluzione di sistemi lineari, approssimazione delle radici di funzioni non lineari, approssimazione di funzioni e dati sperimentali, calcolo di integrali definiti. Tali algoritmi di base molto spesso non sono altro se non un piccolo ingranaggio nella risoluzione di problemi ben più complessi.

1.1.1 Elementi di Algebra Lineare

Sia \mathbb{R} l'insieme dei numeri reali. Generalmente si indica con $\mathbb{R}^{m \times n}$ l'insieme delle *matrici* ad elementi reali aventi m righe ed n colonne. Quindi una matrice è una *tabella a doppia entrata* di numeri reali. Per esempio:

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}$$

con $a_{ij} \in \mathbb{R}$ ed $m, n \in \mathbb{N}$. I numeri interi m ed n si dicono *dimensioni della matrice*, ovvero A si dice matrice di *dimensioni* $m \times n$ o di *ordine* $m \times n$. Se $m = n$ allora la matrice A si dice *quadrata* di dimensione n o di ordine n altrimenti si dice *rettangolare*. Se i e j sono numeri interi con $1 \leq i \leq m$ e $1 \leq j \leq n$ allora l'elemento della matrice A di dimensione $m \times n$ che si trova in posizione (i, j) viene indicato con a_{ij} . Gli elementi a_{ij} di una matrice quadrata A di ordine n tali che $i = j$ sono detti *elementi principali o diagonali* e formano la cosiddetta diagonale principale di A .

Assegnata una matrice $A \in \mathbb{R}^{m \times n}$ si definisce *matrice trasposta di A* la matrice $B = A^T \in \mathbb{R}^{n \times m}$ tale che

$$b_{ij} = a_{ji}, \quad i = 1, \dots, n, \quad j = 1, \dots, m.$$

Se accade che $A = A^T$ allora la matrice è detta *simmetrica*. Gli elementi di una matrice che si trovano al di sopra della diagonale principale sono detti *sopradiagonali*, mentre quelli che si trovano al di sotto della stessa diagonale principale sono detti *sottodiagonali*. Se una matrice ha tutti gli elementi sopradiagonali e sottodiagonali uguali a zero viene detta *matrice diagonale*. Se invece ha solo gli elementi sopradiagonali nulli allora viene detta *triangolare inferiore*. Se ha gli elementi sottodiagonali nulli allora è detta *triangolare superiore*.

Assegnate due matrici $A, B \in \mathbb{R}^{m \times n}$ si definisce *somma* di A e B , e si denota con $C = A + B$, la matrice $C \in \mathbb{R}^{m \times n}$ i cui elementi sono:

$$c_{ij} = a_{ij} + b_{ij} \quad i = 1, \dots, m, \quad j = 1, \dots, n.$$

In modo analogo si definisce la *differenza* tra matrici, infatti $D = A - B$ è la matrice avente elementi:

$$d_{ij} = a_{ij} - b_{ij} \quad i = 1, \dots, m, \quad j = 1, \dots, n.$$

Se $\alpha \in \mathbb{R}$ ed $A \in \mathbb{R}^{m \times n}$ allora la matrice $C = \alpha A$ è definita da:

$$c_{ij} = \alpha a_{ij}.$$

Se $A \in \mathbb{R}^{m \times p}$ e $B \in \mathbb{R}^{p \times n}$ si definisce *prodotto* di A per B la matrice $C \in \mathbb{R}^{m \times n}$ i cui elementi sono

$$c_{ij} = \sum_{k=1}^p a_{ik} b_{kj} \quad i = 1, \dots, m, \quad j = 1, \dots, n.$$

Si noti che affinché tale prodotto abbia senso è necessario che il numero delle colonne di A coincida con il numero delle righe di B . Quando ciò accade le matrici si dicono *conformabili*, altrimenti si dicono *non conformabili*. Ad esempio nel nostro caso se $m \neq n$ allora il prodotto BA non ha senso. Ha sempre significato considerare i prodotti AB e BA se A e B sono matrici quadrate dello stesso ordine ($m = n$).

È facile verificare che il prodotto tra matrici gode della proprietà *associativa* ma in generale non di quella *commutativa*. Vale invece la seguente proprietà:

$$(AB)^T = B^T A^T.$$

Esempio 1.1.1 Siano A e B le seguenti matrici:

$$A = \begin{pmatrix} 3 & 1 & 0 \\ -1 & 2 & 1 \\ 3 & 1 & 1 \end{pmatrix}; \quad B = \begin{pmatrix} 2 & 1 & -1 \\ 0 & 1 & 1 \\ 2 & 1 & 1 \end{pmatrix}.$$

Calcoliamo la matrice $C = AB$. L'elemento c_{ij} è uguale alla somma dei prodotti degli elementi della i -esima riga di A per la j -esima colonna di B .

$$\begin{aligned} c_{11} &= 3 \cdot 2 + 1 \cdot 0 + 0 \cdot 2 = 6 \\ c_{12} &= 3 \cdot 1 + 1 \cdot 1 + 0 \cdot 1 = 4 \\ c_{13} &= 3 \cdot (-1) + 1 \cdot 1 + 0 \cdot 1 = -2 \\ c_{21} &= -1 \cdot 2 + 2 \cdot 0 + 1 \cdot 2 = 0 \\ c_{22} &= -1 \cdot 1 + 2 \cdot 1 + 1 \cdot 1 = 2 \\ c_{23} &= -1 \cdot (-1) + 2 \cdot 1 + 1 \cdot 1 = 4 \\ c_{31} &= 3 \cdot 2 + 1 \cdot 0 + 1 \cdot 2 = 8 \\ c_{32} &= 3 \cdot 1 + 1 \cdot 1 + 1 \cdot 1 = 5 \\ c_{33} &= 3 \cdot (-1) + 1 \cdot 1 + 1 \cdot 1 = -1. \end{aligned}$$

In definitiva

$$C = \begin{pmatrix} 6 & 4 & -2 \\ 0 & 2 & 4 \\ 8 & 5 & -1 \end{pmatrix}.$$

Calcolando il prodotto $D = BA$ si trova invece:

$$D = \begin{pmatrix} 2 & 3 & 0 \\ 2 & 3 & 2 \\ 8 & 5 & 2 \end{pmatrix}$$

da cui risulta evidente che $AB \neq BA$.

Siano $A, B \in \mathbb{R}^{n \times n}$ le seguenti matrici

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

e

$$B = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}$$

dove $A_{11}, B_{11} \in \mathbb{R}^{p \times p}$, $A_{12}, B_{12} \in \mathbb{R}^{p \times (n-p)}$, $A_{21}, B_{21} \in \mathbb{R}^{(n-p) \times p}$ e infine $A_{22}, B_{22} \in \mathbb{R}^{(n-p) \times (n-p)}$, con $p < n$, rappresentano a loro volta matrici e non semplici elementi. Si dice cioè che A e B sono state suddivise a blocchi. Il prodotto AB può essere calcolato utilizzando tale decomposizione delle matrici:

$$AB = \begin{pmatrix} A_{11}B_{11} + A_{12}B_{21} & A_{11}B_{12} + A_{12}B_{22} \\ A_{21}B_{11} + A_{22}B_{21} & A_{21}B_{12} + A_{22}B_{22} \end{pmatrix}.$$

Si definisce *matrice identità di ordine n* la matrice quadrata diagonale I_n avente tutti gli elementi principali uguali a 1:

$$I_n = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \dots & \dots & 0 & 1 \end{pmatrix}.$$

La matrice identità è l'elemento neutro per il prodotto, cioè se $A \in \mathbb{R}^{n \times n}$ si ha

$$AI_n = I_nA = A.$$

Definizione 1.1.1 Una matrice che si ottiene da I_n scambiando alcune righe (o colonne) viene detta matrice di permutazione.

Esempio 1.1.2 Sia P la seguente matrice di permutazione:

$$P = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

che è stata ottenuta da I_3 scambiando la prima riga con la terza. Consideriamo la seguente matrice A

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

e calcoliamo il prodotto PA :

$$PA = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} = \begin{pmatrix} 7 & 8 & 9 \\ 4 & 5 & 6 \\ 1 & 2 & 3 \end{pmatrix}.$$

La moltiplicazione a sinistra di una matrice di permutazione per A ha l'effetto di scambiare le righe di A esattamente nello stesso modo con cui erano state scambiate le righe dell'identità per ottenere P . Calcoliamo ora il prodotto AP :

$$AP = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 3 & 2 & 1 \\ 6 & 5 & 4 \\ 9 & 8 & 7 \end{pmatrix}.$$

Invece la moltiplicazione a destra di una matrice di permutazione per A ha l'effetto di scambiare le colonne di A .

