APPUNTI DI CALCOLO NUMERICO

Enrico Bertolazzi[1] e Gianmarco Manzini[2]

[1] *Dipartimento di Ingegneria Meccanica e Strutturale*
Università degli Studi di Trento
via Mesiano 77, I – 38050 Trento, Italia
`Enrico.Bertolazzi@ing.unitn.it`

[2] *Istituto di Analisi Numerica*
del C.N.R. di Pavia
via Ferrata 1, I – 27100 Pavia, Italia
`Gianmarco.Manzini@ian.pv.cnr.it`

INDICE

PREFAZIONE

Questa raccolta di appunti è frutto dell'attività didattica svolta negli ultimi anni dagli autori presso le Università di Trento e di Pavia.

Gli argomenti trattati riguardano essenzialmente l'Algebra Lineare, che viene presentata sia da un punto di vista teorico che computazionale, ed alcuni settori "classici" dell'Analisi Numerica, come l'interpolazione, l'approssimazione ai Minimi Quadrati, l'integrazione numerica, il calcolo di zeri di funzione.

Gli appunti sono stati raccolti in questa forma per essere proposti sia agli studenti dei corsi di Laurea che di Diploma in Ingegneria. Di fatto, sono stati variamente utilizzati negli anni come materiale didattico di supporto alle lezioni di Analisi Numerica, Calcolo Numerico, e Geo-Calcolo.

Questi appunti sono stati quindi pensati come materiale di lavoro che si rivolge a studenti che non saranno formati per essere "professionisti della matematica", ma piuttosto "utenti della matematica". Si e' così privilegiato nella impostazione l'aspetto procedurale, cioè di utilizzo nella pratica dei risultati teorici.

Si è comunque cercato di mantenere un minimo di rigore formale nella esposizione, per esempio sviluppando i vari argomenti attraverso l'introduzione di definizioni, osservazioni, proposizioni, lemmi, e teoremi con relative dimostrazioni.

Ci rendiamo conto, tuttavia, che alcune scelte didattiche potrebbero disturbare i puristi del settore; per esempio, pur trattando argomenti teorici di algebra lineare con il linguaggio dei vettori, e quindi pur parlando di combinazioni lineari, dipendenza ed indipendenza lineare, basi, etc etc non viene mai introdotta una definizione formale di spazi vettoriali.

Infine, si sono esposte nel testo quasi tutte le dimostrazioni dei risultati importanti proposti (evitando soltanto quelle che ci sono sembrate troppo "tecniche" o comunque non essenziali nell'economia di questo materiale didattico). Ci rendiamo conto che ciò ha prodotto conseguentemente un

testo con molto piuù materiale di quello che può essere ragionevolmente esposto per esempio in un corso semestrale di Calcolo Numerico per i Diplomi. Tuttavia, ci è sembrato importante farlo, quanto meno per lasciare agli studenti più interessati la possibilita‘ di un approfondimento (quasi) immediato degli argomenti trattati nelle lezioni.

Infine, vogliamo ringraziare tutte le persone che ci hanno consigliato e dato suggerimenti, segnalato errori ed imprecisioni nel testo, aiutandoci a migliorare la qualità del nostro lavoro.

Enrico Bertolazzi,
Gianmarco Manzini.

CAPITOLO

UNO

MATRICI E VETTORI

1.1 Matrici e Vettori

1.1.1 Notazioni

In questi appunti i vettori saranno indicati con lettere in grassetto minuscole, ad esempio

$$\mathbf{a}, \qquad \mathbf{b}, \qquad \mathbf{c},$$

sono vettori. Le componenti dei vettori saranno indicate con la stessa lettera del vettore in corsivo, ad esempio

$$a_1, \qquad a_2, \qquad c_i, \qquad b_j,$$

sono componenti dei vettori $\mathbf{a}$, $\mathbf{b}$ e $\mathbf{c}$. Le matrici saranno indicate con lettere in grassetto maiuscole, ad esempio

$$\mathbf{A}, \qquad \mathbf{B} \qquad \mathbf{C},$$

sono matrici. Le componenti delle matrici saranno indicate con la stessa lettera in corsivo, ad esempio

$$A_{13}, \qquad B_{i3}, \qquad C_{j2},$$

sono componenti delle matrici $\mathbf{A}$, $\mathbf{B}$ e $\mathbf{C}$. Gli scalari saranno normalmente indicati con lettere greche, ad esempio

$$\alpha, \qquad \beta, \qquad \gamma, \qquad \dots$$

Con il simbolo $\mathbb{K}$ indicheremo sia il campo dei numeri reali $\mathbb{R}$ che il campo dei numeri complessi $\mathbb{C}$. Questo significa che si può sostituire a $\mathbb{K}$ sia $\mathbb{R}$ che $\mathbb{C}$ ed evitare duplicazioni nelle definizioni.

1.1.2 Matrici e Vettori

Definizione 1. Una matrice $\mathbf{A} \in \mathbb{K}^{m\times n}$ è un insieme di $m \times n$ numeri reali o complessi organizzati in m-righe ed n-colonne. Ad esempio

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ -2 & 2 & 0 \end{bmatrix},$$

è una matrice 2×3 cioè una matrice con 2 righe e 3 colonne. Di solito si indica con $\mathbf{A} \in \mathbb{R}^{m\times n}$ quando la matrice $\mathbf{A}$ è a valori reali. Analogamente si indica con $\mathbf{A} \in \mathbb{C}^{m\times n}$

quando la matrice $\mathbf{A}$ è a valori complessi. Definiremo con $\mathbf{A} \in \mathbb{K}^{m \times n}$ una matrice sia a valori reali che complessi. Se non diversamente specificato le matrici e i vettori devono intendersi a valori reali.

Definizione 2. Una matrice con lo stesso numero di righe e di colonne si dice *matrice quadrata*.

Esempio 1. Le più semplici matrici sono la matrice nulla, indicata con $\mathbf{0}$, cioè la matrice con tutti gli elementi nulli e la matrice identità indicata con $\mathbf{I}$, che è una matrice quadrata con tutti gli elementi nulli tranne quelli sulla diagonale che valgono 1:

$$\mathbf{0} = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 0 \end{bmatrix}, \qquad \mathbf{I} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{bmatrix}.$$

Definizione 3. Una matrice $1 \times n$ è detta *vettore riga* di dimensione n. In modo analogo una matrice $m \times 1$ è detta *vettore colonna* di dimensione m. Normalmente saranno considerati vettori colonna, così, quando faremo riferimento ad un vettore senza specificare se riga o colonna, intenderemo sempre vettore colonna.

Esempio 2. Le due seguenti matrici:

$$\mathbf{a} = \begin{bmatrix} 1 & 1 & 2 & 3 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} 1 \\ 3 \\ 4 \end{bmatrix},$$

si possono considerare vettori: $\mathbf{a}$ vettore riga di dimensione 4 e $\mathbf{b}$ vettore colonna di dimensione 3.

Quando si indica una componente di un vettore riga, si usa omettere l'indice di riga. Analogamente, si fa con i vettori colonna. Ad esempio, a_{12} è la seconda componente del vettore $\mathbf{a}$, ma si scrive a_2. In modo analogo dato un vettore colonna ad esempio $\mathbf{b}$ volendo indicare la terza componente che sarebbe b_{31} si omette l'indice di colonna cioè si scrive b_3.

Definizione 4. Una matrice quadrata $\mathbf{A}$ si dice *triangolare superiore* se

$$A_{ij} = 0, \qquad \text{per ogni } i > j$$

dove i è l'indice di riga e j è l'indice di colonna. Si può visualizzare come segue

$$\mathbf{A} = \begin{bmatrix} * & * & \cdots & * \\ 0 & * & \ddots & \vdots \\ \vdots & \ddots & \ddots & * \\ 0 & \cdots & 0 & * \end{bmatrix},$$

dove sono indicati esplicitamente gli elementi che sono sicuramente zero, detti anche "zeri strutturali", mentre $*$ indica un qualunque elemento che può assumere valori diversi da zero, detto per l'appunto "un non-zero" della matrice [1]. In modo analogo si definisce una matrice triangolare inferiore.

Esempio 3. Le matrici

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 0 & 2 & 3 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

sono tutte triangolari superiori, la matrice $\mathbf{C}$ è anche triangolare inferiore.

essere

1.1.3 Somma, differenza e prodotto per uno scalare

Definizione 5. Date le matrici $\mathbf{A}, \mathbf{B} \in \mathbb{K}^{m\times n}$ si definisce con $\mathbf{C} = \mathbf{A} + \mathbf{B}$ la matrice in $\mathbb{K}^{m\times n}$ risultato della somma di $\mathbf{A}$ e $\mathbf{B}$ dove

$$C_{ij} = A_{ij} + B_{ij}, \qquad i = 1, 2, \ldots, m \quad j = 1, 2, \ldots, n\,;$$

analogamente si definisce la differenza.

[1] Si noti bene, tuttavia, che la proprietà di un elemento di essere un "non-zero" non esclude necessariamente che tale elemento possa essere nullo.

Esempio 4. Date le matrici 2×3

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ -2 & 2 & 0 \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} 0 & -1 & 1 \\ 4 & 3 & 2 \end{bmatrix},$$

otteniamo

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} 1 & 1 & 4 \\ 2 & 5 & 2 \end{bmatrix}, \qquad \mathbf{A} - \mathbf{B} = \begin{bmatrix} 1 & 3 & 2 \\ -6 & -1 & -2 \end{bmatrix}.$$

Definizione 6. Data la matrice $\mathbf{A} \in \mathbb{K}^{m \times n}$ e lo scalare $\alpha \in \mathbb{K}$ si definisce con $\mathbf{B} = \alpha\mathbf{A}$ la matrice $\mathbb{K}^{m \times n}$ prodotto dello scalare α con la matrice $\mathbf{A}$ dove

$$B_{ij} = \alpha A_{ij}, \qquad i = 1, 2, \ldots, m, \quad j = 1, 2, \ldots, n$$

Esempio 5. Data la matrice $\mathbf{A} \in \mathbb{R}^{2 \times 3}$

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ -2 & 2 & 0 \end{bmatrix},$$

otteniamo

$$2\mathbf{A} = \begin{bmatrix} 2 & 4 & 6 \\ -4 & 4 & 0 \end{bmatrix}, \qquad -2.5\mathbf{A} = \begin{bmatrix} -2.5 & -5 & -7.5 \\ 5 & -5 & 0 \end{bmatrix}.$$

1.1.4 Confronto di matrici

Definizione 7. Date le matrici $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m \times n}$ con la scrittura $\mathbf{A} \geq \mathbf{B}$ si intende

$$A_{ij} \geq B_{ij} \qquad i = 1, 2, \ldots, m \quad j = 1, 2, \ldots, n.$$

In questa definizione sono inclusi i vettori riga e colonna come casi particolari di matrici $m \times 1$ e $1 \times n$.

Esempio 6. Date le matrici

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & -1 \\ 0 & 2 & 2 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 0 & 1 & -1 \\ 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 1 \end{bmatrix},$$

abbiamo $\mathbf{A} \geq \mathbf{B}$ e $\mathbf{C} \geq \mathbf{B}$ mentre $\mathbf{A}$ e $\mathbf{C}$ *non sono* confrontabili.

Definizione 8. Date le matrici $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m\times n}$ con la scrittura $\mathbf{A} > \mathbf{B}$ si intende

$$A_{ij} > B_{ij} \qquad i = 1, 2, \ldots, m \quad j = 1, 2, \ldots, n.$$

Esempio 7. Date le matrici

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix},$$

abbiamo $\mathbf{A} > \mathbf{B}$.

Definizione 9. Data una matrice $\mathbf{A} \in \mathbb{K}^{m\times n}$ si definisce con $|\mathbf{A}|$ la matrice[2] le cui componenti sono i valori assoluti (o moduli se $\mathbf{A}$ è a valori complessi) delle componenti della matrice $\mathbf{A}$, cioè[3]

$$\mathbf{B} = |\mathbf{A}| \quad \Rightarrow \quad B_{ij} = |A_{ij}| \qquad i = 1, 2, \ldots, m \quad j = 1, 2, \ldots, n.$$

Esempio 8. Date le matrici

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 1 & -2 & 1 \\ -2 & -3 & 2 \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} 1+\imath & 3+\imath 4 & 3 \\ 1 & -2 & 1 \\ -2 & -3-\imath & 1 \end{bmatrix},$$

abbiamo

$$|\mathbf{A}| = \begin{bmatrix} 1 & 2 & 3 \\ 1 & 2 & 1 \\ 2 & 3 & 2 \end{bmatrix}, \qquad |\mathbf{B}| = \begin{bmatrix} \sqrt{2} & 5 & 3 \\ 1 & 2 & 1 \\ 2 & 2 & 1 \end{bmatrix}.$$

[2] Per definire il determinante si usa a volte la stessa notazione. In ogni caso dovrebbe essere chiaro dal contesto se $|\mathbf{A}|$ indica la matrice dei valori assoluti o il suo determinante.

[3] Ricordiamo che se $\mathrm{z} = a + \imath b$ è un numero complesso allora

$$|z| = |a + \imath b| = \sqrt{a^2 + b^2}.$$

1.1.5 Norme di vettori

Dato un vettore $\mathbf{x} \in \mathbb{R}^3$ le sue tre componenti x_1, x_2, x_3 si possono interpretare come coordinate di un punto in $\mathbb{R}^3$. La distanza di questo punto dall'origine è la lunghezza del segmento che unisce l'origine con il punto $\mathbf{x}$. Questa lunghezza può essere interpretata come la lunghezza del vettore $\mathbf{x}$. Osserviamo che dal teorema di Pitagora[4] questa lunghezza è:

$$\sqrt{x_1^2 + x_2^2 + x_3^2}.$$

Possiamo generalizzare la nozione di lunghezza di un vettore in $\mathbb{R}^n$ o $\mathbb{C}^n$ come segue. Per ogni vettore $\mathbf{x} \in \mathbb{R}^n$ o $\mathbb{C}^n$ possiamo definire

$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2},$$

e chiamare questo numero "lunghezza del vettore". Questa funzione gode delle seguenti proprietà:

1. è una funzione non negativa, infatti $\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2} \geq 0$. Inoltre $\|\mathbf{x}\|_2 = 0$ solo se $x_i = 0$ per $i = 1, 2, \ldots, n$ cioè $\mathbf{x} = \mathbf{0}$;

2. "dilatando" o "contraendo" il vettore cioè moltiplicando ogni sua componente per una costante (rispettivamente maggiore o minore di 1) otteniamo:

$$\begin{aligned} \|\alpha\mathbf{x}\|_2 &= \sqrt{\sum_{i=1}^{n} |\alpha x_i|^2} = \sqrt{\sum_{i=1}^{n} |\alpha|^2 |x_i|^2} = |\alpha| \sqrt{\sum_{i=1}^{n} |x_i|^2} \\ &= |\alpha| \, \|\mathbf{x}\|_2; \end{aligned}$$

3. per ogni $\mathbf{x}, \mathbf{y} \in \mathbb{K}^n$ vale la seguente disuguaglianza:

$$\|\mathbf{x} + \mathbf{y}\|_2 \leq \|\mathbf{x}\|_2 + \|\mathbf{y}\|_2, \tag{1.1}$$

   detta *disuguaglianza triangolare*. Il nome disuguaglianza triangolare deriva dalla nota disuguaglianza sui lati dei triangoli. In particolare la lunghezza di un lato è sempre minore o uguale alla somma degli altri due. Questo fatto è schematizzato in figura 1.1.

[4] Pitagora 580a.c–500a.c.

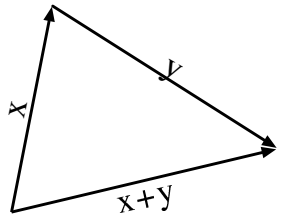


Figura 1.1: Disuguaglianza triangolare

Dimostrare la (1.1) è un po' laborioso e necessita della conoscenza di alcune disuguaglianze classiche. Iniziamo con la disuguaglianza di Young.

Lemma 1. *Dati due numeri reali p e q tali che*

$$\frac{1}{p}+\frac{1}{q}=1, \qquad 1<p,q<\infty$$

allora per ogni coppia di numeri reali non negativi a e b si ha

$$ab \le \frac{a^p}{p}+\frac{b^q}{q}, \tag{1.2}$$

Inoltre la disuguaglianza diventa uguaglianza se $a^p = b^q$.

Dimostrazione. Consideriamo la funzione

$$f(t)=\frac{t}{p}-t^{1/p},$$

allora

$$f'(t)=\frac{1}{p}-\frac{t^{1/p-1}}{p}=\frac{1}{p}-\frac{t^{-1/q}}{p}=\frac{1}{p}\left(1-t^{-1/q}\right),$$

poiché $1/q<1$ abbiamo che $f'(t)<0$ per $0<t<1$ e $f'(t)>0$ per $t>1$. Quindi $t=1$ è un punto di minimo per $f(t)$ in $(0,\infty)$ e di conseguenza $f(t)\ge f(1)$ per $t>0$. Quindi

$$f(t)\ge f(1), \qquad \Rightarrow \qquad \frac{t}{p}-t^{1/p}\ge\frac{1}{p}-1=-\frac{1}{q},$$

da cui

$$t^{1/p}\le\frac{1}{q}+\frac{t}{p}. \tag{1.3}$$

Osserviamo che se $a = 0$ o $b = 0$ la disuguaglianza è banalmente vera. Consideriamo quindi $a, b > 0$ e calcoliamo (1.3) in $t = a^p b^{-q}$ ottenendo:

$$ab^{-q/p} \leq \frac{1}{q} + \frac{a^p b^{-q}}{p},$$

moltiplicando la disuguaglianza per b^q ed osservando che $q - q/p = 1$ otteniamo il risultato cercato. Osserviamo che se $a^p = b^q$ allora calcoliamo la disuguaglianza (1.3) in $t = 1$ dove risulta essere una uguaglianza. ■

Possiamo ora dimostrare la disuguaglianza di Hölder.[5]

Teorema 2. *Dati due numeri reali p e q tali che $1 < p, q < \infty$ e $1/p + 1/q = 1$, ed $a_1, a_2, \ldots, a_n \geq 0$ e $b_1, b_2, \ldots, b_n \geq 0$, allora*

$$\sum_{k=1}^{n} a_k\, b_k \leq \left(\sum_{k=1}^{n} a_k^p\right)^{1/p} \left(\sum_{k=1}^{n} b_k^q\right)^{1/q}.$$

Dimostrazione. Siano

$$A = \left(\sum_{k=1}^{n} a_k^p\right)^{1/p}, \qquad B = \left(\sum_{k=1}^{n} b_k^q\right)^{1/q}.$$

Sia $AB = 0$. Ne consegue che o si ha $A = 0$ oppure $B = 0$ [6]. Supponiamo, per esempio, che sia $A = 0$. Ciò implica che $a_1 = a_2 = \cdots = a_n = 0$ e quindi la disuguaglianza è banalmente vera. Per $B = 0$ si ripete lo stesso ragionamento. Sia quindi $AB > 0$, dalla disuguaglianza (1.2) del lemma 1 otteniamo per ogni k

$$\frac{a_k}{A} \cdot \frac{b_k}{B} \leq \frac{a_k^p}{pA^p} + \frac{b_k^q}{qB^q},$$

cosicché

$$\frac{\displaystyle\sum_{k=1}^{n} a_k\, b_k}{AB} \leq \frac{\displaystyle\sum_{k=1}^{n} a_k^p}{pA^p} + \frac{\displaystyle\sum_{k=1}^{n} b_k^q}{qB^q} = \frac{A^p}{pA^p} + \frac{B^q}{qB^q} = \frac{1}{p} + \frac{1}{q} = 1.$$

Infine, dimostriamo la disuguaglianza di Minkowski.[7]

[5] Ludwig Otto Hölder 1859–1937.

[6] O entrambe contemporaneamente, ma a noi ne basta una sola per procedere!

[7] Hermann Minkowski 1864–1909.

Teorema 3. *Sia* $1 \leq p < \infty$*;* $a_1, a_2, \ldots, a_n \geq 0$ *e* $b_1, b_2, \ldots, b_n \geq 0$*. Allora*

$$\left(\sum_{k=1}^{n}(a_k + b_k)^p\right)^{1/p} \leq \left(\sum_{k=1}^{n} a_k^p\right)^{1/p} + \left(\sum_{k=1}^{n} b_k^p\right)^{1/p}.$$

Dimostrazione. Il caso $p = 1$ è banale. Supponiamo $p > 1$, allora

$$\sum_{k=1}^{n}(a_k + b_k)^p = \sum_{k=1}^{n} a_k(a_k + b_k)^{p-1} + \sum_{k=1}^{n} b_k(a_k + b_k)^{p-1}.$$

Applicando la disuguaglianza di Hölder ad ogni somma, usando q definito dalla $1/p + 1/q = 1$

$$\sum_{k=1}^{n}(a_k + b_k)^p \leq \left(\sum_{k=1}^{n} a_k^p\right)^{1/p}\left(\sum_{k=1}^{n}(a_k + b_k)^{q(p-1)}\right)^{1/q} + \left(\sum_{k=1}^{n} b_k^p\right)^{1/p}\left(\sum_{k=1}^{n}(a_k + b_k)^{q(p-1)}\right)^{1/q}.$$

Dividendo per $\left(\sum_{k=1}^{n}(a_k + b_k)^p\right)^{1/q}$ e osservando che $q(p-1) = p$ otteniamo la disuguaglianza cercata. ■

Osserviamo che la disuguaglianza di Minkowski per $p = 2$ è proprio la (1.1).

Possiamo generalizzare la nozione di lunghezza di un vettore tramite una generalizzazione della funzione $\|\cdot\|$ che conservi le tre proprietà precedenti.

Definizione 10. Una funzione $\|\cdot\| : \mathbb{K}^n \mapsto \mathbb{R}$ è una *norma* se per ogni $\mathbf{x}, \mathbf{y} \in \mathbb{K}^n$ e per ogni $\lambda \in \mathbb{K}$ verifica

1. $\|\mathbf{x}\| \geq 0$ e $\mathbf{x} = \mathbf{0} \Leftrightarrow \|\mathbf{x}\| = 0$,
2. $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$,
3. $\|\lambda\mathbf{x}\| = |\lambda|\,\|\mathbf{x}\|$.

Definizione 11. Utilizzando la disuguaglianza di Minkowski è facile dimostrare che per $1 \leq p < \infty$ la seguente funzione

$$\|\mathbf{x}\|_p = \left(\sum_{k=1}^{n} |x_k|^p\right)^{1/p},$$

è una norma. Tale norma è chiamata p-norma. Due casi di particolare interesse della $p-$norma si hanno per $p = 1, 2$

$$\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i|,$$

$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2}.$$

Con un procedimento di passaggio al limite si può anche definire

$$\|\mathbf{x}\|_\infty = \max_{k=1,n} |x_k|,$$

ed è immediato verificare che anche questa funzione è una norma.

Esempio 9. Dati i vettori

$$\mathbf{x} = \begin{bmatrix} 1 \\ 2 \\ -2 \end{bmatrix}, \qquad \mathbf{y} = \begin{bmatrix} 1+\imath \\ \imath \\ 1 \\ -1 \end{bmatrix},$$

otteniamo

$$\begin{aligned}
\|\mathbf{x}\|_\infty &= \max\{|1|, |2|, |-2|\} = \max\{1, 2, 2\} = 2, \\
\|\mathbf{y}\|_\infty &= \max\{|1+\imath|, |\imath|, |1|, |-1|\} = \max\left\{\sqrt{2}, 1, 1, 1\right\} = \sqrt{2}, \\
\|\mathbf{x}\|_1 &= |1| + |2| + |-2| = 1 + 2 + 2 = 5, \\
\|\mathbf{y}\|_1 &= |1+\imath| + |\imath| + |1| + |-1| = \sqrt{2} + 1 + 1 + 1 = 3 + \sqrt{2}, \\
\|\mathbf{x}\|_2 &= \sqrt{|1|^2 + |2|^2 + |-2|^2} = \sqrt{1+4+4} = 3, \\
\|\mathbf{y}\|_2 &= \sqrt{|1+\imath|^2 + |\imath|^2 + |1|^2 + |-1|^2} = \sqrt{2+1+1+1} = \sqrt{5}.
\end{aligned}$$

1.1.6 Prodotti scalari

Il prodotto scalare tra due vettori è molto usato in fisica ed ha la seguente definizione

$$\mathbf{a} \cdot \mathbf{b} = \|\mathbf{a}\|_2 \|\mathbf{b}\|_2 \cos\theta_{\mathbf{ab}}, \tag{1.4}$$

dove $\theta_{\mathbf{ab}}$ è l'angolo formato dai due vettori $\mathbf{a}$ e $\mathbf{b}$.

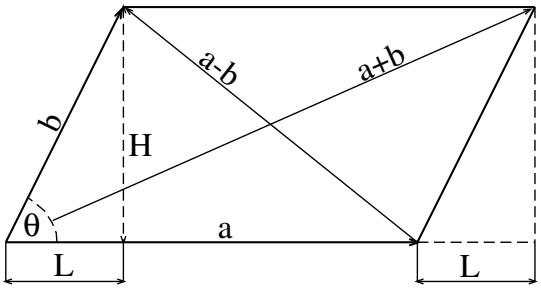


Figura 1.2: prodotto scalare e parallelogrammo associato

Dalla figura 1.2 [8]. possiamo ricavare una formula che non usa coseni ma solo la norma. Osserviamo innanzitutto che

$$L = \|\mathbf{b}\|_2 \cos\theta_{\mathbf{ab}}, \qquad H = \|\mathbf{b}\|_2 \sin\theta_{\mathbf{ab}},$$

e dal teorema di Pitagora

$$\|\mathbf{a}+\mathbf{b}\|_2^2 - H^2 = (\|\mathbf{a}\|_2 + L)^2 = \|\mathbf{a}\|_2^2 + L^2 + 2\|\mathbf{a}\|_2 L, \tag{1.5}$$

$$\|\mathbf{a}-\mathbf{b}\|_2^2 - H^2 = (\|\mathbf{a}\|_2 - L)^2 = \|\mathbf{a}\|_2^2 + L^2 - 2\|\mathbf{a}\|_2 L, \tag{1.6}$$

sottraendo la (1.6) dalla (1.5) otteniamo

$$\|\mathbf{a}+\mathbf{b}\|_2^2 - \|\mathbf{a}-\mathbf{b}\|_2^2 = 4\|\mathbf{a}\|_2 L = 4\|\mathbf{a}\|_2\|\mathbf{b}\|_2 \cos\theta_{\mathbf{ab}},$$

e quindi con la (1.4)

$$\mathbf{a}\cdot\mathbf{b} = \frac{\|\mathbf{a}+\mathbf{b}\|_2^2 - \|\mathbf{a}-\mathbf{b}\|_2^2}{4}.$$

Osserviamo che

$$\|\mathbf{a}+\mathbf{b}\|_2^2 - \|\mathbf{a}-\mathbf{b}\|_2^2 = \sum_{k=1}^{n}\left[|a_k+b_k|^2 - |a_k-b_k|^2\right] = 4\sum_{k=1}^{n} a_k\, b_k\,,$$

per cui il prodotto scalare prende la forma

$$\mathbf{a}\cdot\mathbf{b} = \sum_{k=1}^{n} a_k\, b_k. \tag{1.7}$$

[8] Nella figura l'angolo tra i vettori $\mathbf{a}$ e $\mathbf{b}$ è indicato col simbolo θ e non $\theta_{\mathbf{ab}}$ per ragioni esclusivamente tipografiche

Osserviamo che la formula che abbiamo ricavato vale per vettori reali e inoltre

$$\mathbf{a}\cdot\mathbf{a}=\sum_{k=1}^{n}a_k^2=\sum_{k=1}^{n}|a_k|^2=\|\mathbf{a}\|_2^2\,,$$

cioè il prodotto scalare di un vettore con se stesso restituisce il quadrato della sua lunghezza. Se $\mathbf{a}$ è un vettore complesso la formula (1.7) non restituisce il quadrato della sua lunghezza infatti se a_i sono numeri complessi:

$$\mathbf{a}\cdot\mathbf{a}=\sum_{k=1}^{n}a_k^2\neq\sum_{k=1}^{n}|a_k|^2\,.$$

Si può comunque modificare la definizione (1.7) in modo che applicata a vettori reali sia equivalente a (1.4) e nel caso di vettori complessi valga $\mathbf{a}\cdot\mathbf{a}=\|\mathbf{a}\|_2^2$.

Definizione 12. Consideriamo la funzione che dati due vettori di dimensione n restituisce un numero

$$\mathbf{x}\cdot\mathbf{y}=\sum_{i=1}^{n}x_i\overline{y_i},\tag{1.8}$$

dove con $\overline{\mathbf{z}}$ intendiamo l'operazione di coniugazione nel campo complesso[9] La (1.8) prende il nome di *prodotto scalare euclideo*.

[9] Ricordiamo che

$$\overline{a+\imath b}=a-\imath b$$

e la coniugazione gode delle seguenti proprietà: posto $z=a+\imath b$ e $w=c+\imath d$ abbiamo

$$\begin{aligned}z\overline{z}&=(a+\imath b)(a-\imath b)=a^2+b^2=|z|^2\,,\\ \overline{\overline{z}}&=\overline{\overline{a+\imath b}}=\overline{a-\imath b}=a+\imath b=z\\ \overline{z+w}&=\overline{z}+\overline{w},\end{aligned}$$

inoltre

$$\begin{aligned}\overline{zw}&=\overline{(a+\imath b)(c+\imath b)}=\overline{(ac-bd)+(bc+ad)\imath}=ac-bd-(bc+ad)\imath,\\ \overline{z}\;\overline{w}&=(\overline{a+\imath b})(\overline{c+\imath d})=(a-\imath b)(c-\imath d)=ac-bd-(bc+ad)\imath,\end{aligned}$$

da cui

$$\overline{zw}=\overline{z}\;\overline{w}.$$

Osserviamo che la funzione (1.8) ha le seguenti proprietà

① Calcolando $\mathbf{x} \cdot \mathbf{x}$ otteniamo

$$\mathbf{x} \cdot \mathbf{x} = \sum_{i=1}^{n} x_i \overline{x_i} = \sum_{i=1}^{n} |x_i|^2 \geq 0,$$

inoltre $\mathbf{x} \cdot \mathbf{x} = 0$ se e solo se $x_i = 0$ con $i = 1, 2, \ldots, n$ e quindi $\mathbf{x} = \mathbf{0}$.

② Calcolando $\mathbf{x} \cdot \mathbf{y}$ e tenendo conto del fatto che $\overline{\overline{z}} = z$

$$\mathbf{x} \cdot \mathbf{y} = \sum_{i=1}^{n} x_i \overline{y_i}, = \sum_{i=1}^{n} \overline{\overline{x_i} y_i}, = \overline{\sum_{i=1}^{n} \overline{x_i} y_i}, = \overline{\mathbf{y} \cdot \mathbf{x}}.$$

③ Calcolando $(\mathbf{x} + \mathbf{y}) \cdot \mathbf{z}$ otteniamo

$$(\mathbf{x} + \mathbf{y}) \cdot \mathbf{z} = \sum_{i=1}^{n} (x_i + y_i) \overline{z_i} = \sum_{i=1}^{n} x_i \overline{z_i} + \sum_{i=1}^{n} y_i \overline{z_i} = \mathbf{x} \cdot \mathbf{z} + \mathbf{y} \cdot \mathbf{z}.$$

④ Calcolando $(\alpha \mathbf{x}) \cdot \mathbf{y}$ otteniamo

$$(\alpha \mathbf{x}) \cdot \mathbf{y} = \sum_{i=1}^{n} \alpha x_i \overline{y_i} = \alpha \sum_{i=1}^{n} x_i \overline{y_i} = \alpha \mathbf{x} \cdot \mathbf{y}.$$

Le proprietà ①–④ si possono utilizzare per dare la definizione di prodotto scalare in maniera assiomatica.

Per $z = a + \imath b \in \mathbb{C}$ abbiamo le seguenti funzioni

$$Re\,\{z\} = \frac{z + \overline{z}}{2} = a, \qquad Im\,\{z\} = \frac{z - \overline{z}}{2} = b.$$

Definizione 13. Una funzione $(\cdot,\cdot) : \mathbb{K}^n \times \mathbb{K}^n \mapsto \mathbb{K}$ è un *prodotto scalare* se per ogni $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{K}^n$ e per ogni $\alpha \in \mathbb{K}$ soddisfa:

1. $(\mathbf{x}, \mathbf{x}) \geq 0$ e $(\mathbf{x}, \mathbf{x}) = 0$ per ogni $\mathbf{x}$ se e solo se $\mathbf{x} = \mathbf{0}$;
2. $(\mathbf{x}, \mathbf{y}) = \overline{(\mathbf{y}, \mathbf{x})}$;
3. $(\mathbf{x} + \mathbf{y}, \mathbf{z}) = (\mathbf{x}, \mathbf{z}) + (\mathbf{y}, \mathbf{z})$;
4. $(\alpha\mathbf{x}, \mathbf{y}) = \alpha(\mathbf{x}, \mathbf{y})$.

Osservazione 1. Nella definizione assiomatica appena enunciata, si è indicato il prodotto scalare tra due vettori $\mathbf{x}, \mathbf{y} \in \mathbb{K}^n$ utilizzando la simbologia $(\mathbf{x}, \mathbf{y})$, mentre nella discussione precedente il prodotto scalare era stato indicato con $\mathbf{x} \cdot \mathbf{y}$. Ovviamente ne la definizione del prodotto scalare ne le sue proprietà dipendono dalla notazione scelta. Esiste una terza notazione di uso abituale e cioè $\mathbf{x}^T\mathbf{y}$. Questa notazione assume implicitamente che tutti i vettori siano vettori colonna ed indica con l'apice ^{T} l'operazione di trasposizione che trasforma un vettore colonna in un vettore riga. La definizione generale di "trasposto" di una matrice e di un vettore, di cui abbiamo anticipato l'idea, sarà introdotta tra poco.

Esempio 10. E' facile verificare che anche la seguente funzione

$$[\mathbf{x}, \mathbf{y}] = \sum_{k=1}^{n} k x_k \overline{y_k},$$

definisce un prodotto scalare per ogni k reale positivo.

Osservazione 2. Notiamo che la funzione $\|\cdot\|_2$ si può esprimere tramite il prodotto scalare euclideo "$\cdot$" come segue

$$\|\mathbf{x}\|_2 = \sqrt{\mathbf{x} \cdot \mathbf{x}}.$$

Analogamente, dato un prodotto scalare generico $(\cdot,\cdot)$, si puó sempre definire l'applicazione

$$\|\mathbf{x}\| = \sqrt{(\mathbf{x}, \mathbf{x})},$$

che ha le proprietà di una norma. Questa applicazione prende il nome di *norma indotta dal prodotto scalare*.

Teorema 4 (Disuguaglianza di Cauchy-Schwarz). *Per un prodotto scalare generico vale la disuguaglianza di Cauchy*[10]*-Schwarz*[11]

$$\boxed{|(\mathbf{x},\mathbf{y})| \leq \|\mathbf{x}\|\,\|\mathbf{y}\|} \tag{1.9}$$

dove $\|\cdot\|$ è la norma indotta dal prodotto scalare e la disuguaglianza è stretta a meno che $\mathbf{x} = \gamma\mathbf{y}$ per uno scalare γ, cioè i vettori sono allineati.

Dimostrazione. Se $\mathbf{x} = \mathbf{0}$ o $\mathbf{y} = \mathbf{0}$ la disuguaglianza è banale. Supponiamo quindi che entrambi i vettori siano non nulli. Applicando la proprietà ① della definizione 13 al vettore $\mathbf{x} - \alpha\mathbf{y}$ otteniamo

$$(\mathbf{x} - \alpha\mathbf{y}, \mathbf{x} - \alpha\mathbf{y}) \geq 0,$$

da cui segue

$$\begin{aligned} 0 &\leq (\mathbf{x} - \alpha\mathbf{y}, \mathbf{x} - \alpha\mathbf{y}), \\ &\leq (\mathbf{x},\mathbf{x}) - \alpha(\mathbf{y},\mathbf{x}) - \overline{\alpha}(\mathbf{x},\mathbf{y}) + \alpha\overline{\alpha}(\mathbf{y},\mathbf{y}), \\ &= (\mathbf{x},\mathbf{x}) - \alpha\overline{(\mathbf{x},\mathbf{y})} - \overline{\alpha}\left[(\mathbf{x},\mathbf{y}) - \alpha\,(\mathbf{y},\mathbf{y})\right]. \end{aligned} \tag{1.10}$$

Scegliendo α in modo da annullare l'espressione tra parentesi quadre,

$$\alpha = \frac{(\mathbf{x},\mathbf{y})}{(\mathbf{y},\mathbf{y})},$$

otteniamo

$$0 \leq \|\mathbf{x}\|^2 - \frac{|(\mathbf{x},\mathbf{y})|^2}{\|\mathbf{y}\|^2},$$

che è equivalente alla (1.9). Osserviamo che se $\mathbf{x} - \alpha\mathbf{y} \neq \mathbf{0}$ allora la disuguaglianza in (1.10) è stretta e di conseguenza anche la (1.9). ■

[10] Augustin Louis Cauchy 1789–1857.
[11] Karl Herman Amandus Schwarz 1843–1921.

1.1.7 Ortogonalità e angolo tra vettori

Tramite il concetto di prodotto scalare è possibile introdurre il concetto di ortogonalità e angolo tra vettori. Dal prodotto scalare euclideo (1.8) nella forma (1.4) otteniamo che l'angolo formato tra due vettori $\mathbf{a}$ e $\mathbf{b}$ è dato dalla seguente formula

$$\theta_{\mathbf{ab}} = \arccos \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\|_2 \|\mathbf{b}\|_2}.$$

Esempio 11. I vettori $\mathbf{a}$ e $\mathbf{b}$ definiti come segue

$$\mathbf{a} = \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix},$$

formano un angolo di circa 30^0 o circa 0.5236 radianti, infatti

$$\cos \theta_{\mathbf{ab}} = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\|_2 \|\mathbf{b}\|_2} = \frac{3}{\sqrt{12}}.$$

Se l'angolo tra i vettori è 90^0 allora $\cos 90^0 = 0$ implica che il loro prodotto scalare è nullo. Questo suggerisce la seguente definizione.