Data una matrice $A \in \mathbb{R}^{m \times n}$, una matrice $B \in \mathbb{R}^{h \times k}$, $0 < h \leq m$, $0 < k \leq n$, è detta *sottomatrice* di A se è ottenuta da A eliminando $m - h$ righe ed $n - k$ colonne. Data una matrice $A \in \mathbb{R}^{m \times n}$, una sottomatrice quadrata B di ordine $k \leq n$ di A è detta *principale* se gli elementi principali di B sono anche gli elementi principali di A . Una sottomatrice B principale di ordine k di A è detta *principale di testa* se è formata dagli elementi a_{ij} , $i, j = 1, \dots, k$.

Definizione 1.1.2 Se $A \in \mathbb{R}^{n \times n}$ è una matrice di ordine 1, si definisce determinante di A il numero

$$\det A = a_{11}.$$

Se la matrice A è quadrata di ordine n allora fissata una qualsiasi riga (colonna) di A , diciamo la i -esima (j -esima) allora applicando la cosiddetta regola di Laplace il determinante di A è:

$$\det A = \sum_{j=1}^n a_{ij} (-1)^{i+j} \det A_{ij}$$

dove A_{ij} è la matrice che si ottiene da A cancellando la i -esima riga e la j -esima colonna.

Il determinante è pure uguale a

$$\det A = \sum_{i=1}^n a_{ij} (-1)^{i+j} \det A_{ij},$$

cioè il determinante è indipendente dall'indice di riga (o di colonna) fissato. Se A è la matrice di ordine 2

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}.$$

allora

$$\det A = a_{11}a_{22} - a_{21}a_{12}.$$

Il determinante ha le seguenti proprietà:

1. Se A è una matrice triangolare o diagonale allora

$$\det A = \prod_{i=1}^n a_{ii};$$

2. $\det I = 1$;
3. $\det A^T = \det A$;
4. $\det AB = \det A \det B$ (Regola di Binet);
5. se $\alpha \in \mathbb{R}$ allora $\det \alpha A = \alpha^n \det A$.
6. $\det A = 0$ se una riga (o una colonna) è nulla, oppure una riga (o una colonna) è proporzionale ad un'altra riga (o colonna) oppure è combinazione lineare di due (o più) righe (o colonne) di A .
7. Se A è una matrice triangolare a blocchi

$$A = \begin{pmatrix} B & C \\ O & D \end{pmatrix}$$

con B e D matrici quadrate, allora

$$\det A = \det B \det D. \tag{1.1}$$

Una matrice A di ordine n si dice *non singolare* se il suo determinante è diverso da zero, in caso contrario viene detta *singolare*. Si definisce *inversa di A* la matrice A^{-1} tale che:

$$AA^{-1} = A^{-1}A = I_n$$

Per quello che riguarda il determinante della matrice inversa vale la seguente proprietà:

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

Ricordiamo che se A e B sono due matrici non singolari in base alla regola di Binet anche il loro prodotto è una matrice invertibile e inoltre:

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

Questa proprietà vale anche se abbiamo più di due matrici, cioè:

$$(ABC)^{-1} = C^{-1}B^{-1}A^{-1}.$$

Vettori

Se $A \in \mathbb{R}^{m \times 1}$ (o $A \in \mathbb{R}^{1 \times m}$), la matrice si riduce ad una sola colonna (o una sola riga) e viene detta *vettore colonna* (o *riga*) *ad m elementi o componenti*. Solitamente il termine vettore viene associato a vettori colonna e l'insieme dei vettori ad m componenti viene indicato con \mathbb{R}^m . Per le operazioni tra vettori valgono le stesse regole viste per le matrici, cioè la somma e la differenza sono possibili tra vettori dello stesso tipo e con lo stesso numero di componenti. Se \mathbf{x} è un vettore colonna di m elementi allora \mathbf{x}^T è un vettore riga sempre di m elementi. Se $A \in \mathbb{R}^{m \times n}$ e $\mathbf{x} \in \mathbb{R}^n$ è possibile definire il prodotto matrice per vettore nel seguente modo:

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

quindi $\mathbf{y} \in \mathbb{R}^m$. Non è possibile effettuare il prodotto $A\mathbf{x}^T$ perchè le dimensioni non sono compatibili.

Esempio 1.1.3 *Sia*

$$A = \begin{pmatrix} 5 & 1 & 0 \\ -1 & 1 & 2 \\ 5 & -5 & 1 \end{pmatrix}$$

e sia \mathbf{x} il vettore

$$\mathbf{x} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}.$$

Calcoliamo il vettore prodotto $\mathbf{y} = A\mathbf{x}$:

$$\begin{aligned} y_1 &= 5 \cdot 1 + 1 \cdot 2 + 0 \cdot 3 = 7 \\ y_2 &= -1 \cdot 1 + 1 \cdot 2 + 2 \cdot 3 = 7 \\ y_3 &= 5 \cdot 1 - 5 \cdot 2 + 1 \cdot 3 = -2. \end{aligned}$$

Tra vettori sono consentite le seguenti operazioni:

1. *prodotto interno*;
2. *prodotto esterno*.

Il prodotto interno (o scalare), che viene spesso indicato come (\cdot, \cdot) , è definito nel seguente modo: siano $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, allora

$$(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T \mathbf{y} = \sum_{i=1}^n x_i y_i = \alpha$$

e il risultato è un numero reale. Il prodotto scalare soddisfa le seguenti proprietà:

1. $\mathbf{x}^T \mathbf{x} \geq 0$ per ogni $\mathbf{x} \in \mathbb{R}^n$ e $(\mathbf{x}, \mathbf{x}) = 0$ se e solo se $\mathbf{x} = \mathbf{0}$;
2. $\mathbf{x}^T \mathbf{y} = \mathbf{y}^T \mathbf{x}$ per ogni $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$;
3. $(\alpha \mathbf{x})^T \mathbf{y} = \alpha (\mathbf{x}^T \mathbf{y})$ per ogni $\alpha \in \mathbb{R}$ e per ogni $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$;
4. $(\mathbf{x} + \mathbf{y})^T \mathbf{z} = \mathbf{x}^T \mathbf{z} + \mathbf{y}^T \mathbf{z}$ per ogni $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^n$.
5. se $\mathbf{x}^T \mathbf{y} = 0$ allora i due vettori si dicono *ortogonali*.

Se $\mathbf{x} \in \mathbb{R}^n$ e $\mathbf{y} \in \mathbb{R}^m$ allora il prodotto esterno viene definito nel seguente modo:

$$A = \mathbf{x}\mathbf{y}^T$$

e il risultato è una matrice di dimensione $n \times m$ i cui elementi sono:

$$a_{ij} = x_i y_j, \quad i = 1, \dots, n, \quad j = 1, \dots, m.$$

Esempio 1.1.4 Siano \mathbf{x} e \mathbf{y} i seguenti vettori:

$$\mathbf{x} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$$

e

$$\mathbf{y} = \begin{pmatrix} -1 \\ -2 \\ 4 \end{pmatrix}.$$

Calcoliamo prima il prodotto interno:

$$\mathbf{x}^T \mathbf{y} = 1 \cdot (-1) + 2 \cdot (-2) + 3 \cdot 4 = 7.$$

Osserviamo che tale operazione gode della proprietà commutativa, poichè $\mathbf{y}^T \mathbf{x} = 7$.

Per quello che riguarda il prodotto esterno, il risultato è la matrice

$$A = \mathbf{x}\mathbf{y}^T = \begin{pmatrix} -1 & -2 & 4 \\ -2 & -4 & 8 \\ -3 & -6 & 12 \end{pmatrix}.$$

Tale prodotto non gode della proprietà commutativa, infatti:

$$B = \mathbf{y}\mathbf{x}^T = \begin{pmatrix} -1 & -2 & -3 \\ -2 & -4 & -6 \\ 4 & 8 & 12 \end{pmatrix}.$$

Infatti $B \neq A$, anche se va osservato che $B = A^T$.

Norme Vettoriali

La funzione $\|\cdot\| : \mathbb{R}^n \rightarrow \mathbb{R}$ si dice *norma* se per ogni vettore $\mathbf{x} \in \mathbb{R}^n$ $\|\mathbf{x}\|$ soddisfa:

1. $\|\mathbf{x}\| \geq 0$ e $\|\mathbf{x}\| = 0$ se e solo se $\mathbf{x} = 0$;
2. $\|\alpha\mathbf{x}\| = |\alpha|\|\mathbf{x}\|$ per ogni $\alpha \in \mathbb{C}$;
3. $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ per ogni $\mathbf{x}, \mathbf{y} \in \mathbb{C}^n$ (*disuguaglianza triangolare*).

Si può dimostrare che per ogni fissato p , $1 \leq p \leq \infty$, la funzione $\|\cdot\|_p$ che a $\mathbf{x} \in \mathbb{C}^n$ associa

$$\|\mathbf{x}\|_p = \left(\sum_{j=1}^n |x_j|^p \right)^{\frac{1}{p}}$$

è una norma su vettori. Queste norme prendono il nome di *norme Hölderiane*. Tra queste le più utilizzate sono:

$$\|\mathbf{x}\|_1 = \sum_{j=1}^n |x_j| \quad \text{norma 1}$$

$$\|\mathbf{x}\|_2 = \sqrt{\sum_{j=1}^n |x_j|^2} \quad \text{norma 2 o norma euclidea}$$

$$\|\mathbf{x}\|_\infty = \max_{1 \leq j \leq n} |x_j| \quad \text{norma infinito.}$$

Norme su Matrici

Definizione 1.1.3 Una funzione $\|\cdot\| : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}$ tale che per ogni matrice $A \in \mathbb{R}^{n \times n}$, $\|A\|$ soddisfa:

1. $\|A\| \geq 0$ e $\|A\| = 0$ se e solo se $A = 0$;
2. $\|\alpha A\| = |\alpha| \|A\|$ per ogni $\alpha \in \mathbb{C}$;
3. $\|A + B\| \leq \|A\| + \|B\|$ per ogni $A, B \in \mathbb{C}^{n \times n}$;
4. $\|A \cdot B\| \leq \|A\| \cdot \|B\|$ per ogni $A, B \in \mathbb{C}^{n \times n}$;

si dice norma di matrice.

Definizione 1.1.4 Si dice che una norma di matrice è compatibile con una norma di vettore se per ogni matrice A e per ogni vettore \mathbf{x} risulta

$$\|A\mathbf{x}\| \leq \|A\| \|\mathbf{x}\|.$$

Un modo per definire le norme di matrici compatibili con norme di vettori è il seguente. Sia $\mathbf{x} \neq 0$ con norma $\|\mathbf{x}\|$. Considerata la norma del vettore $A\mathbf{x}$, $\|A\mathbf{x}\|$, definiamo come norma di A il numero $\|A\|$ dato da

$$\|A\| = \sup_{\mathbf{x} \neq 0} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|}. \quad (1.2)$$

$\|A\|$ è detta *norma naturale di A* oppure *norma di A indotta* dalla norma di vettore $\|\mathbf{x}\|$.

La (1.2) può anche scriversi:

$$\|A\| = \sup_{\|\mathbf{y}\|=1} \|A\mathbf{y}\|$$

anzi è possibile dimostrare che

$$\|A\| = \max_{\|\mathbf{y}\|=1} \|A\mathbf{y}\|.$$

Le norme matriciali indotte dalle norme su vettori $\|\mathbf{x}\|_1$ e $\|\mathbf{x}\|_\infty$ sono

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

$$\|A\|_\infty = \max_{1 \leq i \leq n} \sum_{j=1}^n |a_{ij}|.$$

1.2 Metodi diretti per sistemi lineari

Ci poniamo il seguente problema:

Siano assegnati $n^2 + n$ numeri reali, che indichiamo rispettivamente con a_{ij} e b_i , $i, j = 1, \dots, n$. Determinare, se esistono, n numeri reali x_i , con $i = 1, \dots, n$, tali che:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ \vdots & \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= b_n. \end{aligned} \quad (1.3)$$

Le (1.3) definiscono un *sistema di n equazioni algebriche lineari* nelle n incognite x_1, x_2, \dots, x_n . Se definiamo la matrice $A = [a_{ij}]$ con $i, j = 1, \dots, n$, ed

i vettori $\mathbf{x}^T = (x_1, x_2, \dots, x_n)$ e $\mathbf{b}^T = (b_1, b_2, \dots, b_n)$, cosicchè le equazioni (1.3) assumono la forma:

$$A\mathbf{x} = \mathbf{b}. \quad (1.4)$$

Il vettore \mathbf{x} viene detto *vettore soluzione*, A è detta *matrice dei coefficienti* e \mathbf{b} *vettore dei termini noti*. Nel seguito assumeremo che il determinante di A sia diverso da 0. Un metodo universalmente noto per risolvere il problema (1.4) è l'applicazione della cosiddetta *Regola di Cramer* la quale fornisce:

$$x_i = \frac{\det A_i}{\det A} \quad i = 1, \dots, n, \quad (1.5)$$

dove A_i è la matrice ottenuta da A sostituendo la sua i -esima colonna con il termine noto \mathbf{b} . Dalla (1.5) è evidente che per ottenere una componente della soluzione è necessario il calcolo di $n+1$ determinanti di ordine n . Si può facilmente dedurre che il numero di operazioni necessarie per il calcolo del determinate di una matrice di ordine n è circa $n!$, quindi questa strada non permette di poter determinare velocemente la soluzione del nostro sistema.

1.2.1 Sistemi triangolari

Prima di affrontare la soluzione algoritmica di un sistema lineare vediamo qualche particolare sistema che può essere agevolmente risolto. Assumiamo che il sistema da risolvere abbia la seguente forma:

$$\begin{array}{ccccccc} a_{11}x_1 & +a_{12}x_2 & \dots & +a_{1i}x_i & \dots & +a_{1n}x_n & = b_1 \\ & a_{22}x_2 & \dots & +a_{2i}x_i & \dots & +a_{2n}x_n & = b_2 \\ & & \ddots & \vdots & & \vdots & \vdots \\ & & & a_{ii}x_i & \dots & +a_{in}x_n & = b_i \\ & & & & \ddots & \vdots & \vdots \\ & & & & & a_{nn}x_n & = b_n \end{array} \quad (1.6)$$

con $a_{ii} \neq 0$ per ogni i . In questo caso la matrice A è detta *triangolare superiore*. È evidente che in questo caso, la soluzione è immediatamente

calcolabile. Infatti:

$$\begin{cases} x_n = \frac{b_n}{a_{nn}} \\ x_i = \frac{b_i - \sum_{j=i+1}^n a_{ij}x_j}{a_{ii}} \quad i = n-1, \dots, 1. \end{cases} \quad (1.7)$$

Il metodo (1.7) prende il nome di *metodo di sostituzione all'indietro*, poichè il vettore \mathbf{x} viene calcolato partendo dall'ultima componente. Anche per il seguente sistema il vettore soluzione è calcolabile in modo analogo.

$$\begin{array}{rcccccc} a_{11}x_1 & & & & & = b_1 \\ a_{21}x_1 & +a_{22}x_2 & & & & = b_2 \\ \vdots & \vdots & \ddots & & & \vdots \\ a_{i1}x_1 & +a_{i2}x_2 & \dots & +a_{ii}x_i & & = b_i \\ \vdots & \vdots & & & \ddots & \vdots \\ a_{n1}x_1 & +a_{n2}x_2 & \dots & +a_{ni}x_i & \dots & +a_{nn}x_n = b_n \end{array} \quad (1.8)$$

In questo caso la matrice dei coefficienti è *triangolare inferiore* e la soluzione viene calcolata con il *metodo di sostituzione in avanti*:

$$\begin{cases} x_1 = \frac{b_1}{a_{11}} \\ x_i = \frac{b_i - \sum_{j=1}^{i-1} a_{ij}x_j}{a_{ii}} \quad i = 2, \dots, n-1. \end{cases}$$

1.2.2 Metodo di Eliminazione di Gauss

L'idea di base del metodo di Gauss è appunto quella di operare delle opportune trasformazioni sul sistema originale $A\mathbf{x} = \mathbf{b}$, che non costino eccessivamente, in modo da ottenere un sistema equivalente, cioè un sistema che ammetta la stessa soluzione di quello di partenza, ma che sia facile da risolvere, per esempio uno avente come matrice dei coefficienti una matrice

triangolare superiore. Prima di descrivere il metodo vediamo un esempio. Supponiamo che il sistema da risolvere sia:

$$\begin{aligned} 2x_1 + x_2 + x_3 &= -1 \\ 6x_1 + 2x_2 + x_3 &= 1 \\ 4x_1 - 2x_2 + x_3 &= 2 \end{aligned}$$