Definizione 14. Dati due vettori $\mathbf{a}$ e $\mathbf{b}$ diremo che $\mathbf{a}$ e $\mathbf{b}$ sono ortogonali e scriveremo $\mathbf{a} \perp \mathbf{b}$ quando il loro prodotto scalare è nullo, cioè $\mathbf{a} \cdot \mathbf{b} = 0$.

Osservazione 3. Mentre la definizione di angolo tra vettori è valida solo per vettori reali, la definizione di ortogonalità è valida anche per vettori complessi e prodotti scalare qualunque. Infatti la definizione è puramente algebrica.

Esempio 12. I vettori

$$\mathbf{a} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix},$$

sono ortogonali (cioè $\mathbf{a} \perp \mathbf{b}$) infatti

$$\mathbf{a} \cdot \mathbf{b} = 1 \cdot 1 + 2 \cdot (-2) + 3 \cdot 1 = 0,$$

analogamente i vettori

$$\mathbf{x} = \begin{bmatrix} 1+\imath \\ 1 \\ 1 \end{bmatrix}, \qquad \mathbf{y} = \begin{bmatrix} 1+\imath \\ -1 \\ -1 \end{bmatrix},$$

sono ortogonali infatti

$$\mathbf{x} \cdot \mathbf{y} = (1+\imath)(\overline{1+\imath}) + 1 \cdot (-1) + 1 \cdot (-1) = 0.$$

Prodotto vettoriale

Dati due vettori $\mathbf{a}$ e $\mathbf{b}$ in $\mathbb{R}^3$ ci poniamo il problema di trovare un terzo vettore $\mathbf{x}$ ortogonale ad entrambi. Algebricamente il problema diventa:

$$\begin{cases} \text{trovare } \mathbf{x} \in \mathbb{R}^3 \text{ tale che} \\ \mathbf{a} \cdot \mathbf{x} = 0, \\ \mathbf{b} \cdot \mathbf{x} = 0, \end{cases}$$

che scritto usando le componenti dei vettori si esprime come

$$\begin{cases} a_1 x_1 + a_2 x_2 + a_3 x_3 = 0, \\ b_1 x_1 + b_2 x_2 + b_3 x_3 = 0. \end{cases}$$

Una possibile soluzione, che si può verificare per sostituzione diretta, è data da

$$\begin{aligned} x_1 &= a_2 b_3 - a_3 b_2, \\ x_2 &= a_3 b_1 - a_1 b_3, \\ x_3 &= a_1 b_2 - a_2 b_1. \end{aligned}$$

Questa soluzione prende il nome di *prodotto vettoriale* e si indica normalmente con l'espressione

$$\mathbf{x} = \mathbf{a} \wedge \mathbf{b}.$$

Si può anche verificare che

$$(\|\mathbf{a} \wedge \mathbf{b}\|_2)^2 + (\mathbf{a} \cdot \mathbf{b})^2 = (\|\mathbf{a}\|_2)^2 (\|\mathbf{b}\|_2)^2, \tag{1.11}$$

e dalla (1.4) tramite (1.11) ottenere

$$\|\mathbf{a} \wedge \mathbf{b}\|_2 = \|\mathbf{a}\|_2 \|\mathbf{b}\|_2 \sin \theta_{\mathbf{ab}}.$$

Per mezzo del prodotto vettoriale è facile risolvere alcuni problemi di geometria nello spazio, come ad esempio il calcolo del piano passante per 3 punti. Siano infatti $\mathbf{a}$, $\mathbf{b}$ e $\mathbf{c}$ tre punti distinti, allora i vettori

$$\mathbf{v} = \mathbf{b} - \mathbf{a}, \qquad \mathbf{w} = \mathbf{c} - \mathbf{a},$$

sono vettori complanari al piano, il vettore normale al piano $\mathbf{N}$ diventa semplicemente

$$\mathbf{N} = \mathbf{v} \wedge \mathbf{w},$$

e l'equazione del piano

$$\mathbf{N} \cdot \mathbf{x} = \mathbf{N} \cdot \mathbf{a}.$$

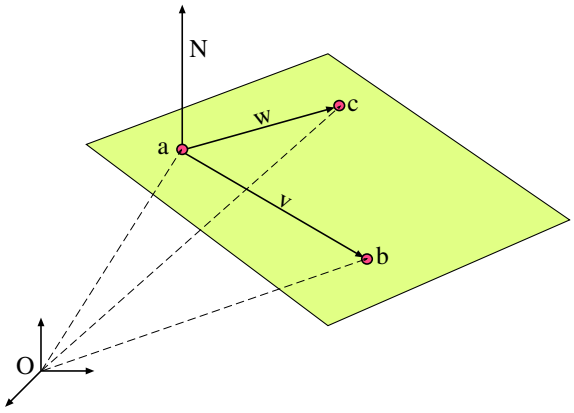


Figura 1.3: piano per 3 punti

Esempio 13. Dati i punti

$$\mathbf{a} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} 0 \\ 2 \\ 1 \end{bmatrix}, \qquad \mathbf{c} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix},$$

trovare il piano passante per $\mathbf{a}$, $\mathbf{b}$ e $\mathbf{c}$. Calcoliamo innanzitutto

$$\mathbf{v} = \mathbf{b} - \mathbf{a} = \begin{bmatrix} 0 \\ 2 \\ 1 \end{bmatrix} - \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} -1 \\ 2 \\ 0 \end{bmatrix}, \qquad \mathbf{w} = \mathbf{c} - \mathbf{a} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} - \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 2 \\ 2 \end{bmatrix},$$

da cui

$$\mathbf{N} = \mathbf{v} \wedge \mathbf{w} = \begin{bmatrix} -1 \\ 2 \\ 0 \end{bmatrix} \wedge \begin{bmatrix} 0 \\ 2 \\ 2 \end{bmatrix} = \begin{bmatrix} 4 \\ 2 \\ -2 \end{bmatrix},$$

e infine

$$\mathbf{N} \cdot \mathbf{x} = 4x_1 + 2x_2 - 2x_3, \qquad \mathbf{N} \cdot \mathbf{a} = 2,$$

ponendo $\mathbf{x} = [x, y, z]^T$ otteniamo l'equazione del piano

$$2x + y - z = 1.$$

1.1.8 Indipendenza lineare e basi in $\mathbb{K}^n$

Il concetto di dipendenza ed indipendenza lineare è estremamente importante nell'algebra lineare.

Definizione 15. Dati k vettori non nulli $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$ se esistono k scalari $\alpha_1, \alpha_2, \ldots, \alpha_k$, con almeno uno scalare non nullo, per cui vale

$$\alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2 + \cdots + \alpha_k\mathbf{x}_k = 0,$$

allora diremo che $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$ sono vettori *linearmente dipendenti*, viceversa se tali scalari non esistono diremo che $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$ sono vettori *linearmente indipendenti*.

Consideriamo k vettori *linearmente indipendenti* di $\mathbb{K}^n$, che indicheremo con $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$. Dato un vettore $\mathbf{w} \in \mathbb{K}^n$, può succedere che questi sia combinazione lineare dei precedenti, per cui dalla definizione si può scrivere

$$\mathbf{w} = \alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2 + \cdots + \alpha_k\mathbf{x}_k,$$

con una opportuna scelta degli scalari α_i.

Definizione 16. Se questa proprietà è vera *per ogni* vettore di $\mathbb{K}^n$, allora diremo che i vettori $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$ formano *una base*. Necessariamente, si deve avere $k = n$. Questa condizione è anche sufficiente, nel senso che scelti n vettori qualunque di $\mathbb{K}^n$, purché linearmente indipendenti, essi formano sempre una base di $\mathbb{K}^n$.

Esempio 14. Gli n vettori $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$ in $\mathbb{K}^n$

$$\mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \qquad \mathbf{e}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \qquad \cdots \qquad \mathbf{e}_n = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix},$$

sono ovviamente linearmente indipendenti. Infatti,

$$\alpha_1\mathbf{e}_1 + \alpha_2\mathbf{e}_2 + \cdots + \alpha_n\mathbf{e}_n = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix},$$

e la combinazione lineare è nulla se e solo se $\alpha_i = 0$ per $i = 1, 2, \ldots, n$.

Esempio 15. I vettori $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$ definiti nell'esempio 14 formano una base in $\mathbb{K}^n$. Dato un vettore $\mathbf{w}$ qualsiasi di componenti $w_1, w_2, \ldots, w_n$, possiamo scrivere

$$\mathbf{w} = w_1\mathbf{e}_1 + w_2\mathbf{e}_2 + \cdots + w_n\mathbf{e}_n,$$

La base $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$ è detta *base canonica*.

Esempio 16. I vettori

$$\mathbf{a} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix}, \qquad \mathbf{c} = \begin{bmatrix} -1 \\ 6 \\ 1 \end{bmatrix},$$

non sono linearmente indipendenti, infatti si verifica immediatamente che

$$\mathbf{a} - 2\mathbf{b} + \mathbf{c} = \mathbf{0}.$$

Teorema 5. *I k vettori $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$ a due a due ortogonali,*

$$\mathbf{x}_i \perp \mathbf{x}_j \qquad i \neq j,$$

sono necessariamente linearmente indipendenti.

Dimostrazione. Supponiamo che esista una scelta di k scalari α_i, tali che

$$\alpha_1 \mathbf{x}_1 + \alpha_2 \mathbf{x}_2 + \cdots + \alpha_k \mathbf{x}_k = 0.$$

Facendo il prodotto scalare con ogni vettore $\mathbf{x}_i$ per $i = 1, 2, \ldots, k$ e tenendo conto delle relazioni di ortogonalità si ottiene

$$\begin{aligned} 0 &= \mathbf{x}_i \cdot (\alpha_1 \mathbf{x}_1 + \alpha_2 \mathbf{x}_2 + \cdots + \alpha_k \mathbf{x}_k), \\ &= \alpha_1 \mathbf{x}_i \cdot \mathbf{x}_1 + \alpha_2 \mathbf{x}_i \cdot \mathbf{x}_2 + \cdots + \alpha_i \mathbf{x}_i \cdot \mathbf{x}_i + \cdots + \alpha_k \mathbf{x}_i \cdot \mathbf{x}_k, \\ &= \alpha_i \mathbf{x}_i \cdot \mathbf{x}_i, \end{aligned}$$

e quindi poiché $\mathbf{x}_i \cdot \mathbf{x}_i > 0$ segue che $\alpha_i = 0$. ■

1.1.9 Ortonormalizzazione di Gram-Schmidt

Definizione 17. Dati k vettori $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$, diremo che gli stessi formano un *sistema ortogonale* se sono a due a due ortogonali, cioè

$$\mathbf{v}_i \perp \mathbf{v}_j, \qquad i \neq j.$$

Definizione 18. Dati k vettori $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k$, diremo che gli stessi formano un *sistema ortonormale* se sono a due a due ortogonali e di norma 1, cioè

$$\|\mathbf{u}_i\|_2 = 1, \qquad \mathbf{u}_i \perp \mathbf{u}_j, \quad i \neq j.$$

Definizione 19. Dati k vettori $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$, definiremo con $\operatorname{span}\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\}$ lo spazio vettoriale generato dalle loro combinazioni lineari

$$\operatorname{span}\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\} = \{\alpha_1 \mathbf{v}_1 + \alpha_2 \mathbf{v}_2 + \cdots + \alpha_k \mathbf{v}_k \mid \alpha_1, \alpha_2, \ldots, \alpha_k \in \mathbb{K}\}$$

Dati k vettori $\mathbf{v}_1$, $\mathbf{v}_2$,..., $\mathbf{v}_k$, linearmente indipendenti è possibile costruire k vettori $\mathbf{u}_1$, $\mathbf{u}_2$,..., $\mathbf{u}_k$ a due a due ortogonali e di norma unitaria tali che

$$\text{span}\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\} = \text{span}\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\}.$$

Teorema 6 (Ortonormalizzazione di Gram-Schmidt). [12] *Dati k vettori $\mathbf{v}_1$, $\mathbf{v}_2$,..., $\mathbf{v}_k$, linearmente indipendenti possiamo costruire k vettori $\mathbf{u}_1$, $\mathbf{u}_2$,..., $\mathbf{u}_k$, con le seguenti proprietà:*

1. $\mathbf{u}_1 = \dfrac{\mathbf{v}_1}{\|\mathbf{v}_1\|_2}$;
2. $\mathbf{u}_i \perp \mathbf{u}_j$ *per ogni* $i \neq j$;
3. $\|\mathbf{u}_i\|_2 = 1$ *per* $i = 1, 2, \ldots, k$;
4. $V_j = U_j$ *per* $j = 1, 2, \ldots, k$ *dove* $V_j = \text{span}\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_j\}$ *e* $U_j = \text{span}\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_j\}$.

Dimostrazione. Osserviamo innanzitutto che dall'indipendenza lineare dei vettori $\mathbf{v}_1$, $\mathbf{v}_2$,..., $\mathbf{v}_k$ segue che $\|\mathbf{v}_i\|_2 \neq 0$ per ogni $i = 1, 2, \ldots, k$. Quindi è sempre possibile scegliere $\mathbf{u}_1 = \dfrac{\mathbf{v}_1}{\|\mathbf{v}_1\|_2}$. La dimostrazione procede per induzione.

- Se $k = 1$ il teorema è ovviamente vero con $\mathbf{u}_1 = \dfrac{\mathbf{v}_1}{\|\mathbf{v}_1\|_2}$.
- Se $k > 1$, assumiamo come *ipotesi induttiva* di avere già determinato $k - 1$ vettori ortonormali $\mathbf{u}_1$, $\mathbf{u}_2$,..., $\mathbf{u}_{k-1}$ tali che

$$\mathbf{u}_1 = \frac{\mathbf{v}_1}{\|\mathbf{v}_1\|_2}, \qquad V_j = U_j, \quad j = 1, 2, \ldots, k-1.$$

Definiamo quindi un vettore ausiliario $\mathbf{w}_k$ ed il vettore $\mathbf{u}_k$ come segue

$$\mathbf{w}_k = \mathbf{v}_k - \sum_{i=1}^{k-1} \beta_i \mathbf{u}_i, \tag{1.12a}$$

$$\mathbf{u}_k = \alpha \mathbf{w}_k, \tag{1.12b}$$

dove dovremo determinare i k coefficienti α e β_i in modo che $\mathbf{u}_k \perp \mathbf{u}_i$ per $i = 1, 2, \ldots, k-1$ e $\|\mathbf{u}_k\|_2 = 1$. Facendo il prodotto scalare di (1.12a) con $\mathbf{u}_k$ ed utilizzando la proprietà di ortogonalità tra i vettori, che vale per ipotesi induttiva, abbiamo l'espressione

$$\mathbf{w}_k \cdot \mathbf{u}_j = \mathbf{v}_k \cdot \mathbf{u}_j - \sum_{i=1}^{k-1} \beta_i \mathbf{u}_i \cdot \mathbf{u}_j = \mathbf{v}_k \cdot \mathbf{u}_j - \beta_j, \qquad j = 1, 2, \ldots, k-1$$

[12] Jorgen Pedersen Gram 1850–1916.
Erhard Schmidt 1876–1959.

ed imponendo $\mathbf{w}_k \cdot \mathbf{u}_j = 0$ si ottiene

$$\beta_j = \mathbf{v}_k \cdot \mathbf{u}_j, \qquad j = 1, 2, \ldots, k-1.$$

Per determinare α imponiamo che $\|\mathbf{u}_k\|_2 = 1$ come segue

$$1 = \|\mathbf{u}_k\|_2^2 = \mathbf{u}_k \cdot \mathbf{u}_k = \alpha^2 \mathbf{w}_k \cdot \mathbf{w}_k = \alpha^2 \|\mathbf{w}_k\|_2^2,$$

da cui risulta $\alpha = 1/\|\mathbf{w}_k\|_2$; ovviamente deve essere $\mathbf{w}_k \neq \mathbf{0}$. Ragionando per assurdo se fosse $\mathbf{w}_k = \mathbf{0}$, allora si avrebbe

$$\mathbf{0} = \mathbf{v}_k - \sum_{i=1}^{k-1} \beta_i \mathbf{u}_i,$$

e poichè per l'ipotesi induttiva $U_{k-1} = V_{k-1}$ esisterebbero $k-1$ scalari γ_1, γ_2,..., γ_{k-1} per cui

$$\mathbf{v}_k = \sum_{i=1}^{k-1} \beta_i \mathbf{u}_i = \sum_{i=1}^{k-1} \gamma_i \mathbf{v}_i,$$

contraddicendo l'indipendenza lineare dei vettori $\mathbf{v}_i$, assunta nell'enunciato del teorema.

Per concludere la dimostrazione, bisogna ancora verificare che $U_k = V_k$.

Consideriamo una qualsiasi combinazione lineare dei vettori $\mathbf{v}_i$, cioè un generico vettore $\mathbf{z} \in V_k$, che scriveremo quindi come

$$\mathbf{z} = \sum_{i=1}^{k} \eta_i \mathbf{v}_i, \tag{1.13}$$

e mostriamo che $\mathbf{z} \in U_k$. Per l'ipotesi induttiva esistono $k-1$ scalari ζ_1, ζ_2,..., ζ_{k-1} tali che

$$\sum_{i=1}^{k-1} \eta_i \mathbf{v}_i = \sum_{i=1}^{k-1} \zeta_i \mathbf{u}_i. \tag{1.14}$$

Dalla (1.12a) possiamo scrivere

$$\mathbf{v}_k = \frac{\mathbf{u}_k}{\alpha} + \sum_{i=1}^{k-1} \beta_i \mathbf{u}_i, \tag{1.15}$$

e con la (1.14) e (1.13) abbiamo

$$\mathbf{z} = \frac{\eta_k}{\alpha}\mathbf{u}_k + \eta_k \sum_{i=1}^{k-1} \beta_i \mathbf{u}_i + \sum_{i=1}^{k-1} \zeta_i \mathbf{u}_i = \frac{\eta_k}{\alpha}\mathbf{u}_k + \sum_{i=1}^{k-1} (\zeta_i + \eta_k \beta_i)\, \mathbf{u}_i.$$

Quindi $\mathbf{z} \in U_k$ e poichè $\mathbf{z}$ è arbitrario segue che

$$V_k \subset U_k.$$

Viceversa sia $\mathbf{z} \in U_k$ allora

$$\begin{aligned} \mathbf{z} &= \sum_{i=1}^{k} \zeta_i \mathbf{u}_i, \\ &= \zeta_k \alpha \left(\mathbf{v}_k - \sum_{i=1}^{k-1} \beta_i \mathbf{u}_i \right) + \sum_{i=1}^{k-1} \zeta_i \mathbf{u}_i, \\ &= \zeta_k \alpha \mathbf{v}_k - \sum_{i=1}^{k-1} (\zeta_i - \zeta_k \alpha \beta_i)\, \mathbf{u}_i. \end{aligned}$$

Per l'ipotesi induttiva esistono $k-1$ scalari $\omega_1, \omega_2, \ldots, \omega_{k-1}$ tali che

$$\sum_{i=1}^{k-1} (\zeta_i - \zeta_k \alpha \beta_i)\, \mathbf{u}_i = \sum_{i=1}^{k-1} \omega_i \mathbf{v}_i,$$

e quindi $\mathbf{z} \in V_k$. Poichè $\mathbf{z}$ è arbitrario, $U_k \subset V_k$, che con la precedente inclusione $V_k \subset U_k$, permette di concludere che $U_k = V_k$. ∎

Questo teorema suggerisce il seguente algoritmo:

Algorithm *Ortonormalizzazione di Gram-Schmidt*
Input: k vettori $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$ linearmente indipendenti.
1. $\mathbf{u}_1 \leftarrow \mathbf{v}_1 / \|\mathbf{v}_1\|_2$
2. **for** $k \leftarrow 2$ **to** k
3. **do** $\mathbf{w}_k \leftarrow \mathbf{v}_k - \sum_{i=1}^{k-1} (\mathbf{v}_k \cdot \mathbf{u}_i)\, \mathbf{u}_i$
4. $\mathbf{u}_k \leftarrow \mathbf{w}_k / \|\mathbf{w}_k\|_2$
5. ($*$ I vettori $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k$ sono ortonormali. $*$)

Osservazione 4. Se indichiamo con

$$\mathbf{Q} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k]$$

la matrice rettangolare in $\mathbb{R}^{m\times k}$ le cui colonne sono i vettori colonna $\mathbf{u}_j$ prodotti dal procedimento di Gram-Schmidt, l'ortonormalità tra i vettori si può esprimere con

$$\mathbf{Q}^T\mathbf{Q} = \mathbf{I} \in \mathbb{R}^{k\times k}$$

Tuttavia si osservi che se $k < m$, non si ha l'ortogonalità tra le righe, cioè non vale $\mathbf{Q}\mathbf{Q}^T = \mathbf{I} \in \mathbb{R}^{m\times m}$. In realtà si può osservare che per la matrice prodotto $\mathbf{P} = \mathbf{Q}\mathbf{Q}^T$ valgono le proprietà seguenti:

$$\mathbf{P}^2 = \mathbf{P},$$

$$\mathbf{P}^T = \mathbf{P}.$$

Quindi, possiamo concludere che la matrice prodotto $\mathbf{Q}\mathbf{Q}^T$ è un *proiettore ortogonale*.