La soluzione di un sistema lineare non cambia se un'equazione viene sostituita dalla combinazione lineare di due (o più) equazioni dello stesso sistema. L'idea alla base del metodo di Gauss è quella di ottenere un sistema lineare con matrice dei coefficienti triangolare superiore effettuando opportune combinazioni lineari tra le equazioni. Poniamo

$$A^{(1)} = \begin{pmatrix} 2 & 1 & 1 \\ 6 & 2 & 1 \\ 4 & -2 & 1 \end{pmatrix}, \quad \mathbf{b}^{(1)} = \begin{pmatrix} -1 \\ 1 \\ 2 \end{pmatrix}$$

rispettivamente la matrice dei coefficienti e il vettore dei termini noti del sistema di partenza. Cerchiamo ora di determinare un sistema lineare equivalente a quello iniziale ma che abbia gli elementi sottodiagonali della prima colonna uguali a zero. Lasciamo inalterata la prima equazione. Poniamo

$$l_{21} = -\frac{a_{21}}{a_{11}} = -\frac{6}{2} = -3$$

e moltiplichiamo la prima equazione per l_{21} ottenendo:

$$-6x_1 - 3x_2 - 3x_3 = 3$$

La nuova seconda equazione sarà la somma tra la seconda equazione e la prima moltiplicata per l_{21} :

$$\begin{array}{r} 6x_1 + 2x_2 + x_3 = 1 \\ -6x_1 - 3x_2 - 3x_3 = 3 \\ \hline -x_2 - 2x_3 = 4 \quad \text{[Nuova seconda equazione].} \end{array}$$

Poniamo

$$l_{31} = -\frac{a_{31}^{(1)}}{a_{11}^{(1)}} = -\frac{4}{2} = -2$$

e moltiplichiamo la prima equazione per l_{31} ottenendo:

$$-4x_1 - 2x_2 - 2x_3 = 2$$

La nuova terza equazione sarà la somma tra la terza equazione e la prima moltiplicata per l_{31} :

$$\begin{array}{rcl} 4x_1 & -2x_2 & +x_3 & = 2 \\ -4x_1 & -2x_2 & -2x_3 & = 2 \\ \hline & -4x_2 & -x_3 & = 4 \quad [\text{Nuova terza equazione}]. \end{array}$$

Al secondo passo il sistema lineare è diventato:

$$\begin{array}{rcl} 2x_1 & +x_2 & +x_3 & = -1 \\ & -x_2 & -2x_3 & = 4 \\ & -4x_2 & -x_3 & = 4 \end{array}$$

La matrice dei coefficienti e il vettore dei termini noti sono diventati:

$$A^{(2)} = \begin{pmatrix} 2 & 1 & 1 \\ 0 & -1 & -2 \\ 0 & -4 & -1 \end{pmatrix}, \quad \mathbf{b}^{(2)} = \begin{pmatrix} -1 \\ 4 \\ 4 \end{pmatrix}.$$

Cerchiamo ora di azzerare gli elementi sottodiagonali della seconda colonna. Lasciamo inalterata le prime due equazioni del sistema. Poniamo

$$l_{32} = -\frac{a_{32}^{(2)}}{a_{22}^{(2)}} = -\frac{-4}{-1} = -4$$

e moltiplichiamo la seconda equazione per l_{32} ottenendo:

$$4x_2 + 8x_3 = -16$$

La nuova terza equazione sarà la somma tra la terza equazione e la seconda appena modificata

$$\begin{array}{rcl} -4x_2 & -x_3 & = 4 \\ 4x_2 & +8x_3 & = -16 \\ \hline & 7x_3 & = -12 \quad [\text{Nuova terza equazione}]. \end{array}$$

Abbiamo ottenuto un sistema triangolare superiore:

$$\begin{array}{rcl} 2x_1 & +x_2 & +x_3 & = -1 \\ & -x_2 & -2x_3 & = 4 \\ & & 7x_3 & = -12. \end{array}$$

La matrice dei coefficienti e il vettore dei termini noti sono diventati:

$$A^{(3)} = \begin{pmatrix} 2 & 1 & 1 \\ 0 & -1 & -2 \\ 0 & 0 & 7 \end{pmatrix}, \quad \mathbf{b}^{(3)} = \begin{pmatrix} -1 \\ 4 \\ -12 \end{pmatrix}.$$

Vediamo ora di calcolare le formule che consentano di calcolare gli elementi della matrice dei coefficienti e del vettore dei termini noti ad ogni passo del metodo di Gauss. Abbiamo detto che $A^{(1)}$ e $\mathbf{b}^{(1)}$ sono assegnati inizialmente, ipotizziamo per il momento che $a_{11}^{(1)} \neq 0$. Calcoliamo ora gli stessi dati al passo 2 tenendo presente che:

1. La prima equazione del sistema resta invariata;
2. Gli elementi sottodiagonali della prima colonna di $A^{(2)}$ sono nulli;
3. La i -esima riga del sistema ($i \geq 2$) è ottenuta sommando alla medesima riga la prima moltiplicata per $-a_{i1}^{(1)}/a_{11}^{(1)}$.

Fissiamo quindi una riga i , $i \geq 2$, e calcoliamo gli elementi $a_{ij}^{(2)}$:

$$\begin{array}{cccccccccc} a_{i1}^{(1)} & a_{i2}^{(1)} & a_{i3}^{(1)} & \dots & a_{ij}^{(1)} & \dots & a_{in}^{(1)} & b_i^{(1)} & + \\ -\frac{a_{i1}^{(1)}}{a_{11}^{(1)}} \times & a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \dots & a_{1j}^{(1)} & \dots & a_{1n}^{(1)} & b_1^{(1)} & = \\ \hline 0 & a_{i2}^{(2)} & a_{i3}^{(2)} & \dots & a_{ij}^{(2)} & \dots & a_{in}^{(2)} & b_i^{(2)} & & \end{array}$$

dove

$$a_{ij}^{(2)} = a_{ij}^{(1)} - \frac{a_{i1}^{(1)}}{a_{11}^{(1)}} a_{1j}^{(1)}, \quad i, j = 2, \dots, n$$

e

$$b_i^{(2)} = b_i^{(1)} - \frac{a_{i1}^{(1)}}{a_{11}^{(1)}} b_1^{(1)}, \quad i = 2, \dots, n.$$

Se ipotizziamo che $a_{22}^{(2)} \neq 0$ possiamo calcolare gli elementi del sistema al passo 3 tenendo presente che:

1. Le prime 2 equazioni del sistema restano invariate;

2. Gli elementi sottodiagonali della prime 2 colonna di $A^{(3)}$ sono nulli;
3. La i -esima riga del sistema ($i \geq 3$) è ottenuta sommando alla medesima riga la seconda moltiplicata per $-a_{i2}^{(2)}/a_{22}^{(2)}$.

Fissiamo quindi una riga i , $i \geq 3$, e calcoliamo gli elementi $a_{ij}^{(3)}$:

$$\begin{array}{cccccccc}
 0 & a_{i2}^{(2)} & a_{i3}^{(2)} & \dots & a_{ij}^{(2)} & \dots & a_{in}^{(2)} & b_i^{(2)} & + \\
 -\frac{a_{i2}^{(2)}}{a_{22}^{(2)}} \times & 0 & a_{22}^{(2)} & a_{23}^{(2)} & \dots & a_{2j}^{(2)} & \dots & a_{2n}^{(2)} & b_2^{(2)} & = \\
 \hline
 0 & 0 & a_{i3}^{(3)} & \dots & a_{ij}^{(3)} & \dots & a_{in}^{(3)} & b_i^{(3)} & &
 \end{array}$$

dove

$$a_{ij}^{(3)} = a_{ij}^{(2)} - \frac{a_{i2}^{(2)}}{a_{22}^{(2)}} a_{2j}^{(2)}, \quad i, j = 3, \dots, n$$

e

$$b_i^{(3)} = b_i^{(2)} - \frac{a_{i2}^{(2)}}{a_{22}^{(2)}} b_2^{(2)}, \quad i = 3, \dots, n.$$

Avendo ricavato esplicitamente le formule per i primi due passi del metodo di Gauss è semplice ricavare quelle per un generico passo k . La matrice $A^{(k)}$ ha gli elementi sottodiagonali delle prime $k - 1$ colonne uguali a zero, e, supposto $a_{kk}^{(k)} \neq 0$, gli elementi di $A^{(k+1)}$ e di $\mathbf{b}^{(k+1)}$ sono quindi:

$$a_{ij}^{(k+1)} = a_{ij}^{(k)} - \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} a_{kj}^{(k)}, \quad i, j = k + 1, \dots, n \quad (1.9)$$

e

$$b_i^{(k+1)} = b_i^{(k)} - \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} b_k^{(k)}, \quad i = k + 1, \dots, n. \quad (1.10)$$

Il valore di k varia da 1 (matrice dei coefficienti e vettori dei termini noti iniziali) fino a $n - 1$, infatti la matrice $A^{(n)}$ avrà gli elementi sottodiagonali delle prime $n - 1$ colonne uguali a zero.

Tutto il discorso fatto finora va bene se gli elementi $a_{kk}^{(k)}$ sono diversi da zero

per ogni k , prima di affrontare come modificare il metodo di Gauss qualora tale situazione non si verifichi consideriamo una formulazione alternativa dello stesso metodo.

Sia $L^{(1)}$ una matrice triangolare inferiore $n \times n$ così definita:

$$L^{(1)} = \begin{pmatrix} 1 & & & & \\ l_{21} & 1 & & & 0 \\ l_{31} & 0 & 1 & & \\ \vdots & \vdots & \ddots & \ddots & \\ l_{n1} & 0 & \dots & 0 & 1 \end{pmatrix},$$

con

$$l_{i1} = -\frac{a_{i1}^{(1)}}{a_{11}^{(1)}} \quad i = 2, \dots, n. \quad (1.11)$$

Per calcolare la matrice prodotto $L^{(1)}A^{(1)}$ decomponiamo a blocchi le due matrici nel seguente modo:

$$L^{(1)} = \begin{pmatrix} 1 & 0^T \\ \mathbf{l}_1 & I_{n-1} \end{pmatrix} \quad A^{(1)} = \begin{pmatrix} a_{11}^{(1)} & \mathbf{a}_1^T \\ \widehat{\mathbf{a}}_1 & A_{22}^{(1)} \end{pmatrix}$$

avendo indicato con \mathbf{l}_1 il vettore i cui elementi sono definiti da (1.11), con \mathbf{a}_1^T gli elementi della prima riga di $A^{(1)}$, escluso il primo, con $\widehat{\mathbf{a}}_1^T$ gli elementi della prima colonna di $A^{(1)}$, escluso il primo, e con I_{n-1} la matrice identità di ordine $n - 1$. Calcoliamo ora il prodotto tra queste due matrici:

$$\begin{aligned} L^{(1)}A^{(1)} &= \begin{pmatrix} 1 & 0^T \\ \mathbf{l}_1 & I_{n-1} \end{pmatrix} \begin{pmatrix} a_{11}^{(1)} & \mathbf{a}_1^T \\ \widehat{\mathbf{a}}_1 & A_{22}^{(1)} \end{pmatrix} = \\ &= \begin{pmatrix} a_{11}^{(1)} & \mathbf{a}_1^T \\ \mathbf{l}_1 a_{11}^{(1)} + \widehat{\mathbf{a}}_1 & A_{22}^{(1)} + \mathbf{l}_1 \mathbf{a}_1^T \end{pmatrix}. \end{aligned}$$

Osserviamo preliminarmente che la prima riga della matrice prodotto coincide con la prima riga di $A^{(1)}$. Per gli elementi della prima colonna:

$$\mathbf{l}_1 a_{11}^{(1)} + \widehat{\mathbf{a}}_1 = \begin{pmatrix} -a_{21}^{(1)} \\ -a_{31}^{(1)} \\ \vdots \\ -a_{n1}^{(1)} \end{pmatrix} + \begin{pmatrix} a_{21}^{(1)} \\ a_{31}^{(1)} \\ \vdots \\ a_{n1}^{(1)} \end{pmatrix} = 0.$$

e gli elementi $a_{ij}^{(k)}$ e $b_i^{(k)}$ sono dati dalle relazioni (1.9) e (1.10).

Nell' eseguire il metodo di Gauss si è fatta l'implicita ipotesi (vedi formule (1.9) e (1.10)) che i cosiddetti *elementi pivotali* $a_{kk}^{(k)}$ siano non nulli per $k = 1, 2, \dots, n - 1$. In vero questa non è un'ipotesi limitante in quanto la non singolarità di A permette, con un opportuno scambio di righe in $A^{(k)}$, di ricondursi a questo caso. Infatti scambiare due righe in $A^{(k)}$ significa sostanzialmente scambiare due equazioni nel sistema $A^{(k)}\mathbf{x} = \mathbf{b}^{(k)}$ e ciò non altera la natura del sistema stesso.

Consideriamo la matrice $A^{(k)}$ e supponiamo $a_{kk}^{(k)} = 0$. Se $a_{ik}^{(k)} = 0$ per $i = k + 1, \dots, n$, allora $A^{(k)}$ ha la seguente struttura:

$$A^{(k)} = \begin{pmatrix} a_{11}^{(1)} & \cdots & a_{1,k-1}^{(1)} & a_{1k}^{(1)} & a_{1,k+1}^{(1)} & \cdots & a_{1n}^{(1)} \\ & \ddots & \vdots & \vdots & \vdots & & \vdots \\ & & a_{k-1,k-1}^{(k-1)} & a_{k-1,k}^{(k-1)} & a_{k-1,k+1}^{(k-1)} & \cdots & a_{k-1,n}^{(k-1)} \\ & & & 0 & a_{k,k+1}^{(k)} & & a_{kn}^{(k)} \\ & 0 & & \vdots & \vdots & & \vdots \\ & & & 0 & a_{n,k+1}^{(k)} & \cdots & a_{nn}^{(k)} \end{pmatrix}$$

Partizioniamo $A^{(k)}$ nel seguente modo

$$A^{(k)} = \begin{pmatrix} T_{k-1} & * \\ 0 & \hat{T}_{n-k+1} \end{pmatrix}$$

segue che $\det A^{(k)} = 0$ e di conseguenza dovrebbe essere anche $\det A = 0$ e questo contrasta con l'ipotesi fatta. Quindi possiamo concludere che se $a_{kk}^{(k)} = 0$ e $\det A \neq 0$ deve necessariamente esistere un elemento $a_{ik}^{(k)} \neq 0$, con $i \in \{k + 1, k + 2, \dots, n\}$.

1.2.3 Equivalenza tra Fattorizzazione LU e Metodo di Gauss

Vogliamo ora provare che il metodo di eliminazione di Gauss senza scambio di righe equivale a fattorizzare la matrice A dei coefficienti nel prodotto di una matrice triangolare inferiore, che denotiamo con L avente elementi diagonali tutti uguali a 1, e una matrice triangolare superiore U :

$$A = LU.$$

A tal fine risultano utili i seguenti risultati.

1.2.4 Esistenza e Unicità della Fattorizzazione LU

Per le condizioni di esistenza della fattorizzazione LU si deve considerare un importante risultato preliminare che però non dimostriamo.

Detta A_k la sottomatrice principale di testa di ordine k della matrice A esiste una relazione che lega gli elementi pivotali e i minori principali di una matrice (cioè i determinanti delle sottomatrici di testa di A). In particolare se

$$A_k = \begin{pmatrix} a_{11} & \dots & a_{1k} \\ \vdots & & \vdots \\ a_{k1} & \dots & a_{kk} \end{pmatrix}$$

allora

$$a_{kk}^{(k)} = \frac{\det A_k}{\det A_{k-1}} \quad k = 2, \dots, n$$

e $a_{11}^{(1)} = \det A_1 = a_{11}$.

Una naturale conseguenza di questo risultato è il seguente teorema di esistenza e unicità della fattorizzazione LU .

Teorema 1.2.1 *Sia A una matrice quadrata di ordine n non singolare. Sia A_k la sottomatrice principale di testa di ordine k di A e sia $\det A_k \neq 0$ per $k = 1, 2, \dots, n$, cioè siano i minori principali di testa non nulli.*

Allora esistono un'unica matrice L triangolare inferiore con elementi diagonali uguali a 1 ed un'unica U triangolare superiore tali che: $A = LU$.