Osservazione 5. Dato un insieme di k vettori linearmente indipendenti $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\}$ in $\mathbb{K}^n$ a due a due ortonormali, è sempre possibile trovare altri $n-k$ vettori $\{\mathbf{u}_{k+1}, \mathbf{u}_{k+2}, \ldots, \mathbf{u}_n\}$ in modo che $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$ sia una base ortonormale. Infatti data una qualunque base basta togliere ad essa i vettori linearmente dipendenti da $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k$. I rimanenti, uniti ai k vettori di partenza, formano una base. Utilizzando il procedimento di ortonormalizzazione otteniamo una base ortonormale. E' facile verificare che da questo processo i vettori $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k$ non vengono modificati.

1.1.10 Operazioni con le matrici

Definizione 20 (Prodotto). Date le matrici $\mathbf{A} \in \mathbb{K}^{m\times n}$ e $\mathbf{B} \in \mathbb{K}^{n\times p}$ si definisce con $\mathbf{C} = \mathbf{A}\mathbf{B}$ la matrice $\mathbb{K}^{m\times p}$ prodotto di $\mathbf{A}$ e $\mathbf{B}$ nella quale

$$C_{ij} = \sum_{k=1}^{n} A_{ik}B_{kj}, \qquad i = 1, 2, \ldots, m, \quad j = 1, 2, \ldots, p.$$

Osserviamo che C_{ij} è il "prodotto scalare" della i-esima riga di $\mathbf{A}$ con la j-esima colonna di $\mathbf{B}$.

Esempio 17. Date le matrici 2×3 e 3×1

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ -2 & 2 & 0 \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix},$$

otteniamo

$$\mathbf{AB} = \begin{bmatrix} 1 & 2 & 3 \\ -2 & 2 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ -2 \end{bmatrix}.$$

Osservazione 6. Data una coppia di matrici $\mathbf{A}$ e $\mathbf{B}$ qualsiasi, non è detto che si possano sempre moltiplicare tra loro. Infatti, esse devono essere *compatibili*, cioè il prodotto $\mathbf{AB}$ è definito solo se $\mathbf{A}$ ha tante colonne quante sono le righe di $\mathbf{B}$. Analogamente è richiesto per il prodotto $\mathbf{BA}$. Si noti che potrebbe essere definito il prodotto $\mathbf{AB}$ ma non $\mathbf{BA}$ o viceversa, ed inoltre le matrici prodotto $\mathbf{AB}$ e $\mathbf{BA}$ se definite entrambe possono tuttavia avere dimensioni differenti, quindi un diverso numero di righe e colonne. Nel seguito, ogni volta che si parla di prodotto di matrici, anche se non esplicitamente dichiarato, si intenderà sempre che si tratta di matrici compatibili.

Esempio 18. Esemplifichiamo il fatto che la moltiplicazione di matrici non è commutativa, considerando le due seguenti matrici 2×2

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 1 & -1 \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}.$$

Possiamo definire sia il prodotto $\mathbf{AB}$ che il prodotto $\mathbf{BA}$, ma si vede che $\mathbf{AB} \neq \mathbf{BA}$. Infatti,

$$\mathbf{AB} = \begin{bmatrix} 2 & 3 \\ -1 & 0 \end{bmatrix}, \qquad \mathbf{BA} = \begin{bmatrix} 1 & -1 \\ 2 & 1 \end{bmatrix}.$$

Esempio 19. Dati i vettori riga e colonna

$$\mathbf{a} = \begin{bmatrix} 1 & 2 & 3 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix},$$

abbiamo che $\mathbf{ab}$ è una matrice 1×1 identificabile con un numero o scalare;

$$\mathbf{ab} = \begin{bmatrix} 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} = 1 \cdot 1 + 2 \cdot 0 + 3 \cdot 1 = 4,$$

mentre $\mathbf{ba}$ è una matrice 3×3

$$\mathbf{ba} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 0 & 0 \\ 1 & 2 & 3 \end{bmatrix}.$$

Osservazione 7. Non vale per le matrici la regola di annullamento, nel senso che se $\mathbf{A}$ e $\mathbf{B}$ sono due matrici e $\mathbf{AB} = \mathbf{0}$ non è detto che $\mathbf{A} = \mathbf{0}$ o $\mathbf{B} = \mathbf{0}$. Infatti basta considerare il seguente esempio

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

per il quale abbiamo

$$\mathbf{AB} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \qquad \mathbf{BA} = \begin{bmatrix} 0 & 0 \\ 2 & 0 \end{bmatrix}.$$

Osserviamo che in $\mathbf{AB} \neq \mathbf{BA}$ conferma la non commutatività del prodotto di matrici nel caso piú generale.

Esempio 20. Nell'insieme delle matrici la matrice identità $\mathbf{I}$ rappresenta l'elemento neutro della moltiplicazione. Infatti sia $\mathbf{A} \in \mathbb{K}^{m \times n}$ e $\mathbf{I} \in \mathbb{K}^{m \times m}$ la matrice identità allora

$$\mathbf{AI} = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & \ddots & \ddots & \vdots \\ \vdots & \ddots & A_{m-1n-1} & A_{m-1n} \\ A_{m1} & \cdots & A_{mn-1} & A_{mn} \end{bmatrix} \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & 1 & 0 \\ 0 & \cdots & 0 & 1 \end{bmatrix},$$

osserviamo che $I_{ij} = \delta_{i,j}$ dove

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

è il simbolo di Kroneker[13], e quindi

$$(\mathbf{AI})_{ij} = \sum_{k=1}^{n} A_{ik} I_{kj} = \sum_{k=1}^{n} A_{ik}\delta_{k,j} = A_{ij}\delta_{j,j} = A_{ij},$$

in modo del tutto analogo si prova che $\mathbf{IA} = \mathbf{A}$ dove questa volta $\mathbf{I} \in \mathbb{K}^{m\times m}$.

Definizione 21 (Inversa). Si definisce matrice inversa di una matrice *quadrata* $\mathbf{A}$ la matrice quadrata $\mathbf{B}$ (se esiste) che soddisfa $\mathbf{AB} = \mathbf{BA} = \mathbf{I}$. La matrice inversa se esiste si denota con $\mathbf{A}^{-1}$. La matrice $\mathbf{A}$ di dice invertibile.

Esempio 21. Data la matrice $\mathbf{A} \in \mathbb{K}^{2\times 2}$

$$\mathbf{A} = \begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix},$$

con $\alpha\delta \neq \beta\gamma$, allora si può verificare direttamente che sia l'inversa destra che l'inversa sinistra sono date dalla matrice

$$\mathbf{A}^{-1} = \frac{1}{\alpha\delta - \beta\gamma} \begin{bmatrix} \delta & -\beta \\ -\gamma & \alpha \end{bmatrix}.$$

Che cosa succede se $\alpha\delta = \beta\gamma$?

La situazione dell'esempio precedente, in cui l'inversa destra e quella sinistra coincidono, è del tutto generale. Infatti, vale il seguente teorema.

Teorema 7. *L'inversa di una matrice invertibile è unica.*

Dimostrazione. Dimostriamo prima di tutto che l'inversa destra e sinistra di una matrice invertibile coincidono. Sia $\mathbf{A}$ una matrice quadrata, $\mathbf{B}$ e $\mathbf{C}$ le sue due inverse destra e sinistra. In tal caso abbiamo

$$\mathbf{AB} = \mathbf{I}.$$

Moltiplicando a sinistra per $\mathbf{C}$ otteniamo

$$\mathbf{CAB} = \mathbf{CI},$$

[13] Leopold Kronecker 1823–1891.

ma $\mathbf{CA} = \mathbf{I}$ e quindi

$$\mathbf{IB} = \mathbf{CI} \quad \Rightarrow \quad \mathbf{B} = \mathbf{C}.$$

Siano ora $\mathbf{B}_1$ e $\mathbf{B}_2$ sono due inverse destre di $\mathbf{A}$, cioè assumiamo che valgano le relazioni

$$\mathbf{AB}_1 = \mathbf{AB}_2 = \mathbf{I}.$$

Dato che l'inversa destra e sinistra coincidono, possiamo dire che $\mathbf{B}_1$ è anche inversa sinistra,

$$\mathbf{B}_1\mathbf{A} = \mathbf{AB}_2 = \mathbf{I}.$$

Ma siccome l'inversa destra e sinistra coincidono, ne segue che $\mathbf{B}_1 = \mathbf{B}_2$. ■

Teorema 8. *Siano $\mathbf{A}, \mathbf{B} \in \mathbb{K}^{n\times n}$ invertibili. Allora il prodotto delle due matrici è invertibile e vale la formula $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$.*

Dimostrazione. Si verifica con un calcolo diretto che la formula fornisce l'espressione dell'inversa destra e sinistra:

$$\Big(\mathbf{AB}\Big)\Big(\mathbf{AB}\Big)^{-1} = \mathbf{ABB}^{-1}\mathbf{A}^{-1} = \mathbf{I}$$
$$\Big(\mathbf{AB}\Big)^{-1}\Big(\mathbf{AB}\Big) = \mathbf{B}^{-1}\mathbf{A}^{-1}\mathbf{AB} = \mathbf{I}$$

L'unicità dell'inversa è garantita dal teorema precedente. ■

Definizione 22 (Trasposta). Data la matrice $\mathbf{A} \in \mathbb{K}^{m\times n}$ si definisce *matrice trasposta* e la si indica con $\mathbf{A}^T$ la matrice $\mathbb{K}^{m\times n}$ definita come segue

$$\Big(\mathbf{A}^T\Big)_{ij} = A_{ji}, \qquad i = 1, 2, \ldots, m, \quad j = 1, 2, \ldots, n.$$

Esempio 22. Data la matrice 2×3

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 0 & 4 \end{bmatrix},$$

la sua trasposta è la seguente matrice 3×2

$$\mathbf{A}^T = \begin{bmatrix} 1 & 2 \\ 2 & 0 \\ 3 & 4 \end{bmatrix}.$$

Definizione 23 (Coniugata). Data la matrice $\mathbf{A} \in \mathbb{C}^{m\times n}$ si definisce *matrice coniugata* e la si denota con $\overline{\mathbf{A}}$ la matrice $\mathbb{C}^{m\times n}$ definita come segue

$$\left(\overline{\mathbf{A}}\right)_{ij} = \overline{A_{ij}}, \qquad i = 1, 2, \ldots, m \quad j = 1, 2, \ldots, n$$

Ovviamente se le componenti di $\mathbf{A}$ sono tutte reali, si ha $\mathbf{A} = \overline{\mathbf{A}}$.

Esempio 23. Data la matrice 2×2 a valori complessi

$$\mathbf{A} = \begin{bmatrix} 1+2\imath & 2+4\imath \\ 2-\imath & 0 \end{bmatrix},$$

la sua coniugata è la seguente matrice 2×2

$$\overline{\mathbf{A}}^T = \begin{bmatrix} 1-2\imath & 2-4\imath \\ 2+\imath & 0 \end{bmatrix}.$$

Definizione 24 (Trasposta coniugata). Data la matrice $\mathbf{A} \in \mathbb{K}^{m\times n}$ si definisce *matrice trasposta coniugata* e la si denota con $\mathbf{A}^H$ la matrice $\mathbf{B} \in \mathbb{K}^{m\times n}$ definita come segue

$$\left(\mathbf{A}^H\right)_{ij} = \overline{A_{ji}}, \qquad i = 1, 2, \ldots, m \quad j = 1, 2, \ldots, n.$$

Ovviamente si ha $\mathbf{A}^H = \overline{\mathbf{A}}^T$.

Esempio 24. Data la matrice 2×2 a valori complessi

$$\mathbf{A} = \begin{bmatrix} 1+2\imath & 2+4\imath \\ 2-\imath & 0 \end{bmatrix},$$

la sua trasposta coniugata è la seguente matrice 2×2

$$\overline{\mathbf{A}}^T = \begin{bmatrix} 1-2\imath & 2+\imath \\ 2-4\imath & 0 \end{bmatrix}.$$

Definizione 25 (Matrice simmetrica). La matrice *quadrata* $\mathbf{A} \in \mathbb{K}^{m\times m}$ si dice **simmetrica** se coincide con la sua trasposta, cioè

$$\mathbf{A} = \mathbf{A}^T.$$

Definizione 26 (Matrice hermitiana). La matrice *quadrata* $\mathbf{A} \in \mathbb{K}^{m \times m}$ si dice **hermitiana** se coincide con la sua trasposta coniugata, cioè

$$\mathbf{A} = \overline{\mathbf{A}}^T = \mathbf{A}^H.$$

1.2 Sistemi Lineari

1.2.1 Determinante: definizione assiomatica

In questi appunti introdurremo i determinanti in maniera assiomatica. Definiremo come determinante una particolare funzione

$$|\cdot| : \mathbb{K}^{n\times n} \mapsto \mathbb{K},$$

cioè una legge che ad ogni matrice quadrata $\mathbf{A} \in \mathbb{K}^n$ associa uno scalare, indicato nel testo col simbolo $|\,\mathbf{A}\,|$. Il determinante sarà definito in modo che alcune proprietà che enunceremo come assiomi siano sempre verificate.

Per semplificare l'esposizione introduciamo una notazione matriciale per colonne. Indicheremo con $\mathbf{A}_{\bullet j}$ la colonna j-esima della matrice $\mathbf{A}$,

$$\mathbf{A} = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & A_{n-1n} \\ A_{n1} & \cdots & A_{nn-1} & A_{nn} \end{bmatrix}, \quad \mathbf{A}_{\bullet j} = \begin{bmatrix} A_{1j} \\ A_{2j} \\ \vdots \\ A_{nj} \end{bmatrix},$$

in modo che la si possa pensare *partizionata per colonne*, cioè scritta come segue

$$\mathbf{A} = (\mathbf{A}_{\bullet 1}, \ldots, \mathbf{A}_{\bullet n}),$$

Ogni colonna $\mathbf{A}_{\bullet j}$ della matrice $\mathbf{A}$ è un vettore colonna e quindi si può esprimere come combinazione lineare dei vettori della base canonica di $\mathbb{K}^n$,

$$\mathbf{A}_{\bullet j} = \sum_{k=1}^{n} A_{kj}\mathbf{e}_k,$$

dove i coefficienti sono le stesse componenti della matrice sulla colonna considerata. La dipendenza della funzione determinante dalle colonne della matrice si scrive con la notazione

$$|\,\mathbf{A}\,| := |\,\mathbf{A}_{\bullet 1}, \ldots, \mathbf{A}_{\bullet n}\,|.$$

Enunciamo ora le tre proprietà fondamentali che definiscono assiomaticamente la funzione determinante.

Definizione 27 (Determinante). .

1. Il determinante è una funzione *multi-lineare* nelle colonne

$$|\ \ldots, \lambda\mathbf{a}, \ldots| = \lambda|\ \ldots, \mathbf{a}, \ldots|,$$

$$|\ \ldots, \mathbf{a}+\mathbf{b}, \ldots| = |\ \ldots, \mathbf{a}, \ldots| + |\ \ldots, \mathbf{b}, \ldots|.$$

2. Il determinante è nullo se due colonne consecutive sono uguali

$$|\ \ldots, \mathbf{a}, \mathbf{a}, \ldots| = 0.$$

3. Il determinante della matrice identità vale 1:

$$|\,\mathbf{I}\,| = |\,\mathbf{e}_1, \ldots, \mathbf{e}_n\,| = 1,$$

dove $\mathbf{e}_i$ sono i vettori della base canonica in $\mathbb{K}^n$.

Queste tre proprietà sono sufficienti per determinare l'esistenza e l'unicità della funzione determinante. Esse inoltre implicano un gran numero di conseguenze importanti, che verranno man mano discusse. Molte tra queste proprietà conseguenti dipendono tuttavia solo dalle prime due condizioni enunciate e sono indipendenti dalla terza. Per meglio evidenziare questo fatto introdurremo un simbolo alternativo, e cioè $\mathcal{D}(A)$, che rappresenta una funzione determinante piú generale. La funzione $\mathcal{D}(A)$ soddisfa le proprietà 1 e 2, ma non necessariamente la 3, potendo assumere un qualsiasi valore non nullo invece che l'unità. Useremo il simbolo $\mathcal{D}(A)$ al posto di $|\,A\,|$ ogni volta che non sarà necessaria la proprietà 3.

Osservazione 8. Mostriamo che in alcuni casi particolari è possibile definire facilmente una formula che esprime una funzione con le proprietà di un determinante.

$n = 1$

$$|\,A_{11}\,| = A_{11};$$

$n = 2$

$$\begin{vmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{vmatrix} = A_{11}\,A_{22} - A_{21}\,A_{12};$$

$n = 3$

$$\begin{vmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{vmatrix} = \begin{cases} A_{11}A_{22}A_{33} + A_{12}A_{23}A_{31} + A_{21}A_{32}A_{13} \\ -A_{13}A_{22}A_{31} - A_{12}A_{21}A_{33} - A_{11}A_{23}A_{32}. \end{cases}$$

1. *Matrici triangolari superiori*

$$\begin{vmatrix} A_{11} & A_{12} & \dots & A_{1n} \\ 0 & A_{22} & \dots & A_{2n} \\ \vdots & & & \\ 0 & 0 & \dots & A_{nn} \end{vmatrix} = A_{11}A_{22}\dots A_{nn}.$$

   Il determinante è il prodotto degli elementi sulla diagonale. Lo stesso vale per le matrici triangolari *inferiori* e per le matrici *diagonali*.

Si lascia per esercizio al lettore la verifica che tutte e tre le proprietà assiomatiche dei determinanti sono verificate dalle tre formule appena esposte.

1.2.2 Alcune proprietà dei determinanti

Lemma 9 (Prodotto per uno scalare). *Dalla proprietà 1 segue immediatamente che per ogni matrice* $\mathbf{A} \in \mathbb{K}^n$ *e per ogni scalare* λ *vale*

$$\mathcal{D}(\lambda\mathbf{A}) = \lambda^n\mathcal{D}(\mathbf{A})$$

Dimostrazione. Si osservi che valgono le seguenti uguaglianze

$$\begin{aligned} \mathcal{D}(\lambda\mathbf{A}_{\bullet 1}, \lambda\mathbf{A}_{\bullet 2}, \lambda\mathbf{A}_{\bullet 3}, \dots, \lambda\mathbf{A}_{\bullet n}) &= \lambda\mathcal{D}(\mathbf{A}_{\bullet 1}, \lambda\mathbf{A}_{\bullet 2}, \lambda\mathbf{A}_{\bullet 3}, \dots, \lambda\mathbf{A}_{\bullet n}) \\ &= \lambda^2\mathcal{D}(\mathbf{A}_{\bullet 1}, \mathbf{A}_{\bullet 2}, \lambda\mathbf{A}_{\bullet 3}, \dots, \lambda\mathbf{A}_{\bullet n}) \\ &= \dots \\ &= \lambda^n\mathcal{D}(\mathbf{A}_{\bullet 1}, \mathbf{A}_{\bullet 2}, \dots, \mathbf{A}_{\bullet n}) \end{aligned}$$

da cui il lemma segue immediatamente. ■

Osservazione 9 (Somma di matrici). Sempre dalla multilinearità espressa nella proprietà 1 si ricava un risultato negativo ma assai importante a proposito del determinante di una somma di matrici, e cioè

$$\mathcal{D}(\mathbf{A}+\mathbf{B}) \neq \mathcal{D}(\mathbf{A}) + \mathcal{D}(\mathbf{B})$$

Infatti, se la somma di due matrici si esprime mediante una matrice partizionata per colonne come segue

$$\mathbf{A}+\mathbf{B} = [\mathbf{A}_{\bullet 1} + \mathbf{B}_{\bullet 1}, \mathbf{A}_{\bullet 2} + \mathbf{B}_{\bullet 2}, \ldots, \mathbf{A}_{\bullet n} + \mathbf{B}_{\bullet n}]$$

dalla proprietà 1 di multilinearità si ha che

$$\begin{aligned}
\mathcal{D}(\mathbf{A}+\mathbf{B}) &= \mathcal{D}(\mathbf{A}_{\bullet 1}, \mathbf{A}_{\bullet 2} + \mathbf{B}_{\bullet 2}, \mathbf{A}_{\bullet 3} + \mathbf{B}_{\bullet 3}, \ldots, \mathbf{A}_{\bullet n} + \mathbf{B}_{\bullet n}) \\
&\quad + \mathcal{D}(\mathbf{B}_{\bullet 1}, \mathbf{A}_{\bullet 2} + \mathbf{B}_{\bullet 2}, \mathbf{A}_{\bullet 3} + \mathbf{B}_{\bullet 3}, \ldots, \mathbf{A}_{\bullet n} + \mathbf{B}_{\bullet n}), \\
&= \mathcal{D}(\mathbf{A}_{\bullet 1}, \mathbf{A}_{\bullet 2}, \mathbf{A}_{\bullet 3} + \mathbf{B}_{\bullet 3}, \ldots, \mathbf{A}_{\bullet n} + \mathbf{B}_{\bullet n}) \\
&\quad + \mathcal{D}(\mathbf{A}_{\bullet 1}, \mathbf{B}_{\bullet 2}, \mathbf{A}_{\bullet 3} + \mathbf{B}_{\bullet 3}, \ldots, \mathbf{A}_{\bullet n} + \mathbf{B}_{\bullet n}) \\
&\quad + \mathcal{D}(\mathbf{B}_{\bullet 1}, \mathbf{A}_{\bullet 2}, \mathbf{A}_{\bullet 3} + \mathbf{B}_{\bullet 3}, \ldots, \mathbf{A}_{\bullet n} + \mathbf{B}_{\bullet n}) \\
&\quad + \mathcal{D}(\mathbf{B}_{\bullet 1}, \mathbf{B}_{\bullet 2}, \mathbf{A}_{\bullet 3} + \mathbf{B}_{\bullet 3}, \ldots, \mathbf{A}_{\bullet n} + \mathbf{B}_{\bullet n}), \\
&= \ldots
\end{aligned}$$

ed è evidente che la relazione che ne risulta è in generale assai piú complicata di quella che esprime la somma dei due determinanti

$$\mathcal{D}(\mathbf{A}) + \mathcal{D}(\mathbf{B}) = \mathcal{D}(\mathbf{A}_{\bullet 1}, \mathbf{A}_{\bullet 2}, \ldots, \mathbf{A}_{\bullet n}) + \mathcal{D}(\mathbf{B}_{\bullet 1}, \mathbf{B}_{\bullet 2}, \ldots, \mathbf{B}_{\bullet n}).$$

Invitiamo il lettore a costruirsi un suo controesempio.

Sempre dalla proprietà 1 segue immediatamente il seguente lemma.

Lemma 10. *Se una colonna è nulla, il determinante è nullo.*

Dimostrazione. Se $\mathbf{v}_i = \mathbf{0}$, il vettore nullo, allora avremo

$$\begin{aligned}
\mathcal{D}(\ldots, \mathbf{0}, \ldots) &= \mathcal{D}(\ldots, 0 \cdot \mathbf{0}, \ldots), \\
&= 0 \cdot \mathcal{D}(\ldots, \mathbf{0}, \ldots), \\
&= 0.
\end{aligned}$$

Dalla proprietà 2 e combinando insieme le proprietà 1 e 2 si hanno una serie di conseguenze sul comportamento del determinante per scambio di colonne.

Lemma 11. *Se due colonne consecutive sono scambiate, il determinante cambia segno.*

Dimostrazione. Basta osservare che per la proprietà 2

$$\mathcal{D}(\ldots, \mathbf{w}+\mathbf{z}, \mathbf{w}+\mathbf{z}, \ldots) = 0,$$

ed usando la multilinearità

$$\begin{aligned} 0 &= \mathcal{D}(\ldots, \mathbf{w}+\mathbf{z}, \mathbf{w}+\mathbf{z}, \ldots), \\ &= \mathcal{D}(\ldots, \mathbf{w}, \mathbf{w}+\mathbf{z}, \ldots) + \mathcal{D}(\ldots, \mathbf{z}, \mathbf{w}+\mathbf{z}, \ldots), \\ &= \mathcal{D}(\ldots, \mathbf{w}, \mathbf{w}, \ldots) + \mathcal{D}(\ldots, \mathbf{w}, \mathbf{z}, \ldots) + \mathcal{D}(\ldots, \mathbf{z}, \mathbf{w}, \ldots) + \mathcal{D}(\ldots, \mathbf{z}, \mathbf{z}, \ldots), \end{aligned} \tag{1.16}$$

osserviamo che per la proprietà 2

$$\mathcal{D}(\ldots, \mathbf{w}, \mathbf{w}, \ldots) = 0, \qquad \mathcal{D}(\ldots, \mathbf{z}, \mathbf{z}, \ldots) = 0,$$

e quindi la (1.16) diventa

$$0 = \mathcal{D}(\ldots, \mathbf{w}, \mathbf{z}, \ldots) + \mathcal{D}(\ldots, \mathbf{z}, \mathbf{w}, \ldots),$$

cioè

$$\mathcal{D}(\ldots, \mathbf{w}, \mathbf{z}, \ldots) = -\mathcal{D}(\ldots, \mathbf{z}, \mathbf{w}, \ldots).$$

Questa proprietà può essere estesa anche allo scambio di due colonne in qualunque posizione e non necessariamente adiacenti. Osserviamo prima di tutto che vale il seguente, utilissimo, risultato.

Lemma 12. *Se due colonne sono uguali, il determinante è nullo.*

Dimostrazione. Supponiamo che $\mathbf{v}_i = \mathbf{v}_j$ per le due colonne di indice $i < j$. Possiamo scambiare il vettore colonna $\mathbf{v}_i$ con i vettori vicini, fino a portarlo adiacente al vettore v^j.

$$\begin{aligned} \mathcal{D}(\ldots, \mathbf{v}_i, \mathbf{v}_{i+1}, \mathbf{v}_{i+2}, \ldots, \mathbf{v}_j, \ldots) &= (-1)\mathcal{D}(\ldots, \mathbf{v}_{i+1}, \mathbf{v}_i, \mathbf{v}_{i+2}, \ldots, \mathbf{v}_j, \ldots), \\ &= (-1)^2\mathcal{D}(\ldots, \mathbf{v}_{i+1}, \mathbf{v}_{i+2}, \mathbf{v}_i, \ldots, \mathbf{v}_j, \ldots), \\ &= \ldots \\ &= \sigma\mathcal{D}(\ldots, \mathbf{v}_{i+1}, \mathbf{v}_{i+2} \ldots, \mathbf{v}_i, \mathbf{v}_j, \ldots) \end{aligned}$$

dove $\sigma = (-1)^{j-i}$, cioè può essere solo ± 1. Ma allora se sostituiamo $\mathbf{v}_i = v^j = \mathbf{a}$, per la proprietà 2 dei determinanti, si ha che

$$\mathcal{D}(\ldots, \mathbf{v}_{i+1}, \ldots, \mathbf{a}, \mathbf{a}, \ldots) = 0.$$

Lemma 13. *Se due colonne qualunque, ad esempio la colonna i-esima e j-esima (con $i \neq j$), sono scambiate, il determinante cambia segno.*

Dimostrazione. Si ripete lo stesso argomento utilizzato nel caso delle colonne consecutive.

$$\begin{aligned} 0 &= \mathcal{D}(\ldots, \mathbf{w}+\mathbf{z}, \ldots, \mathbf{w}+\mathbf{z}, \ldots), \\ &= \mathcal{D}(\ldots, \mathbf{w}, \ldots, \mathbf{w}+\mathbf{z}, \ldots) + \mathcal{D}(\ldots, \mathbf{z}, \ldots, \mathbf{w}+\mathbf{z}, \ldots), \\ &= \mathcal{D}(\ldots, \mathbf{w}, \ldots, \mathbf{w}, \ldots) + \mathcal{D}(\ldots, \mathbf{w}, \ldots, \mathbf{z}, \ldots) + \mathcal{D}(\ldots, \mathbf{z}, \ldots, \mathbf{w}, \ldots) + \\ &\quad \mathcal{D}(\ldots, \mathbf{z}, \ldots, \mathbf{z}, \ldots). \end{aligned}$$

Il primo e l'ultimo termine sono nulli per il lemma 12, per cui si ritrova l'espressione

$$0 = \mathcal{D}(\ldots, \mathbf{w}, \ldots, \mathbf{z}, \ldots) + \mathcal{D}(\ldots, \mathbf{z}, \ldots, \mathbf{w}, \ldots),$$

dove però ora $\mathbf{w}$ e $\mathbf{z}$ sono in qualsiasi posizione e non necessariamente adiacenti. ■

Combinando il lemma 12 con la multi-linearità – proprietà 1 – si ottiene un risultato molto interessante, di cui si farà uso nel seguito.

Lemma 14. *Se ad una colonna si somma una qualunque combinazione lineare delle altre (esclusa la colonna in esame) il valore del determinante non cambia.*

Dimostrazione. Sia [14]

$$\mathbf{b} = \sum_{j=1}^{n}{}^{(i)} \beta_j \mathbf{v}_j,$$

con $\beta_1, \ldots, \beta_n$ scalari qualsiasi. Allora

$$\begin{aligned} \mathcal{D}(\ldots, \mathbf{v}_{i-1}, \mathbf{v}_i + \mathbf{b}, \mathbf{v}_{i+1}, \ldots) &= \mathcal{D}\left(\ldots, \mathbf{v}_{i-1}, \mathbf{v}_i + \sum_{j=1}^{n}{}^{(i)} \beta_j \mathbf{v}_j, \mathbf{v}_{i+1}, \ldots\right), \\ &= \mathcal{D}(\ldots, \mathbf{v}_{i-1}, \mathbf{v}_i, \mathbf{v}_{i+1}, \ldots) + \sum_{j=1}^{n}{}^{(i)} \beta_j \mathcal{D}\left(\ldots, \mathbf{v}_{i-1}, \mathbf{v}_j, \mathbf{v}_{i+1}, \ldots\right). \end{aligned}$$

[14] Il simbolo $\sum_{j=1}^{n}{}^{(i)}$ significa che si sommano tutti i termini per $j = 1, \ldots, n$ escluso $j = i$.

Dato che per il lemma 13

$$\mathcal{D}(\ldots, \mathbf{v}_{i-1}, \mathbf{v}_j, \mathbf{v}_{i+1}, \ldots) = 0, \qquad \text{per} \quad j \neq i$$

si ottiene

$$\mathcal{D}(\ldots, \mathbf{v}_{i-1}, \mathbf{v}_i + \mathbf{b}, \mathbf{v}_{i+1}, \ldots) = \mathcal{D}(\ldots, \mathbf{v}_{i-1}, \mathbf{v}_i, \mathbf{v}_{i+1}, \ldots).$$

1.2.3 Esistenza ed unicità del determinante

Teorema 15. *Esiste una unica funzione determinante che soddisfa le proprietà* 1, 2, 3 *e che si indica col simbolo* $|\cdot|$.

Dimostrazione. Poiché, come si è già osservato nell'introduzione, si può scrivere

$$\mathbf{A}_{\bullet j} = \sum_{k=1}^{n} A_{kj}\mathbf{e}_k,$$

dalla multi-linearità del determinante segue che

$$\begin{aligned}
\mathcal{D}(\mathbf{A}_{\bullet 1}, \ldots, \mathbf{A}_{\bullet n}) &= \mathcal{D}\left(\sum_{i_1=1}^{n} A_{i_1 1}\mathbf{e}_{i_1}, \sum_{i_2=1}^{n} A_{i_2 2}\mathbf{e}_{i_2}, \ldots, \sum_{i_n=1}^{n} A_{i_n n}\mathbf{e}_{i_n}\right), \\
&= \sum_{i_1=1}^{n} A_{i_1 1}\mathcal{D}\left(\mathbf{e}_{i_1}, \sum_{i_2=1}^{n} A_{i_2 2}\mathbf{e}_{i_2}, \ldots, \sum_{i_n=1}^{n} A_{i_n n}\mathbf{e}_{i_n}\right), \qquad (1.17) \\
&= \sum_{i_1=1}^{n} A_{i_1 1} \sum_{i_2=1}^{n} A_{i_2 2} \cdots \sum_{i_n=1}^{n} A_{i_n n}\mathcal{D}(\mathbf{e}_{i_1}, \mathbf{e}_{i_2}, \ldots, \mathbf{e}_{i_n}).
\end{aligned}$$

Non è difficile rendersi conto che la sommatoria finale ottenuta in (1.17) coinvolge n^n termini, di cui però soltanto $n!$ sono non nulli. Infatti tutti i termini del tipo $\mathcal{D}(\ldots, \mathbf{e}_{i_s}, \ldots, \mathbf{e}_{i_t}, \ldots)$ con s e t tali che $i_s = i_t$ sono determinanti di matrici con almeno due colonne uguali e quindi nulli per il lemma 13. Gli unici termini non nulli sono quindi quelli per cui gli indici $i_1, \ldots i_n$ sono compresi tra 1 ed n ma sono tutti distinti, cioè formano una permutazione dei primi n numeri interi. Dato che ogni permutazione può essere ottenuta tramite una serie di scambi, si ha dal lemma 12 che

$$\mathcal{D}(\mathbf{e}_{i_1}, \mathbf{e}_{i_2}, \ldots, \mathbf{e}_{i_n}) = \sigma(i_1, i_2, \ldots, i_n)\mathcal{D}(\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n),$$

dove $\sigma(i_1, i_2, \ldots, i_n)$ può assumere solo i valori $+1$ o -1 ed è chiamata *segno* della permutazione $i_1, i_2, \ldots, i_n$. Osserviamo che il segno della permutazione è semplicemente -1 elevato al numero (minimo) di scambi necessari per trasformare la sequenza 1, 2, $\ldots$, n in i_1, i_2,$\ldots$, i_n ed è dunque indipendente dalla particolare funzione $\mathcal{D}(\,\cdot\,)$ che si sta considerando. Indichiamo con $\Pi(n)$ l'insieme di tutte le possibili permutazioni degli interi $1, \ldots, n$. La (1.17) può essere riscritta come

$$\mathcal{D}(\mathbf{A}_{\bullet 1}, \mathbf{A}_{\bullet 2}, \ldots, \mathbf{A}_{\bullet n}) = \sum_{(i_1, i_2, \ldots, i_n) \in \Pi(n)} \sigma(i_1, i_2, \ldots, i_n) A_{i_1 1} A_{i_2 2} \cdots A_{i_n n} \mathcal{D}(\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n).$$

Essendo $\mathcal{D}(\mathbf{I}) \neq 0$, possiamo introdurre il simbolo

$$|\,\mathbf{A}\,| = \sum_{(i_1, i_2, \ldots, i_n) \in \Pi(n)} \sigma(i_1, i_2, \ldots, i_n) A_{i_1 1} A_{i_2 2} \cdots A_{i_n n} = \frac{\mathcal{D}(\mathbf{A})}{\mathcal{D}(\mathbf{I})}.$$

Si può verificare immediatamente che le tre proprietà assiomatiche dei determinanti sono automaticamente verificate dalla funzione $|\,\mathbf{A}\,|$. Le prime due seguono dal fatto che $|\,\mathbf{A}\,|$ coincide con $\mathcal{D}(\mathbf{A})$ a meno di una costante moltiplicativa, mentre la terza è conseguenza della normalizzazione per $\mathcal{D}(\mathbf{I})$. ■

Si noti che la dimostrazione è in effetti costruttiva, in quanto suggerisce una formula, che, sebbene non praticissima, permette il calcolo del determinante di una matrice. Riprendiamo questo fatto enunciandolo come un corollario separato.

Corollario 16. *Per ogni $\mathbf{A} \in \mathbb{K}^n$ si ha*

$$|\,\mathbf{A}\,| = \sum_{(i_1, i_2, \ldots, i_n) \in \Pi(n)} \sigma(i_1, i_2, \ldots, i_n) A_{i_1 1} A_{i_2 2} \cdots A_{i_n n},$$

dove $\sigma(i_1, i_2, \ldots, i_n)$ è il segno della permutazione, definito come nella dimostrazione del teorema.