1.2.5 Strategie di Pivoting nel Metodo di Eliminazione di Gauss

Le strategie di pivoting nel metodo di Gauss hanno principalmente lo scopo di evitare gli elementi pivotali nulli. Infatti al k -esimo passo la matrice $A^{(k)}$ si presenta nella forma:

$$\begin{pmatrix} a_{11}^{(1)} & \dots & a_{1,k-1}^{(1)} & a_{1k}^{(1)} & \dots & a_{1n}^{(1)} \\ & \ddots & \vdots & \vdots & & \vdots \\ & & a_{k-1,k-1}^{(k-1)} & a_{k-1,k}^{(k-1)} & \dots & a_{k-1,n}^{(k-1)} \\ & & 0 & a_{kk}^{(k)} & \dots & a_{kn}^{(k)} \\ & & \vdots & \vdots & & \vdots \\ & & 0 & a_{nk}^{(k)} & \dots & a_{nn}^{(k)} \end{pmatrix}.$$

Il passo successivo consiste nell'azzerare gli elementi al di sotto dell'elemento $a_{kk}^{(k)}$ situati nella k -esima colonna. La strategia di *Pivoting parziale* prevede che prima di fare ciò si ricerchi l'elemento di massimo modulo tra gli elementi $a_{kk}^{(k)}, a_{k+1,k}^{(k)}, \dots, a_{nk}^{(k)}$ e si scambii la riga in cui si trova questo elemento con la k -esima qualora esso sia diverso da $a_{kk}^{(k)}$. In altri termini il pivoting parziale richiede le seguenti operazioni:

1. determinare l'elemento $a_{rk}^{(k)}$ tale che

$$|a_{rk}^{(k)}| = \max_{k \leq i \leq n} |a_{ik}^{(k)}|;$$

2. effettuare lo scambio tra la r -esima e la k -esima riga.

In alternativa alla strategia di pivoting parziale si può effettuare la strategia di *Pivoting totale* che costa un po' di più ma che rende più stabile il metodo di Gauss. La strategia di pivoting totale è la seguente:

1. determinare gli indici r, s tali che

$$|a_{rs}^{(k)}| = \max_{k \leq i, j \leq n} |a_{ij}^{(k)}|;$$

2. effettuare lo scambio tra la r -esima e la k -esima riga e tra la s -esima e la k -esima colonna.

1.2.6 Fattorizzazione di Matrici Non Singolari

Abbiamo visto che se A , matrice quadrata di ordine n , è non singolare e tutti i suoi minori principali di testa fino all'ordine $n - 1$ sono non nulli allora A ammette un'unica fattorizzazione LU . Se A è solo non singolare non è detto che ammetta fattorizzazione LU . In particolare si può dimostrare che è sempre possibile trovare una *matrice di permutazione* P (vedere definizione 1.1.1) tale che $PA = LU$.

Infatti vale il seguente risultato.

Teorema 1.2.2 *Se A è una matrice quadrata non singolare allora esistono una matrice di permutazione P , una matrice L triangolare inferiore con elementi uguali a 1 sulla diagonale principale ed una matrice U triangolare superiore tali che:*

$$PA = LU.$$

1.2.7 Classi di Matrici che non hanno bisogno di Pivoting

Tra le classi di matrici che non hanno bisogno di alcuna strategia di pivoting è opportuno citare le matrici a predominanza per righe e le matrici simmetriche e definite positive.

Definizione 1.2.1 *A si dice a predominanza diagonale per colonne se:*

$$|a_{jj}| \geq \sum_{i=1, i \neq j}^n |a_{ij}|, \quad j = 1, 2, \dots, n.$$

Infatti è possibile dimostrare che la predominanza diagonale è invariante sotto trasformazioni elementari di Gauss, cioè le sottomatrici trasformate ad ogni passo sono pure a predominanza diagonale per colonne di conseguenza non è necessario utilizzare alcuna strategia di pivoting.

Definizione 1.2.2 *Una matrice quadrata A , simmetrica, di ordine n si dice definita positiva se e solo se:*

$$\mathbf{x}^T A \mathbf{x} > 0 \quad \text{per ogni } \mathbf{x} \in \mathbb{R}^n, \mathbf{x} \neq 0.$$

Uno dei criteri possibili per determinare se una matrice è definita positiva è fornito dal *criterio di Sylvester* che stabilisce che una matrice A simmetrica di ordine n è definita positiva se e solo se $\det A_k > 0$ per $k = 1, 2, \dots, n$, dove A_k è la sottomatrice principale di testa di ordine k . Poichè è noto che

$$a_{kk}^{(k)} = \frac{\det A_k}{\det A_{k-1}} \quad k = 1, \dots, n, \quad \det A_0 = 1$$

dove gli $a_{kk}^{(k)}$ sono gli elementi pivotali nel metodo di Gauss. Ad ogni passo l'elemento pivotale è positivo quindi non è necessaria alcuna strategia di pivoting.

1.2.8 Fattorizzazione Diretta

Quando risolviamo un sistema lineare con il metodo di Gauss dobbiamo calcolare circa $n^3/3$ risultati intermedi. È comunque possibile riarrangiare i calcoli in modo tale da valutare direttamente gli elementi di L ed U . Il

metodo che ne risulta prende il nome di *Metodo di Doolittle*.

Fissata la matrice A , quadrata di ordine n , imponiamo che risulti

$$A = LU$$

con L matrice triangolare inferiore con elementi diagonali uguali a 1 e U matrice triangolare superiore. Una volta note tali matrici il sistema di partenza $A\mathbf{x} = \mathbf{b}$ viene scritto come

$$LU\mathbf{x} = \mathbf{b}$$

e, posto $U\mathbf{x} = \mathbf{y}$, il vettore \mathbf{x} viene trovato prima risolvendo il sistema triangolare inferiore

$$L\mathbf{y} = \mathbf{b}$$

e poi quello triangolare superiore

$$U\mathbf{x} = \mathbf{y}.$$

Imponiamo quindi che la matrice A ammetta fattorizzazione LU :

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} = \begin{pmatrix} l_{11} & 0 & \dots & 0 \\ l_{21} & l_{22} & \ddots & \vdots \\ \vdots & & \ddots & 0 \\ l_{n1} & l_{n2} & \dots & l_{nn} \end{pmatrix} \begin{pmatrix} u_{11} & u_{12} & \dots & u_{1n} \\ 0 & u_{22} & \dots & u_{2n} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & u_{nn} \end{pmatrix}.$$

Deve essere

$$a_{ij} = \sum_{k=1}^{\min(i,j)} l_{ik}u_{kj} \quad i, j = 1, \dots, n. \quad (1.12)$$

Uguagliando la parte triangolare superiore di A abbiamo

$$a_{ij} = \sum_{k=1}^i l_{ik}u_{kj} \quad j \geq i \quad (1.13)$$

ovvero

$$a_{ij} = \sum_{k=1}^{i-1} l_{ik}u_{kj} + l_{ii}u_{ij} \quad j \geq i.$$

Imponendo $l_{ii} = 1$ risulta

$$u_{ij} = a_{ij} - \sum_{k=1}^{i-1} l_{ik}u_{kj} \quad j \geq i \quad (1.14)$$

e ovviamente $u_{1j} = a_{1j}$, per $j = 1, \dots, n$. Uguagliamo ora la parte triangolare strettamente inferiore di A . Sempre da (1.12) risulta:

$$a_{ij} = \sum_{k=1}^j l_{ik}u_{kj} \quad i > j \quad (1.15)$$

ovvero

$$a_{ij} = \sum_{k=1}^{j-1} l_{ik}u_{kj} + l_{ij}u_{jj} \quad i > j$$

da cui

$$l_{ij} = \frac{1}{u_{jj}} \left(a_{ij} - \sum_{k=1}^{j-1} l_{ik}u_{kj} \right) \quad i \geq j. \quad (1.16)$$

Si osservi che le formule (1.14) e (1.16) vanno implementate secondo uno degli schemi riportati in Figura 1.1.