Infine, dalla dimostrazione segue anche un'altra proprietà interessante che sarà utilizzata in seguito.

Corollario 17. *Se $\mathcal{D}(\mathbf{A})$ soddisfa le proprietà 1 e 2 allora*

$$\mathcal{D}(\mathbf{A}) = |\,\mathbf{A}\,|\mathcal{D}(\mathbf{I}).$$

1.2.4 Determinanti del prodotto

Teorema 18 (di Binet). [15] *Siano* $\mathbf{A}$ *e* $\mathbf{B}$ *due matrici quadrate dello stesso ordine allora*

$$|\,\mathbf{AB}\,| = |\,\mathbf{A}\,||\,\mathbf{B}\,|.$$

Dimostrazione. Sia $\mathbf{C}$ la matrice prodotto $\mathbf{AB}$. Si noti che, partizionando $\mathbf{B}$ per colonne, si ha

$$\mathbf{C} = \mathbf{AB} = \mathbf{A}\,[\mathbf{B}_{\bullet 1}, \ldots \mathbf{B}_{\bullet n}] = [\mathbf{AB}_{\bullet 1}, \ldots \mathbf{AB}_{\bullet n}],$$

dato che le componenti di $\mathbf{C}_{\bullet j}$, j-esima colonna della matrice prodotto $\mathbf{C}$, sono

$$\mathbf{C}_{\bullet j} = \mathbf{AB}_{\bullet j} = \begin{bmatrix} \sum_{k=1}^{n} A_{1k} B_{kj} \\ \sum_{k=1}^{n} A_{2k} B_{kj} \\ \vdots \\ \sum_{k=1}^{n} A_{nk} B_{kj} \end{bmatrix}.$$

Il determinante della matrice prodotto è dunque

$$|\,\mathbf{C}\,| = |\,\mathbf{AB}\,| = |\,\mathbf{AB}_{\bullet 1}, \mathbf{AB}_{\bullet 2}, \ldots, \mathbf{AB}_{\bullet n}\,|.$$

Se consideriamo la matrice $\mathbf{A}$ fissata, possiamo introdurre una funzione

$$\mathcal{D}_{\mathbf{A}}(\mathbf{v}_1, \ldots, \mathbf{v}_n) = |\,\mathbf{Av}_1, \ldots, \mathbf{Av}_n\,|,$$

tale che $|\,\mathbf{AB}\,| = \mathcal{D}_{\mathbf{A}}(\mathbf{B})$. Osserviamo ora che valgono le seguenti relazioni

$$\begin{aligned} \mathcal{D}_{\mathbf{A}}(\mathbf{v}_1, \ldots, \lambda\mathbf{v}_k, \ldots, \mathbf{v}_n) &= |\,\mathbf{Av}_1, \ldots, \lambda\mathbf{Av}_k, \ldots, \mathbf{Av}_n\,|, \\ &= \lambda|\,\mathbf{Av}_1, \ldots, \mathbf{Av}_k, \ldots, \mathbf{Av}_n\,|, \\ &= \lambda\mathcal{D}_{\mathbf{A}}(\mathbf{v}_1, \ldots, \mathbf{v}_k, \ldots, \mathbf{v}_n), \end{aligned}$$

$$\begin{aligned} \mathcal{D}_{\mathbf{A}}(\mathbf{v}_1, \ldots, \mathbf{v}_k + \mathbf{w}_k, \ldots, \mathbf{v}_n) &= |\,\mathbf{Av}_1, \ldots, \mathbf{A}(\mathbf{v}_k + \mathbf{w}_k), \ldots, \mathbf{Av}_n\,|, \\ &= |\,\mathbf{Av}_1, \ldots, \mathbf{Av}_k + \mathbf{Aw}_k, \ldots, \mathbf{Av}_n\,|, \\ &= \mathcal{D}_{\mathbf{A}}(\mathbf{v}_1, \ldots, \mathbf{v}_k, \ldots, \mathbf{v}_n) + \mathcal{D}_{\mathbf{A}}(\mathbf{v}_1, \ldots, \mathbf{w}_k, \ldots, \mathbf{v}_n), \end{aligned}$$

[15] Jacques Philippe Marie Binet 1786–1856

ed ancora

$$\mathcal{D}_{\mathbf{A}}(\mathbf{v}_1, \ldots, \mathbf{a}, \mathbf{a}, \ldots, \mathbf{v}_n) = |\, \mathbf{A}\mathbf{v}_1, \ldots, \mathbf{A}\mathbf{a}, \mathbf{A}\mathbf{a}, \ldots, \mathbf{A}\mathbf{v}_n \,| = 0.$$

Quindi $\mathcal{D}_{\mathbf{A}}$ soddisfa le proprietà 1 e 2. Per il corollario del teorema 15 è vero che

$$\mathcal{D}_{\mathbf{A}}(\mathbf{B}_{\bullet 1}, \ldots, \mathbf{B}_{\bullet n}) = |\, \mathbf{B} \,| \mathcal{D}_{\mathbf{A}}(\mathbf{I}),$$

e dato che si ha anche

$$\begin{aligned} \mathcal{D}_{\mathbf{A}}(\mathbf{I}) &= |\, \mathbf{A}\mathbf{I}_{\bullet 1}, \ldots, \mathbf{A}\mathbf{I}_{\bullet n} \,| \\ &= |\, \mathbf{A}\mathbf{e}_1, \ldots, \mathbf{A}\mathbf{e}_n \,|, \\ &= |\, \mathbf{A}_{\bullet 1}, \ldots, \mathbf{A}_{\bullet n} \,|, \\ &= |\, \mathbf{A} \,|, \end{aligned}$$

si ottiene infine

$$|\, \mathbf{A}\mathbf{B} \,| = \mathcal{D}_{\mathbf{A}}(\mathbf{B}) = \mathcal{D}_{\mathbf{A}}(\mathbf{I}) |\, \mathbf{B} \,| = |\, \mathbf{A} \,|\, |\, \mathbf{B} \,|.$$

1.2.5 Determinante della matrice inversa

La formula del prodotto di determinanti permette di correlare immediatamente il determinante di una matrice con il determinante della matrice inversa. Infatti si ha che

$$1 = |\, \mathbf{I} \,| = |\, \mathbf{A}\mathbf{A}^{-1} \,| = |\, \mathbf{A} \,|\, |\, \mathbf{A}^{-1} \,|,$$

da cui si deduce che

$$|\, \mathbf{A}^{-1} \,| = \frac{1}{|\, \mathbf{A} \,|}.$$

1.2.6 Regola di Cramer

I determinanti hanno la loro applicazione principe nella regola di Cramer[16], che permette di formulare *in maniera teorica* la soluzione di un sistema lineare ad n equazioni in n incognite e matrice dei coefficienti non singolare.

[16] Gabriel Cramer 1704–1752.

Dato un sistema lineare, che scriveremo in forma matriciale compatta come $\mathbf{Ax} = \mathbf{b}$, si osservi innanzitutto che il termine noto si esprime come combinazione lineare dei vettori che formano le colonne della matrice

$$\mathbf{b} = \mathbf{Ax} = \mathbf{A}_{\bullet 1}x_1 + \mathbf{A}_{\bullet 2}x_2 + \cdots + \mathbf{A}_{\bullet n}x_n = \sum_{j=1}^{n} \mathbf{A}_{\bullet j}x_j .$$

utilizzando le componenti del vettore incognito $\mathbf{x}$ come coefficienti.

Calcoliamo ora

$$\begin{aligned}
\mathcal{D}(\mathbf{A}_{\bullet 1}, \ldots, \mathbf{A}_{\bullet i-1}, \mathbf{b}, \mathbf{A}_{\bullet i+1}, \ldots, \mathbf{A}_{\bullet n}) &= \mathcal{D}\left(\mathbf{A}_{\bullet 1}, \ldots, \mathbf{A}_{\bullet i-1}, \sum_{j=1}^{n} \mathbf{A}_{\bullet j}x_j, \mathbf{A}_{\bullet i+1}, \ldots, \mathbf{A}_{\bullet n}\right), \\
&= \sum_{j=1}^{n} x_j \mathcal{D}(\mathbf{A}_{\bullet 1}, \ldots, \mathbf{A}_{\bullet i-1}, \mathbf{A}_{\bullet j}, \mathbf{A}_{\bullet i+1}, \ldots, \mathbf{A}_{\bullet n}), \\
&= x_i \mathcal{D}(\mathbf{A}_{\bullet 1}, \ldots, \mathbf{A}_{\bullet i-1}, \mathbf{A}_{\bullet i}, \mathbf{A}_{\bullet i+1}, \ldots, \mathbf{A}_{\bullet n}),
\end{aligned}$$

da cui possiamo ricavare x_i

$$x_i = \frac{\mathcal{D}(\mathbf{A}_{\bullet 1}, \ldots, \mathbf{A}_{\bullet i-1}, \mathbf{b}, \mathbf{A}_{\bullet i+1}, \ldots, \mathbf{A}_{\bullet n})}{\mathcal{D}(\mathbf{A}_{\bullet 1}, \ldots, \mathbf{A}_{\bullet n})},$$

perché per $j \neq i$ ogni termine della sommatoria coinvolge il determinante di una matrice con due colonne della $\mathbf{A}$ ripetute ed è quindi nullo. La regola di Cramer ha senso, cioè è possibile calcolare gli x_i, solo se la matrice $\mathbf{A}$ dei coefficienti è non singolare (cioè le colonne sono linearmente indipendenti). Quindi si richiede che sia

$$\mathcal{D}(\mathbf{A}_{\bullet 1}, \ldots, \mathbf{A}_{\bullet n}) = \mathcal{D}(\mathbf{A}) \neq 0.$$

Osservazione 10. La regola di Cramer richiede il calcolo di $n + 1$ determinanti [17] ognuno dei quali a sua volta richiede $n!$ somme e prodotti – e quindi sono in tutto $(n + 1)!$ operazioni aritmetiche. Nel capitolo successivo si mostrerà che esistono tecniche di risoluzione dei sistemi lineari che forniscono con un costo computazionale $\mathcal{O}(n^3)$ operazioni[18]

[17] Perché se abbiamo solo n incognite?

[18] Ricordiamo che la definizione *corretta* del simbolo $\mathcal{O}(g(n))$, con $g : \mathbb{N} \to \mathbb{R}$ è

$$\mathcal{O}(g) = \Big\{ h : \mathbb{N} \to \mathbb{R} \;\; | \, h(n) \leq C(h)\, g(n), \text{ per } n \geq n_0 \geq 0, \\ \text{con } C(h) \text{ costante reale non negativa indipendente da } n \Big\}$$

di somma e prodotto la soluzione desiderata. Il costo della regola di Cramer è tale che già per $n = 5$ ne è sconsigliabile l'uso, mentre per n piú grandi la crescita del fattoriale rende questa tecnica di risoluzione di fatto impraticabile. La regola di Cramer ha dunque un significato essenzialmente *teorico*.

1.2.7 Dipendenza e indipendenza lineare

I determinanti possono essere usati per vedere se un insieme di vettori è o non è linearmente dipendente. Consideriamo prima il caso di n vettori $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ in $\mathbb{K}^n$. In tal caso possiamo calcolare il determinante della matrice costituita dalle colonne $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ e vale il teorema

Teorema 19. *Gli n vettori $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ in $\mathbb{K}^n$ sono linearmente indipendenti se e solo se*

$$\mathcal{D}(\mathbf{v}_1, \ldots, \mathbf{v}_n) \neq 0.$$

Dimostrazione. Supponiamo che $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ siano linearmente dipendenti, per cui esiste una scelta di coefficienti $\beta_1, \beta_2, \ldots, \beta_n$ non tutti simultaneamente nulli tale che

$$\mathbf{0} = \beta_1\mathbf{v}_1 + \beta_2\mathbf{v}_2 + \cdots + \beta_n\mathbf{v}_n,$$

Senza perdere in generalità possiamo supporre che almeno l'i-esimo coefficiente sia non nullo, cioè $\beta_i \neq 0$, e quindi ottenere

$$\mathbf{v}_i = -\frac{\beta_1}{\beta_i}\mathbf{v}_1 - \cdots - \frac{\beta_{i-1}}{\beta_i}\mathbf{v}_{i-1} - \frac{\beta_{i+1}}{\beta_i}\mathbf{v}_{i+1} - \frac{\beta_n}{\beta_i}\mathbf{v}_n = -\frac{1}{\beta_i}\sum_{j=1}^{n}{}^{(i)}\,\beta_j\mathbf{v}_j.$$

Poiché possiamo aggiungere alla colonna i-esima della matrice $\mathbf{A}$ una qualsiasi combinazione lineare delle altre colonne senza modificare il valore del determinante – vedi lemma 14 – si può scrivere

$$\begin{aligned}\mathcal{D}(\mathbf{v}_1, \ldots, \mathbf{v}_i, \ldots, \mathbf{v}_n) &= \mathcal{D}\left(\mathbf{v}_1, \ldots, \mathbf{v}_i + \frac{1}{\beta_i}\sum_{j=1}^{n}{}^{(i)}\,\beta_j\mathbf{v}_j, \ldots, \mathbf{v}_n\right),\\ &= \mathcal{D}(\mathbf{v}_1 \ldots, \mathbf{0}, \ldots, \mathbf{v}_n),\\ &= 0.\end{aligned}$$

e che l'espressione d'uso abituale $f = \mathcal{O}(g)$, che si legge "f è dell'ordine di g" è un abuso notazionale per $f \in \mathcal{O}(g)$.

Viceversa se gli n vettori $\mathbf{v}_1,\ldots,\mathbf{v}_n$ sono linearmente indipendenti, allora formano una base in $\mathbb{K}^n$ e quindi gli n vettori della base canonica si possono esprimere come loro combinazione lineare. Quindi per ogni $i = 1,\ldots,n$ si ha

$$\mathbf{e}_i = \sum_{k=1}^{n} A_{ik}\mathbf{v}_k.$$

Vale la sequenza di relazioni

$$\begin{aligned}
1 &\neq \mathcal{D}(\mathbf{I}) \\
&= \mathcal{D}(\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n) \\
&= \mathcal{D}(\sum_{k_1=1}^{n} A_{1k_1}\mathbf{v}_{k_1}, \sum_{k_2=1}^{n} A_{2k_2}\mathbf{v}_{k_2}, \ldots, \sum_{k_n=1}^{n} A_{nk_n}\mathbf{v}_{k_n}), \\
&= \sum_{k_1,k_2,\ldots,k_n=1}^{n} A_{1k_1}A_{2k_2}\cdots A_{nk_n}\mathcal{D}(\mathbf{v}_{k_1}, \mathbf{v}_{k_2}, \ldots, \mathbf{v}_{k_n}), \\
&= \sum_{k_1,k_2,\ldots,k_n\in S}^{n} \sigma(k_1, k_2, \ldots, k_n)A_{1k_1}A_{2k_2}\cdots A_{nk_n}\mathcal{D}(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n), \\
&= C\mathcal{D}(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n),
\end{aligned}$$

dove si è dapprima utilizzato lo stesso argomento del teorema 15 per ridurre la sommatoria alle permutazioni degli interi $1,\ldots n$ e quindi si è introdotta la quantità

$$C = \sum_{k_1,\ldots,k_n\in S}^{n} \sigma(k_1,\ldots,k_n)A_{1k_1}\cdots A_{nk_n}\ ,$$

che è ovviamente finita essendo somma di un numero finito di prodotti di quantità finite. Risulta evidentemente che $C \neq 0$ e che $\mathcal{D}(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n) \neq 0$. ■

È possibile generalizzare il risultato appena visto nel caso i vettori siano meno di n? La risposta è positiva, ma richiede l'introduzione di qualche strumento ulteriore.

Definizione 28 (Sottomatrice). Data una matrice $\mathbf{A}$ (non necessariamente quadrata) possiamo costruire una *sottomatrice* selezionando solo gli elementi che appartengono all'intersezione di una scelta fissata di righe e colonne di $\mathbf{A}$. Possiamo indicare una sottomatrice con la notazione esplicita

$$\mathbf{A}_{i_1 i_2 \ldots i_s}^{j_1 j_2 \ldots j_r},$$

dove gli indici $i_1, i_2, \ldots, i_s$ si riferiscono alle righe e gli apici $j_1, j_2, \ldots, j_r$ alle colonne della $\mathbf{A}$.

Ad esempio dalla matrice

$$\mathbf{A} = \begin{bmatrix} 1 & 3 & 3 & 0 \\ 0 & 2 & 2 & -1 \\ 1 & 1 & -5 & 9 \\ 3 & 5 & 7 & 0 \end{bmatrix},$$

possiamo estrarre la sottomatrice

$$\mathbf{A}_{12}^{14} = \begin{bmatrix} 1 & 3 \\ 3 & 5 \end{bmatrix}.$$

Ovviamente una sottomatrice è ancora una matrice e quindi può essere indicata al solito da una qualsiasi lettera maiuscola.

Definizione 29 (Minore). Data una matrice $\mathbf{A} \in \mathbb{K}^{n\times n}$, il determinante di una sua qualsiasi sottomatrice $\mathbf{A}_{i_1 i_2 \ldots i_k}^{j_1 j_2 \ldots j_k} \in \mathbb{K}^{k\times k}$ è detto *minore* di ordine k di $\mathbf{A}$.

Ci proponiamo ora di dimostrare il seguente teorema.

Teorema 20. *Se i k vettori $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$ sono linearmente dipendenti, tutti i minori di ordine k della matrice $\mathbf{A} \in \mathbb{K}^{k\times k}$ con $n \geq k$* [19]

$$\mathbf{A} = [\,\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\,]$$

sono nulli.

Dimostrazione. Un generico minore di ordine k ha la forma seguente

$$\mathbf{A}_{i_1 i_2 \ldots i_k}^{12\ldots k} = [\,\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_k\,],$$

dove $\mathbf{w}_i = (\mathbf{v}_i)_{i_1 i_2 \ldots i_k}$ è il vettore composto solo dalle componenti con indice di riga i_1, i_2, $\ldots$, i_k di $\mathbf{v}_i$. Poiché questi vettori sono linearmente dipendenti, esiste una combinazione lineare non nulla che genera il vettore nullo

$$\mathbf{0} = \beta_1 \mathbf{v}_1 + \beta_2 \mathbf{v}_2 + \cdots + \beta_k \mathbf{v}_k.$$

[19] Perché si suppone $n \geq k$? Cosa succederebbe se fosse invece $n < k$?

Se leggiamo questa combinazione lineare componente per componente sulle righe di indici $i_1, i_2, \ldots, i_k$, si ha che vale anche la relazione seguente

$$\mathbf{0} = \beta_1 \mathbf{w}_1 + \beta_2 \mathbf{w}_2 + \cdots + \beta_k \mathbf{w}_k,$$

da cui si conclude che i vettori $\mathbf{w}_1, \mathbf{w}_2, \ldots \mathbf{w}_k$ sono linearmente dipendenti e quindi che il minore è nullo per il teorema 19. ■

Possiamo "capovolgere" l'enunciato di questo teorema[20]: se esiste almeno un minore di ordine k non nullo, allora necessariamente i k vettori $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$ sono linearmente indipendenti. Si ottiene così in un criterio che permette di verificare l'indipendenza lineare di un insieme di k vettori in $\mathbb{K}^n$.