Ogni schema rappresenta in modo stilizzato una matrice la cui parte triangolare superiore indica la matrice U mentre quella triangolare inferiore la matrice L mentre i numeri indicano l'ordine con cui gli elementi saranno calcolati. Per esempio applicando la tecnica di Crout si segue il seguente ordine:

- 1° Passo: Calcolo della prima riga di U ;
- 2° Passo: Calcolo della seconda riga di L ;
- 3° Passo: Calcolo della seconda riga di U ;
- 4° Passo: Calcolo della terza riga di L ;
- 5° Passo: Calcolo della terza riga di U ;
- 6° Passo: Calcolo della quarta riga di L ;
- 7° Passo: Calcolo della quarta riga di U ;

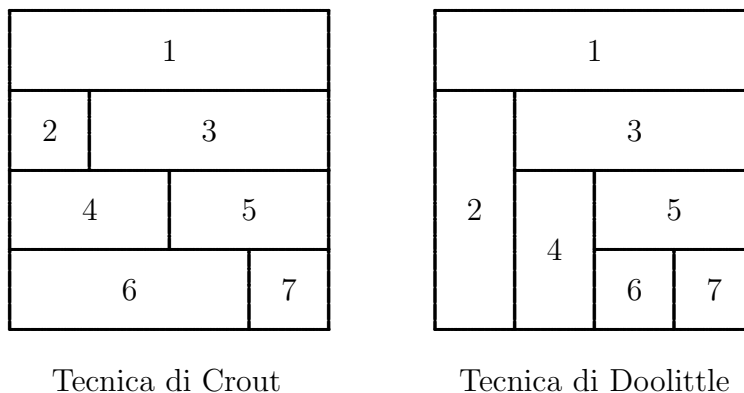


Figura 1.1: Tecniche di calcolo diretto della fattorizzazione di una matrice

e così via procedendo per righe in modo alternato. Nel caso della tecnica di Doolittle si seguono i seguenti passi:

- 1° Passo: Calcolo della prima riga di U ;
- 2° Passo: Calcolo della prima colonna di L ;
- 3° Passo: Calcolo della seconda riga di U ;
- 4° Passo: Calcolo della seconda colonna di L ;
- 5° Passo: Calcolo della terza riga di U ;
- 6° Passo: Calcolo della terza colonna di L ;
- 7° Passo: Calcolo della quarta riga di U .

Lo svantaggio del metodo di fattorizzazione diretto risiede essenzialmente nella maggiore difficoltà, rispetto al metodo di Gauss, di poter programmare una strategia di pivot. Per matrici simmetriche e definite positive, poichè non è necessaria alcuna strategia di pivoting, il metodo diventa particolarmente vantaggioso.

1.2.9 Condizionamento di sistemi lineari

Uno dei più importanti aspetti legati all'uso di algoritmi numerici per la soluzione di problemi matematici è l'affidabilità dei risultati ottenuti. I metodi descritti in questo capitolo funzionerebbero sempre bene (cioè fornirebbero risultati numericamente affidabili) se non si dovesse aver a che fare con una serie di problemi legati alla struttura dell'elaboratore che stiamo utilizzando. Un primo problema che ci troviamo ad affrontare è il modo con cui i numeri reali sono rappresentati nella memoria di un elaboratore. Giusto per rendersi conto delle difficoltà che si incontrano va osservato che i numeri reali sono infiniti mentre la memoria di un calcolatore ha una capacità finita. Una seconda osservazione consiste nel fatto che un numero reale ammette molteplici rappresentazioni. Per esempio il numero 12.47 può essere scritto in diversi modi

$$12.47 = 1.247 \times 10^1 = 0.1247 \times 10^2 = 1247 \times 10^{-2}.$$

Questa prima ambiguità viene risolta convenzionalmente utilizzando come rappresentazione la seguente:

$$x = \text{segno}(x) q \times 10^p$$

dove $\text{segno}(x) = \pm 1$, $q = 0.d_1d_2d_3 \dots d_k \dots$ è un numero reale positivo minore di 1 con le cifre d_i comprese tra 0 e 9 se si utilizza la rappresentazione in base 10, e si impone $d_1 \neq 0$, cioè

$$\frac{1}{10} \leq q < 1.$$

Lo stesso discorso si può ripetere scegliendo una qualsiasi base $\beta \in \mathbb{N}$, cioè

$$x = \text{segno}(x) q \times \beta^p, \quad \text{dove } q = 0.d_1d_2d_3 \dots d_k \dots$$

con le cifre d_i comprese tra 0 e $\beta - 1$, e $d_1 \neq 0$, cioè

$$\frac{1}{\beta} \leq q < 1.$$

Assegnato $x \in \mathbb{R}$, $x \neq 0$, l'espressione

$$x = \text{segno}(x) \beta^p \times 0.d_1d_2 \dots d_k \dots$$

prende il nome di *rappresentazione in base β di x* . Il numero p viene detto *esponente* (o *caratteristica*), i valori d_i sono le cifre della rappresentazione,

mentre $0.d_1d_2\dots d_k\dots$ si dice *mantissa*. Il numero x viene normalmente rappresentato con la cosiddetta *notazione posizionale* $x = \text{segno}(x)(.d_1d_2d_3\dots) \times \beta^p$, che viene detta *normalizzata*. In alcuni casi è ammessa una rappresentazione in notazione posizionale tale che $d_1 = 0$, che viene detta *denormalizzata*. Quindi un qualunque numero reale $x \neq 0$ può essere rappresentato con *infinite cifre* nella mantissa. Inoltre come è noto l'insieme dei numeri reali ha cardinalità infinita. Poichè un elaboratore è dotato di *memoria finita* non è possibile memorizzare:

- a) gli infiniti numeri reali
- b) le infinite (in generale) cifre di un numero reale.

Risulta perciò importante stabilire dei criteri che permettano di rappresentare in macchina i numeri reali che è possibile rappresentare in modo da commettere il minimo errore possibile.

Assegnati i numeri $\beta, t, m, M \in \mathbb{N}$ con $\beta \geq 2, t \geq 1, m, M > 0$, si dice *insieme dei numeri di macchina con rappresentazione normalizzata in base β con t cifre significative* l'insieme:

$$\mathcal{F}(\beta, t, m, M) = \{ x \in \mathbb{R} : x = \pm \beta^p \times 0.d_1d_2\dots d_t \text{ con } 0 \leq d_i \leq \beta - 1, \\ d_1 \neq 0, -m \leq p \leq M \} \cup \{0\}.$$

Infatti poichè zero sfugge alla rappresentazione in base normalizzata viene assegnato per definizione all'insieme \mathcal{F} . Normalmente lo zero viene rappresentato con mantissa nulla ed esponente $-m$.

Osserviamo che un elaboratore che abbia le seguenti caratteristiche:

- t campi di memoria per la mantissa, ciascuno dei quali può assumere β differenti configurazioni (e perciò può memorizzare una cifra d_i),
- un campo di memoria che può assumere $m + M + 1$ differenti configurazioni (e perciò può memorizzare i differenti valori p dell'esponente),
- un campo che può assumere due differenti configurazioni (e perciò può memorizzare il segno $+ o -$),

è in grado di rappresentare esattamente tutti gli elementi dell'insieme $\mathcal{F}(\beta, t, m, M)$.

Assumiamo ora che $x \in \mathbb{R}$ e che abbia la seguente rappresentazione in base β :

$$x = \text{segno}(x) \beta^p \times 0.d_1d_2\dots d_t$$

con $d_1 \neq 0$ e $p \in [-m, M]$. Allora è evidente che $x \in \mathcal{F}(\beta, t, m, M)$ e pertanto verrà rappresentato esattamente su un qualunque elaboratore che

utilizzi $\mathcal{F}(\beta, t, m, M)$ come insieme dei numeri di macchina.

Assumiamo ora che $x \in \mathbb{R}$ ma $x \notin \mathcal{F}(\beta, t, m, M)$. In questo caso si pone il problema di associare ad x un numero di macchina che lo rappresenti in modo da commettere il più piccolo errore possibile. Per risolvere questo problema facciamo innanzitutto, e solo per semplicità di esposizione, le seguenti ipotesi $x \in \mathbb{R}$, $x > 0$ e β numero pari.

Distinguiamo quindi i seguenti casi:

- a) $p \notin [-m, M]$. Se $p < -m$ allora x è più piccolo del più piccolo numero di macchina: in questo caso si dice che si è verificato un *underflow* (l'elaboratore interrompe la sequenza di calcoli e segnala con un messaggio l'underflow). Se $p > M$ allora vuol dire che x è più grande del più grande numero di macchina e in questo caso si dice che si è verificato un *overflow* (anche in questo caso l'elaboratore si ferma e segnala l'overflow).
- b) $p \in [-m, M]$, $x = \beta^p \times 0.d_1d_2 \dots d_t \dots$, ed esiste un $k > t$ tale che $d_k \neq 0$. Anche in questo caso poichè x ha più di t cifre significative $x \notin \mathcal{F}$. È però possibile rappresentare x mediante un numero in \mathcal{F} con un'opportuna operazione di taglio delle cifre decimali che seguono la t -esima. In pratica i criteri di taglio sono i seguenti:

1. *troncamento di x alla t -esima cifra significativa*

$$\tilde{x} = \text{tr}(x) = \beta^p \times 0.d_1d_2 \dots d_t$$

2. *arrotondamento di x alla t -esima cifra significativa*

$$\tilde{x} = \text{arr}(x) = \beta^p \times 0.d_1d_2 \dots \tilde{d}_t$$

dove

$$\tilde{d}_t = \begin{cases} d_t + 1 & \text{se } d_{t+1} \geq \beta/2 \\ d_t & \text{se } d_{t+1} < \beta/2. \end{cases}$$

Se $x \in \mathbb{R}$ e \tilde{x} è la sua rappresentazione di macchina, chiameremo *errore assoluto* la quantità

$$E_a = |x - \tilde{x}|$$

mentre per $x \neq 0$ chiameremo *errore relativo* la quantità

$$E_r = \frac{|x - \tilde{x}|}{|x|}.$$

Nel seguito assumeremo $x > 0$ e supporremo anche che la rappresentazione di x in $\mathcal{F}(\beta, t, m, M)$ non dia luogo ad underflow o overflow.