Vogliamo mostrare ora che la relazione che lega l'esistenza di un qualche minore di ordine k per k vettori colonna di $\mathbb{K}^n$ con la loro indipendenza lineare è di fatto una equivalenza, perché vale anche il seguente teorema.

Teorema 21. *Se i k vettori $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k \in \mathbb{K}^n$ con sono $k \leq n$ linearmente indipendenti, allora esiste almeno un minore non nullo di ordine k della matrice $\mathbf{A} \in \mathbb{K}^{k \times k}$ con $n \geq k$*

$$\mathbf{A} = [\, \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k \,].$$

La dimostrazione di questo teorema è un pochino piú laboriosa e richiede qualche risultato intermedio che enunciamo come lemma.

Lemma 22. *Se i k vettori $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$ sono linearmente indipendenti ed all'i-esimo vettore sostituiamo il vettore*

$$\mathbf{z} = \mathbf{v}_i + \sum_{j=1}^{k}{}^{(i)} \beta_j \mathbf{v}_j,$$

per una qualsiasi scelta degli scalari $\beta_1, \ldots, \beta_k$, si ha che i k vettori $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{i-1}, \mathbf{z}, \mathbf{v}_{i+1}, \ldots, \mathbf{v}_k$ sono ancora linearmente indipendenti.

Dimostrazione. Sia

$$\mathbf{0} = \gamma_1 \mathbf{v}_1 + \gamma_2 \mathbf{v}_2 + \cdots + \gamma_{i-1} \mathbf{v}_{i-1} + \gamma_i \mathbf{z} + \gamma_{i+1} \mathbf{v}_{i+1} + \cdots + \gamma_k \mathbf{v}_k,$$

[20] Ricordiamo che dalla logica elementare se la proposizione $\mathcal{A}$ implica $\mathcal{B}$, allora la negazione di $\mathcal{B}$ implica la negazione di $\mathcal{A}$.

che può essere riscritto come

$$\begin{aligned}\mathbf{0} &= \gamma_1\mathbf{v}_1 + \gamma_2\mathbf{v}_2 + \cdots + \gamma_{i-1}\mathbf{v}_{i-1} + \gamma_i\left(\mathbf{v}_i + \sum_{j=1}^{k}{}^{(i)}\beta_j\mathbf{v}_j\right) + \gamma_{i+1}\mathbf{v}_{i+1} + \cdots + \gamma_k\mathbf{v}_k,\\ &= (\gamma_1 + \gamma_i\beta_1)\mathbf{v}_1 + \cdots + (\gamma_{i-1} + \gamma_i\beta_{i-1})\mathbf{v}_{i-1} +\\ &\qquad \gamma_i\mathbf{v}_i + (\gamma_{i+1} + \gamma_i\beta_{i+1})\mathbf{v}_{i+1} + \cdots + (\gamma_k + \gamma_i\beta_k)\mathbf{v}_k,\end{aligned}$$

poiché $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$ sono linearmente indipendenti segue che

$$\gamma_s + \gamma_i\beta_s = 0, \quad s \neq i \qquad \text{e} \qquad \gamma_i = 0,$$

e quindi

$$\gamma_s = 0, \qquad s = 1, 2, \ldots, k$$

o in altre parole $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{i-1}, \mathbf{z}, \mathbf{v}_{i+1}, \ldots, \mathbf{v}_k$ sono linearmente indipendenti. ■

Lemma 23. *Sia $\mathbf{A}$ una matrice le cui colonne sono costituite dei vettori $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$ allora i minori di ordine k restano invariati se all'i-esimo vettore si aggiunge una combinazione lineare degli altri vettori cioè se si sostituisce al vettore $\mathbf{v}_i$ il vettore*

$$\mathbf{z} = \mathbf{v}_i + \sum_{j=1}^{k}{}^{(i)}\beta_j\mathbf{v}_j.$$

Dimostrazione. Basta osservare che un generico minore di ordine k della matrice trasformata si può scrivere come

$$\mathbf{B} = [\,\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{i-1}, \mathbf{x}, \mathbf{w}_{i+1}, \ldots, \mathbf{w}_k\,],$$

dove $\mathbf{w}_i = (\mathbf{v}_i)_{i_1 i_2 \ldots i_k}$ ed $\mathbf{x}_i = (\mathbf{z}_i)_{i_1 i_2 \ldots i_k}$ osserviamo anche che

$$\mathbf{x} = \mathbf{w}_i + \sum_{j=1}^{k}{}^{(i)}\beta_j\mathbf{w}_j,$$

e quindi per il lemma 14 il determinante non è cambiato. ■

I lemmi 22 e 23 ci danno la possibilità di semplificare la dimostrazione dell'inverso del teorema 20, infatti potremmo cambiare le colonne della matrice $\mathbf{A}$ con opportune combinazioni lineari che non cambiano il valore dei determinanti dei minori ne la dipendenza o indipendenza lineare dei vettori colonna e portano la matrice $\mathbf{A}$ in una forma opportuna che semplifica la dimostrazione.

Lemma 24. *Sia* $\mathbf{A}$ *una matrice le cui colonne sono costituite dai vettori* $\mathbf{v}_1$, $\mathbf{v}_2$,..., $\mathbf{v}_k$ *linearmente indipendenti allora tramite opportune combinazioni lineari dei vettori colonna che non cambiano il valore dei minori di ordine* k *può essere messa nella forma*

$$\mathbf{A} = \begin{bmatrix} \mathbf{T} \\ \mathbf{M} \end{bmatrix}, \quad \mathbf{T} = \begin{bmatrix} T_{11} & 0 & \cdots & 0 \\ T_{21} & T_{22} & & \vdots \\ \vdots & & \ddots & 0 \\ T_{k1} & T_{k2} & \cdots & T_{kk} \end{bmatrix},$$

con $T_{ii} \neq 0$, $i = 1, 2, \ldots, k$ *a meno di opportune permutazioni delle righe.*

Dimostrazione. La dimostrazione è fatta per costruzione. Supponiamo infatti che sia

$$(\mathbf{v}_1)_1 \neq 0;$$

se cosí non fosse basta scambiare la riga 1 con la prima riga i per cui

$$(\mathbf{v}_1)_i \neq 0,$$

osserviamo che un tale i esiste altrimenti il vettore $\mathbf{v}_1$ sarebbe nullo ed i vettori $\mathbf{v}_1$, $\mathbf{v}_2$, $\ldots$, $\mathbf{v}_k$ sarebbero linearmente dipendenti. Adesso ai vettori $\mathbf{v}_i$ sostituiamo i vettori

$$\mathbf{z}_i = \mathbf{v}_i - \frac{(\mathbf{v}_i)_1}{(\mathbf{v}_1)_1}\mathbf{v}_1 = [\,0, (\mathbf{z}_i)_2, (\mathbf{z}_i)_3, \ldots, (\mathbf{z}_i)_k\,]^T,$$

e quindi la matrice $\mathbf{A}$ si trasforma nella matrice $\mathbf{B}$ della forma

$$\mathbf{B} = \left[\begin{array}{c|c} (\mathbf{v}_1)_1 & 0 \cdots 0 \\ \hline (\mathbf{v}_1)_2 & \\ \vdots & \mathbf{C} \\ (\mathbf{v}_1)_n & \end{array}\right],$$

possiamo ora ripetere il ragionamento per la sottomatrice $\mathbf{C}$ perché una combinazione lineare di zeri da sempre zero e quindi la prima riga non viene piú modificata. ■

A questo punto la dimostrazione del teorema 21 è immediata.

Dimostrazione. Si consideri il minore $\mathbf{T}$ del lemma 24. ■

Osservazione 11. Il lettore smaliziato[21] noterà che stiamo utilizzando senza dirlo esplicitamente un processo di eliminazione di Gauss sulla matrice $\mathbf{B}^T$. La triangolarizzazione con scambio di righe può essere effettuata per i k passi necessari proprio perché la matrice $\mathbf{B}$ è supposta di rango massimo, il che garantisce l'esistenza dell'elemento pivotale non nullo. Questo procedimento fornisce anche una tecnica operativa per determinare il rango di una matrice in alternativa a quella "ingenua" di cercare "a occhio" la sottomatrice quadrata non singolare piú grande possibile.

Questi risultati si riassumono nel teorema finale.

Teorema 25. *I k vettori $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$ sono linearmente indipendenti se e solo se la matrice $\mathbf{A}$ le cui colonne sono formate dalle componenti degli stessi vettori rispetto alla base canonica ha almeno un minore di ordine k non nullo.*

1.2.8 Rango di una matrice

Definizione 30 (Rango). Data una matrice rettangolare $\mathbf{A}$ il massimo numero di vettori colonna di $\mathbf{A}$ linearmente indipendenti è detto rango della matrice $\mathbf{A}$ in simboli $\mathcal{R}\{\mathbf{A}\}$.

I teoremi fin qui dimostrati forniscono un procedimento per determinare il rango di una matrice.

Teorema 26. *Data una matrice $\mathbf{A}$ qualunque, il suo rango coincide l'ordine del piú grande minore non nullo.*

Dimostrazione. Sia

$$\mathbf{A} = [\,\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m\,],$$

e siano

$$\mathbf{a}_{i_1}, \mathbf{a}_{i_2}, \ldots, \mathbf{a}_{i_k},$$

vettori linearmente indipendenti, allora per teorema 25 tra tutti i minori di ordine k che sono formati dalle colonne $\mathbf{a}_{i_1}$, $\mathbf{a}_{i_2}$,...,$\mathbf{a}_{i_k}$ ne esiste almeno uno con determinante non nullo. Viceversa se $\mathbf{C}$ è un minore con determinante non nullo costruito dalle colonne $\mathbf{a}_{i_1}$, $\mathbf{a}_{i_2}$,...,$\mathbf{a}_{i_k}$ sempre per il teorema 25 segue che i vettori $\mathbf{a}_{i_1}$, $\mathbf{a}_{i_2}$,...,$\mathbf{a}_{i_k}$ sono linearmente indipendenti. ■

[21] Un lettore smaliziato è per esempio uno studente brillante che ha *letto e capito* gli argomenti presentati nel capitolo successivo.

1.2.9 Il teorema di Rouchè-Capelli

Utilizzando i teoremi che sono stati fin qui introdotti e dimostrati è possibile definire dei criteri per sapere se un dato sistema di equazioni lineari ammette o meno soluzioni. Consideriamo il sistema lineare

$$\mathbf{A}\mathbf{x} = \mathbf{b}, \tag{1.18}$$

dove $\mathbf{A} \in \mathbb{K}^{m\times n}$, $\mathbf{x} \in \mathbb{K}^n$ e $\mathbf{b} \in \mathbb{K}^m$. Supponiamo che $\mathbf{x}$ sia una possibile soluzione, senza preoccuparci in questo momento della sua unicità. Allora sappiamo che si può scrivere

$$\mathbf{b} = x_1\mathbf{A}_{\bullet 1} + x_2\mathbf{A}_{\bullet 2} + \cdots + x_n\mathbf{A}_{\bullet n},$$

cioè il vettore $\mathbf{b}$ è combinazione lineare delle colonne della matrice $\mathbf{A}$ con coefficienti le componenti del vettore soluzione. Quindi il numero di vettori linearmente indipendenti dell' insieme $\mathbf{A}_{\bullet 1}$, $\mathbf{A}_{\bullet 2}$, …, $\mathbf{A}_{\bullet n}$ è lo stesso dell' insieme $\mathbf{A}_{\bullet 1}$, $\mathbf{A}_{\bullet 2}$, …, $\mathbf{A}_{\bullet n}$, $\mathbf{b}$. Questo in simboli si può scrivere come

$$\mathcal{R}\{\mathbf{A}\} = \mathcal{R}\{\mathbf{A}, \mathbf{b}\},$$

dove con $\mathcal{R}\{\mathbf{A}, \mathbf{b}\}$ si intende il rango della matrice

$$[\,\mathbf{A}_{\bullet 1}, \mathbf{A}_{\bullet 2}, \ldots, \mathbf{A}_{\bullet n}, \mathbf{b}\,],$$

cioè la matrice costituita dalle colonne di $\mathbf{A}$ piú il vettore colonna $\mathbf{b}$.

Viceversa se il sistema (1.18) non ammette soluzioni, allora il vettore $\mathbf{b}$ non è una combinazione lineare dei vettori colonna della matrice $\mathbf{A}$. Quindi il vettore $\mathbf{b}$ è linearmente indipendente dai vettori colonna di $\mathbf{A}$ e quindi il numero di vettori colonna linearmente indipendenti dell'insieme $\mathbf{A}_{\bullet 1}$, $\mathbf{A}_{\bullet 2}$, …, $\mathbf{A}_{\bullet n}$ è piú piccolo del numero di vettori linearmente indipendenti dell'insieme $\mathbf{A}_{\bullet 1}$, $\mathbf{A}_{\bullet 2}$, …, $\mathbf{A}_{\bullet n}$, $\mathbf{b}$. Questo in simboli si può scrivere come

$$\mathcal{R}\{\mathbf{A}\} < \mathcal{R}\{\mathbf{A}, \mathbf{b}\}.$$

Questi risultati si sintetizzano nel seguente teorema

Teorema 27. *Sia dato il sistema*

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

dove $\mathbf{A} \in \mathbb{K}^{m\times n}$, $\mathbf{x} \in \mathbb{K}^n$ *e* $\mathbf{b} \in \mathbb{K}^m$. *Allora il sistema ammette soluzioni se e solo se*

$$\mathcal{R}\{\mathbf{A}\} = \mathcal{R}\{\mathbf{A}, \mathbf{b}\}.$$

I risultati sui determinanti permettono di dare una forma operativa del teorema 27, infatti

Teorema 28. *Sia dato il sistema*

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

dove $\mathbf{A} \in \mathbb{K}^{m\times n}$, $\mathbf{x} \in \mathbb{K}^n$ *e* $\mathbf{b} \in \mathbb{K}^m$. *Allora il sistema ha soluzione se e solo se il massimo ordine dei minori non nulli della matrice* $\mathbf{A}$ *è lo stesso della matrice* $[\,\mathbf{A}, \mathbf{b}\,]$.

Dimostrazione. Basta applicare il teorema 26 per il calcolo del rango di una matrice. ■

Sia $\mathbf{x} \in \mathbb{K}^n$ *una* soluzione del sistema lineare $\mathbf{A}\mathbf{x} = \mathbf{b}$, con $\mathbf{A} \in \mathbb{K}^{m\times n}$ e $\mathbf{b} \in \mathbb{K}^m$, con $\mathbf{b} \neq \mathbf{0}$, e $\tilde{\mathbf{x}} \in \mathbb{K}^n$ una soluzione del sistema lineare omogeneo $\mathbf{A}\tilde{\mathbf{x}} = 0$. È interessante notare che $\mathbf{x} + \tilde{\mathbf{x}}$ è ancora soluzione del sistema lineare non omogeneo:

$$\mathbf{A}(\mathbf{x} + \tilde{\mathbf{x}}) = \mathbf{A}\mathbf{x} + \mathbf{A}\tilde{\mathbf{x}} = \mathbf{b} + \mathbf{0} = \mathbf{b}.$$

Quindi, data una soluzione *particolare* $\mathbf{x}$ del sistema lineare non omogeneo — $\mathbf{b} \neq \mathbf{0}$ — *per ogni* soluzione $\tilde{\mathbf{x}}$ del sistema lineare omogeneo si ottiene una diversa soluzione del problema non omogeneo della forma $\mathbf{x} + \tilde{\mathbf{x}}$. È intuitivo che la soluzione di un sistema lineare non omogeneo – una volta che ne sia ammessa l'esistenza – è *unica* solo nei casi in cui possiamo garantire che sia sempre $\tilde{\mathbf{x}} = \mathbf{0}$, cioè che l'unica soluzione possibile del sistema lineare omogeneo $\mathbf{A}\tilde{\mathbf{x}} = \mathbf{0}$ sia il vettore nullo.

Formalizziamo questa considerazione nel seguente teorema per matrici quadrate.

Teorema 29. *Sia* $\mathbf{A} \in \mathbb{K}^{n\times n}$, *tale che* $\mathcal{R}\{\mathbf{A}\} = n$. *Allora la soluzione* $\mathbf{x} \in \mathbb{K}^n$ *del problema* $\mathbf{A}\mathbf{x} = \mathbf{b}$, *con* $\mathbf{b} \in \mathbb{K}^n$ *vettore assegnato,* ***esiste*** *ed è* ***unica***.

Dimostrazione. Dimostriamo separatamente prima l'esistenza e poi l'unicità.

(i) **Esistenza.** Poiché in $\mathbb{K}^n$ possiamo avere al massimo n vettori linearmente indipendenti, $\mathcal{R}\{\mathbf{A}\} = n$ implica che $\mathcal{R}\{\mathbf{A}, \mathbf{b}\} = n$, e l'esistenza segue dal teorema 27.

(ii) **Unicità.** Essendo la matrice $\mathbf{A}$ quadrata di ordine n uguale al suo rango, si deduce immediatamente che le sue colonne $\mathbf{A}_{\bullet 1}, \mathbf{A}_{\bullet 2}, \ldots, \mathbf{A}_{\bullet n}$, sono vettori linearmente indipendenti. Supponiamo ora di avere due soluzioni possibili del sistema lineare non omogeneo, che indichiamo con $\mathbf{x}_1$ e $\mathbf{x}_2$, per cui possiamo scrivere $\mathbf{A}\mathbf{x}_1 = \mathbf{A}\mathbf{x}_2 = \mathbf{b}$. Per differenza si ottiene $\mathbf{A}(\mathbf{x}_1 - \mathbf{x}_2) = 0$, da cui segue che $\mathbf{x}_1 - \mathbf{x}_2 = \mathbf{0}$, cioè $\mathbf{x}_1 = \mathbf{x}_2$. ■

Che cosa succede se la matrice è rettangolare?

Osservazione 12. Se la matrice è rettangolare con piú colonne che righe – $m < n$ l'unicità non è sicuramente possibile. Infatti il massimo rango possibile è m almeno $n - m$ colonne sono linearmente dipendenti dalle altre, per cui esistono almeno $n - m$ vettori linearmente indipendenti di $\mathbb{K}^n$, corrispondenti alle combinazioni lineari dei vettori colonna della matrice $\mathbf{A}$, che sono soluzione del problema omogeneo.

Osservazione 13. Nel caso di matrici rettangolari con piú righe che colonne – $m > n$ – il massimo rango possibile è ovviamente n. Tuttavia, il sistema lineare ammette soluzione soltanto nel caso si abbia $\mathcal{R}\{\mathbf{A}\} = \mathcal{R}\{\mathbf{A}, \mathbf{b}\}$. Viceversa il sistema lineare sarebbe infatti *sovradeterminato*. Questo significa che possiamo eliminare un certo numero di equazioni e ridurre la matrice finale a quadrata o rettangolare con piú colonne che righe. Nel caso sia $\mathcal{R}\{\mathbf{A}\} = n$ eliminando opportunamente $m - n$ equazioni siamo nella situazione del teorema 29. Altrimenti dovremo eliminare un numero maggiore di equazioni e saremo proprio nel caso dell'osservazione precedente.