Teorema 1.2.3 *Sia $x = \pm\beta^p 0.d_1d_2\dots d_t\dots$ tale che la sua rappresentazione macchina non dia luogo a fenomeni di underflow o overflow, allora risulta:*

$$\begin{aligned} |tr(x) - x| &< \beta^{p-t} \\ |arr(x) - x| &\leq \frac{1}{2}\beta^{p-t} \end{aligned}$$

dove il segno di uguaglianza vale se e solo se $d_{t+1} = \frac{\beta}{2}$ e $d_{t+i} = 0$ per $i \geq 2$.

Teorema 1.2.4 *Sia $x = \pm\beta^p 0.d_1d_2\dots d_t\dots$, $x \neq 0$, se \tilde{x} è la sua rappresentazione di macchina cioè $\tilde{x} \in \mathcal{F}(\beta, t, m, M)$, allora*

$$\begin{aligned} \left| \frac{\tilde{x} - x}{x} \right| &< u \\ \left| \frac{\tilde{x} - x}{\tilde{x}} \right| &< u \end{aligned}$$

dove

$$u = \begin{cases} \beta^{-t+1} & \text{se } \tilde{x} = tr(x) \\ \frac{1}{2}\beta^{-t+1} & \text{se } \tilde{x} = arr(x). \end{cases}$$

La quantità u che interviene nel precedente teorema si chiama *precisione di macchina* o *zero macchina*. La rappresentazione di $x \in \mathbb{R}$ attraverso $\tilde{x} \in \mathcal{F}(\beta, t, m, M)$ si dice *rappresentazione in virgola mobile di x* o *rappresentazione floating point*, con troncamento se $\tilde{x} = tr(x)$, con arrotondamento se $\tilde{x} = arr(x)$.

Se \tilde{x} è un'approssimazione di $x \in \mathbb{R}$ con un errore relativo minore di β^{1-t} , si dice che t *cifre della rappresentazione in base β sono significative*.

Un problema simile si presenta anche quando si effettuano delle operazioni aritmetiche su numeri reali. In fatti non è garantito, in generale, che un'operazione aritmetica su due numeri macchina fornisca come risultato un numero di macchina. Per esempio considerati $x = (.11)10^0$ e $y = (.11)10^{-2}$, $x, y \in \mathcal{F}(10, 2, m, M)$, si ha:

$$x + y = (.1111)10^0 \notin \mathcal{F}(10, 2, m, M)$$

Per poter realizzare la naturale ed importante *Proprietà di chiusura* di una operazione in un certo insieme risulta importante definire delle *operazioni di macchina* che permettano appunto di realizzare tale proprietà. Un requisito essenziale che si richiede nel costruire un'aritmetica di macchina è il seguente. Indicata con \cdot una delle quattro operazioni aritmetiche $+$, $-$, \times , \div e con \odot la corrispondente operazione di macchina dev'essere:

$$x \odot y = (x \cdot y)(1 + \varepsilon), \quad |\varepsilon| < u \quad (1.17)$$

per ogni $x, y \in \mathcal{F}(\beta, t, m, M)$ tali che $x \odot y$ non dia luogo ad overflow o underflow. Si può dimostrare che

$$x \odot y = \text{tr}(x \cdot y)$$

e

$$x \odot y = \text{arr}(x \cdot y)$$

soddisfano la (1.17) e dunque danno luogo ad operazioni di macchina. Le quattro operazioni così definite danno luogo alla *aritmetica di macchina* o *aritmetica finita*.

Si può dimostrare che per le operazioni di macchina non valgono alcune proprietà, come per esempio l'associatività dell'addizione e della moltiplicazione o la distributività della moltiplicazione rispetto all'addizione, che invece sono valide per le operazioni tra numeri reali.

Supponiamo ora di voler valutare la differenza tra due numeri reali x e y . Siano $\text{fl}(x)$ e $\text{fl}(y)$ rispettivamente le loro rappresentazioni di macchina. Vogliamo vedere quale è l'errore relativo che viene commesso dall'elaboratore quando calcola $x - y$.

$$\begin{aligned} \text{fl}(x) \ominus \text{fl}(y) &= [\text{fl}(x) - \text{fl}(y)](1 + \varepsilon) = \\ &= [x(1 + \varepsilon_x) - y(1 + \varepsilon_y)](1 + \varepsilon) = \\ &= (x + a\varepsilon_x - y - y\varepsilon_y)(1 + \varepsilon) = \\ &= (x - y) + (x - y)\varepsilon + x\varepsilon_x - y\varepsilon_y + x\varepsilon\varepsilon_x - y\varepsilon\varepsilon_y. \end{aligned}$$

Una maggiorazione per l'errore relativo è la seguente

$$\begin{aligned} \frac{|(fl(x) \ominus fl(y)) - (x - y)|}{|x - y|} \leq & |\varepsilon| + \frac{|x|}{|x - y|} (|\varepsilon_x| + |\varepsilon| |\varepsilon_x|) + \\ & + \frac{|y|}{|x - y|} (|\varepsilon_y| + |\varepsilon| |\varepsilon_y|). \end{aligned} \quad (1.18)$$

Se x e y hanno segno opposto risulta

$$\max(|x|, |y|) \leq |x - y|$$

e dalla (1.18) segue la maggiorazione

$$\frac{|(fl(x) \ominus fl(y)) - (x - y)|}{|x - y|} \leq 3u + O(u^2)$$

dove u è la precisione di macchina ed $O(u^2)$ indica la maggiorazione per i cosiddetti termini quadratici dell'errore (cioè i prodotti del tipo $\varepsilon\varepsilon_x$, infatti tali quantità sono maggiorabili singolarmente in modulo dalla precisione di macchina ed il loro prodotto è una quantità trascurabile).

Se x e y hanno lo stesso segno allora l'errore relativo può essere molto grande quanto più x e y sono vicini. Questo fenomeno prende il nome di *cancellazione di cifre significative*.

Tornando al problema della risoluzione di un sistema lineare è ovvio che gli errori nella rappresentazione dei dati del problema A e \mathbf{b} hanno come conseguenza il fatto che, invece di risolvere il sistema

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

si risolve il sistema perturbato

$$(A + \delta A)(\mathbf{x} + \delta \mathbf{x}) = \mathbf{b} + \delta \mathbf{b}.$$

La quantità $\delta \mathbf{x}$ rappresenta quanto la soluzione del sistema che effettivamente stiamo risolvendo è distante dalla soluzione del sistema che invece vorremmo risolvere. Questa proprietà, che è tipica del problema che vogliamo risolvere e non del metodo che stiamo utilizzando, prende il nome di *condizionamento del problema* e si tratta di qualcosa che non è neanche strettamente legato ai sistemi lineari in quanto può essere definito per un qualsiasi problema

matematico. In particolare si parla di *problema bencondizionato* se piccole perturbazioni sui dati iniziali provocano una piccola perturbazione dei dati di uscita (quindi se le quantità δA , $\delta \mathbf{b}$ e $\delta \mathbf{x}$ sono tutte piccole). Si parla di *problema malcondizionato* se piccole perturbazioni sui dati iniziali provocano un grande cambiamento dei dati di uscita (quindi se le quantità δA e $\delta \mathbf{b}$ sono piccole e invece $\delta \mathbf{x}$ è molto grande). Nel caso dei sistemi lineari si può provare che vale la seguente relazione

$$\frac{\|\delta \mathbf{x}\|}{\|\mathbf{x}\|} \leq \|A\| \|A^{-1}\| \left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta \mathbf{b}\|}{\|\mathbf{b}\|} \right).$$

La quantità $\|A\| \|A^{-1}\|$ prende il nome di *indice di condizionamento di A* e viene considerato a tutti gli effetti come una misura del condizionamento del sistema. In particolare esso misura di quanto una matrice è vicina ad essere singolare.