1.2.10 Cofattori di una matrice quadrata

Sia $\mathbf{A}$ una matrice quadrata di ordine n che scriveremo come segue

$$\mathbf{A} = [\,\mathbf{A}_{\bullet 1}, \mathbf{A}_{\bullet 2}, \ldots, \mathbf{A}_{\bullet n}\,],$$

osserviamo che il determinante di $\mathbf{A}$ si può scrivere come

$$\begin{aligned} |\,\mathbf{A}\,| &= |\,\mathbf{A}_{\bullet 1}, \mathbf{A}_{\bullet 2}, \ldots, \mathbf{A}_{\bullet n}\,|, \\ &= |\,\textstyle\sum_{i=1}^{n} A_{i1}\mathbf{e}_i, \mathbf{A}_{\bullet 2}, \ldots, \mathbf{A}_{\bullet n}\,|, \\ &= \sum_{i=1}^{n} A_{i1}|\,\mathbf{e}_i, \mathbf{A}_{\bullet 2}, \ldots, \mathbf{A}_{\bullet n}\,|. \end{aligned}$$

Definizione 31. Chiameremo *cofattori* i determinanti della forma $|\,\mathbf{e}_i, \mathbf{A}_{\bullet 2}, \ldots, \mathbf{A}_{\bullet n}\,|$ e li indicheremo col simbolo

$$\alpha_{i1} = |\,\mathbf{e}_i, \mathbf{A}_{\bullet 2}, \ldots, \mathbf{A}_{\bullet n}\,|,$$

ed ovviamente per un elemento di posizione i, j qualsiasi

$$\alpha_{ij} = |\,\mathbf{A}_{\bullet 1}, \ldots, \mathbf{A}_{\bullet j-1}, \mathbf{e}_i, \mathbf{A}_{\bullet j+1}, \ldots, \mathbf{A}_{\bullet n}\,|.$$

Definizione 32 (Cofattore). Sia $\mathbf{A}$ una matrice quadrata definiremo matrice cofattore di $\mathbf{A}$ in simboli $\mathrm{cft}\{\mathbf{A}\}$ la seguente matrice

$$\mathrm{cft}\{\mathbf{A}\} = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \ddots & \alpha_{2n} \\ \vdots & \ddots & \ddots & \alpha_{n-1n} \\ \alpha_{n1} & \cdots & \alpha_{nn-1} & \alpha_{nn} \end{bmatrix},$$

cioè è la matrice in cui al posto dell' elemento A_{ij} sostituiamo il cofattore α_{ij}.

Introduciamo una notazione molto comoda per indicare i cofattori:

$$\alpha_{ij} = |\, \mathbf{A}_{\bullet 1}, \ldots, \mathbf{A}_{\bullet j-1}, \mathbf{e}_i, \mathbf{A}_{\bullet j+1}, \ldots, \mathbf{A}_{\bullet n} \,| = \left| \mathbf{A} \overset{j}{\leftarrow} \mathbf{e}_i \right|,$$

dove con $\mathbf{A} \overset{j}{\leftarrow} \mathbf{v}$ si indica la matrice uguale ad $\mathbf{A}$ tranne per la j-esima colonna che è sostituita con il vettore $\mathbf{v}$. Con questa notazione la regola di Cramer si può scrivere come segue

$$\mathbf{A}\mathbf{x} = \mathbf{b} \quad \Rightarrow \quad x_k = \frac{\left| \mathbf{A} \overset{k}{\leftarrow} \mathbf{b} \right|}{|\, \mathbf{A} \,|}.$$

La multi-linearità del determinante diventa

$$i) \qquad \left| \mathbf{A} \overset{k}{\leftarrow} \lambda \mathbf{v} \right| = \lambda \left| \mathbf{A} \overset{k}{\leftarrow} \mathbf{v} \right|,$$

$$ii) \qquad \left| \mathbf{A} \overset{k}{\leftarrow} \mathbf{v} + \mathbf{w} \right| = \left| \mathbf{A} \overset{k}{\leftarrow} \mathbf{v} \right| + \left| \mathbf{A} \overset{k}{\leftarrow} \mathbf{w} \right|.$$

Con questa notazione – tramite la linearità – il determinante si può calcolare con lo *sviluppo sulla prima colonna*

$$|\, \mathbf{A} \,| = \sum_{i=1}^{n} A_{i1} \left| \mathbf{A} \overset{1}{\leftarrow} \mathbf{e}_i \right| = \sum_{i=1}^{n} A_{i1} \alpha_{i1},$$

oppure, in maniera del tutto analoga, con lo *sviluppo sulla k-esima colonna*

$$|\, \mathbf{A} \,| = \sum_{i=1}^{n} A_{ik} \left| \mathbf{A} \overset{k}{\leftarrow} \mathbf{e}_i \right| = \sum_{i=1}^{n} A_{ik} \alpha_{ik}.$$

Enunciamo ora una proprietà che sarà utilizzata nel seguito.

Lemma 30. *Sia* $\mathbf{A}$ *una matrice quadrata e* α_{ij} *i suoi cofattori allora vale*

$$\delta_{jk}|\,\mathbf{A}\,| = \sum_{i=1}^{n} A_{ik}\alpha_{ij}.$$

Dimostrazione. Per $j = k$ il lemma fornisce esattamente la formula dello sviluppo del determinante sulla colonna k. Invece, se $j \neq k$, si ha

$$\begin{aligned}\sum_{i=1}^{n} A_{ik}\alpha_{ij} &= \sum_{i=1}^{n} A_{ik}\left|\,\mathbf{A} \overset{j}{\leftarrow} \mathbf{e}_i\,\right| = \sum_{i=1}^{n}\left|\,\mathbf{A} \overset{j}{\leftarrow} A_{ij}\mathbf{e}_i\,\right| = \left|\,\mathbf{A} \overset{j}{\leftarrow} \sum_{i=1}^{n} A_{ik}\mathbf{e}_i\,\right|, \\ &= \left|\,\mathbf{A} \overset{j}{\leftarrow} \mathbf{A}_{\bullet k}\,\right| = 0.\end{aligned}$$

Infatti la matrice $\mathbf{A} \overset{j}{\leftarrow} \mathbf{A}_{\bullet k}$ si ottiene sostituendo nella matrice $\mathbf{A}$ la colonna j-esima con la colonna k-esima. Se $j = k$ riotteniamo la matrice $\mathbf{A}$, altrimenti otteniamo una matrice con due colonne uguali, il cui determinante è nullo per il lemma 12. ■

1.2.11 Rappresentazione della matrice inversa

Il risultato precedente permette di definire una formula per la rappresentazione formale[22] dell'inversa di una matrice quadrata. Consideriamo, per iniziare, il seguente teorema.

Lemma 31. *Sia* $\mathbf{A}$ *una matrice quadrata di ordine* n *e* $cft\{\mathbf{A}\}$ *la sua matrice cofattore; allora vale*

$$cft\{\mathbf{A}\}^T\,\mathbf{A} = |\,\mathbf{A}\,|\mathbf{I}.$$

Dimostrazione. Scriviamo esplicitamente il prodotto della trasposta della matrice cofattore di $\mathbf{A}$ per la matrice $\mathbf{A}$. Si ottiene

$$\sum_{i=1}^{n}(\mathrm{cft}\{\mathbf{A}\}^T)_{ji}A_{ik} = \sum_{i=1}^{n} A_{ik}(\mathrm{cft}\{\mathbf{A}\})_{ij} = \sum_{i=1}^{n} A_{ik}\alpha_{ij} = \delta_{jk}|\,\mathbf{A}\,|$$

dove l'ultima uguaglianza si ricava applicando il lemma 30. ■

Il teorema seguente l'espressione dell'inversa di una matrice $\mathbf{A}$ con $|\,\mathbf{A}\,| \neq 0$.

[22] Si noti bene che si sta parlando di *rappresentazione formale* e non di *calcolo* della matrice inversa.

Teorema 32. *Sia* $\mathbf{A}$ *una matrice quadrata di ordine* n *con determinante non nullo. Allora* $\mathbf{A}$ *è invertibile e la sua inversa verifica la relazione*

$$\mathbf{A}^{-1} = \frac{1}{|\mathbf{A}|} cft\{\mathbf{A}\}^T .$$

Dimostrazione. Dal lemma 31 si ottiene subito l'espressione dell'inversa sinistra

$$\frac{\text{cft}\{\mathbf{A}\}^T}{|\mathbf{A}|}\mathbf{A} = \mathbf{I}.$$

L'unicità dell'inversa – destra e sinistra – conclude la dimostrazione del teorema. ■

Osservazione 14. Questa relazione fornisce una rappresentazione esplicita della matrice inversa; in particolare si osservi che l'inversa si può costruire solo se il determinante della matrice è non nullo, cioè la matrice $\mathbf{A}$ è di rango massimo.
Vale anche in questo caso la stessa osservazione fatta per la regola di Cramer, e cioè che questa formula ha un significato essenzialmente teorico, ma, dipendendo dal calcolo di n^2 cofattori, che, come vedremo nella prossima sezione, comporta il calcolo di minori di ordine $n-1$, è troppo costosa per essere di qualche utilità computazionale. Anticipiamo, come si è fatto per la regola di Cramer, che esistono tecniche numeriche che calcolano il determinante di una matrice con un costo computazionale assai inferiore.

1.2.12 Calcolo dei cofattori

Teorema 33. *Sia* $\mathbf{A}$ *una matrice quadrata allora*

$$\alpha_{ij} = (-1)^{i+j} |\mathbf{A}_{-i-j}| = (-1)^{i-j} |\mathbf{A}_{-i-j}|,$$

dove con $\mathbf{A}_{-i-j}$ *si intende la sottomatrice ottenuta sopprimendo la riga* i*-esima e la colonna* j*-esima della matrice* $\mathbf{A}$*.*

Dimostrazione. Osserviamo che

$$\alpha_{ij} = \left|\begin{array}{ccc:c:ccc} A_{11} & \cdots & A_{1j-1} & 0 & A_{1j+1} & \cdots & A_{1n} \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ A_{i-11} & \cdots & A_{i-1j-1} & 0 & A_{i-1j+1} & \cdots & A_{i-1n} \\ \hdashline A_{i1} & \cdots & A_{ij-1} & 1 & A_{ij+1} & \cdots & A_{in} \\ \hdashline A_{i+11} & \cdots & A_{i+1j-1} & 0 & A_{i+1j+1} & \cdots & A_{i+1n} \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ A_{n1} & \cdots & A_{nj-1} & 0 & A_{nj+1} & \cdots & A_{nn} \end{array}\right|,$$

e che tramite opportune combinazioni lineari delle colonne possiamo scrivere

$$\alpha_{ij} = \left|\begin{array}{ccc:c:ccc} A_{11} & \cdots & A_{1j-1} & 0 & A_{1j+1} & \cdots & A_{1n} \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ A_{i-11} & \cdots & A_{i-1j-1} & 0 & A_{i-1j+1} & \cdots & A_{i-1n} \\ \hdashline 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ \hdashline A_{i+11} & \cdots & A_{i+1j-1} & 0 & A_{i+1j+1} & \cdots & A_{i+1n} \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ A_{n1} & \cdots & A_{nj-1} & 0 & A_{nj+1} & \cdots & A_{nn} \end{array}\right|.$$

Quindi, definendo i vettori

$$\mathbf{v}_1 = \left[\begin{array}{c} A_{11} \\ \vdots \\ \hdashline A_{i-11} \\ \hdashline A_{i+11} \\ \vdots \\ A_{n1} \end{array}\right], \qquad \mathbf{v}_2 = \left[\begin{array}{c} A_{12} \\ \vdots \\ \hdashline A_{i-12} \\ \hdashline A_{i+12} \\ \vdots \\ A_{n2} \end{array}\right], \cdots, \qquad \mathbf{v}_{j-1} = \left[\begin{array}{c} A_{1j-1} \\ \vdots \\ \hdashline A_{i-1j-1} \\ \hdashline A_{i+1j-1} \\ \vdots \\ A_{nj-1} \end{array}\right],$$

$$\mathbf{v}_{j+1} = \left[\begin{array}{c} A_{1j+1} \\ \vdots \\ \hdashline A_{i-1j+1} \\ \hdashline A_{i+1j+1} \\ \vdots \\ A_{nj+1} \end{array}\right], \qquad \mathbf{v}_{j+2} = \left[\begin{array}{c} A_{1j+2} \\ \vdots \\ \hdashline A_{i-1j+2} \\ \hdashline A_{i+1j+2} \\ \vdots \\ A_{nj+2} \end{array}\right], \cdots, \qquad \mathbf{v}_n = \left[\begin{array}{c} A_{1n} \\ \vdots \\ \hdashline A_{i-1n} \\ \hdashline A_{i+1n} \\ \vdots \\ A_{nn} \end{array}\right],$$

si verifica facilmente che α_{ij} è funzione dei vettori $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{j-1}, \mathbf{v}_{j+1}, \ldots, \mathbf{v}_n$. Possiamo quindi scrivere

$$\alpha_{ij} = f_{ij}(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{j-1}, \mathbf{v}_{j+1}, \ldots, \mathbf{v}_n),$$

e la funzione f_{ij} soddisfa i primi due assiomi sui determinanti. Introduciamo per semplificare le notazioni il simbolo $\mathbf{I}_n \in \mathbb{K}^{n\times n}$ che rappresenta la matrice identità di dimensioni $n \times n$. Quindi per il corollario del teorema 15

$$\begin{aligned}\alpha_{ij} &= f_{ij}(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{j-1}, \mathbf{v}_{j+1}, \ldots, \mathbf{v}_n),\\ &= |\, \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{j-1}, \mathbf{v}_{j+1}, \ldots, \mathbf{v}_n \,| f_{ij}(\mathbf{I}_{n-1}),\end{aligned}$$

osserviamo che

$$\begin{aligned}|\, \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{j-1}, \mathbf{v}_{j+1}, \ldots, \mathbf{v}_n \,| &= \left|\begin{array}{ccc:ccc} A_{11} & \cdots & A_{1j-1} & A_{1j+1} & \cdots & A_{1n} \\ \vdots & & \vdots & \vdots & & \vdots \\ A_{i-11} & \cdots & A_{i-1j-1} & A_{i-1j+1} & \cdots & A_{i-1n} \\ \hdashline A_{i+11} & \cdots & A_{i+1j-1} & A_{i+1j+1} & \cdots & A_{i+1n} \\ \vdots & & \vdots & \vdots & & \vdots \\ A_{n1} & \cdots & A_{nj-1} & A_{nj+1} & \cdots & A_{nn} \end{array}\right|,\\ &= |\, \mathbf{A}_{-i-j} \,|.\end{aligned}$$

Supponendo ad esempio $i < j$ abbiamo

$$f_{ij}(\mathbf{I}_{n-1}) = |\, \mathbf{e}_1, \ldots, \mathbf{e}_{i-1}, \mathbf{e}_{i+1}, \ldots, \mathbf{e}_{j-1}, \mathbf{e}_i, \mathbf{e}_j, \ldots, \mathbf{e}_n \,|,$$

con $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$ base canonica di $\mathbb{K}^{n\times n}$; con $|i-j|$ scambi si ottiene

$$\begin{gathered}\mathbf{e}_1, \ldots, \mathbf{e}_{i-1}, \mathbf{e}_{i+1}, \ldots, \mathbf{e}_{j-1}, \mathbf{e}_i, \mathbf{e}_j, \ldots, \mathbf{e}_n,\\ \Downarrow\\ \mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n,\end{gathered}$$

e quindi

$$\begin{aligned}f_{ij}(\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{n-1}) &= |\, \mathbf{e}_1, \ldots, \mathbf{e}_{i-1}, \mathbf{e}_{i+1}, \ldots, \mathbf{e}_{j-1}, \mathbf{e}_i, \mathbf{e}_j, \ldots, \mathbf{e}_n \,|,\\ &= (-1)^{|i-j|} |\, \mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n \,|,\\ &= (-1)^{|i-j|}.\end{aligned}$$

Corollario 34. *Il determinante ha quindi il seguente sviluppo (detto di Laplace[23])*

$$|\, \mathbf{A} \,| = \sum_{i=1}^{n} (-1)^{i+j} A_{ij} |\, \mathbf{A}_{-i-j} \,|, \qquad j = 1, 2, \ldots, n.$$

[23] Pierre-Simon Laplace 1749–1827

1.2.13 Determinante della trasposta

Ci chiediamo ora che valore assume il determinante della matrice trasposta. Abbiamo già visto lo sviluppo del determinante lungo le colonne, cerchiamo ora una formula analoga per lo sviluppo per righe.

Teorema 35. *Il determinante si sviluppa come*

$$|\,\mathbf{A}\,| = \sum_{j=1}^{n} (-1)^{i+j} A_{ij} |\,A_{-i-j}\,|, \quad i = 1, 2, \ldots, n. \tag{1.19}$$

Dimostrazione. Basta osservare

$$\begin{aligned} |\,\mathbf{A}\,| &= |\,\mathbf{A}_{\bullet 1}, \mathbf{A}_{\bullet 2}, \ldots, \mathbf{A}_{\bullet n}\,|, \\ &= |\,A_{11}\mathbf{e}_1 + \mathbf{b}, \mathbf{A}_{\bullet 2}, \ldots, \mathbf{A}_{\bullet n}\,|, \end{aligned}$$

dove

$$\mathbf{b} = \begin{bmatrix} 0 \\ A_{21} \\ \vdots \\ A_{n1} \end{bmatrix},$$

ed usando la multilinearità del determinante

$$\begin{aligned} |\,\mathbf{A}\,| &= |\,A_{11}\mathbf{e}_1, \mathbf{A}_{\bullet 2}, \ldots, \mathbf{A}_{\bullet n}\,| + |\,\mathbf{b}, \mathbf{A}_{\bullet 2}, \ldots, \mathbf{A}_{\bullet n}\,|, \\ &= A_{11}|\,\mathbf{e}_1, \mathbf{A}_{\bullet 2}, \ldots, \mathbf{A}_{\bullet n}\,| + |\,\mathbf{b}, \mathbf{A}_{\bullet 2}, \ldots, \mathbf{A}_{\bullet n}\,|, \\ &= A_{11}\alpha_{11} + |\,\mathbf{b}, \mathbf{A}_{\bullet 2}, \ldots, \mathbf{A}_{\bullet n}\,|, \end{aligned}$$

in maniera analoga possiamo continuare

$$\begin{aligned} |\,\mathbf{A}\,| &= A_{11}\alpha_{11} + |\,\mathbf{b}, \mathbf{A}_{\bullet 2}, \ldots, \mathbf{A}_{\bullet n}\,|, \\ &= A_{11}\alpha_{11} + A_{12}\alpha_{12} + |\,\mathbf{c}, \mathbf{A}_{\bullet 2}, \ldots, \mathbf{A}_{\bullet n}\,|, \\ &\vdots \\ &= A_{11}\alpha_{11} + \cdots + A_{1n}\alpha_{1n} + |\,\mathbf{0}, \mathbf{A}_{\bullet 2}, \ldots, \mathbf{A}_{\bullet n}\,|, \\ &= A_{11}\alpha_{11} + \cdots + A_{1n}\alpha_{1n}, \end{aligned}$$

ed utilizzando il teorema 33 otteniamo proprio l' espressione (1.19). ■

Siamo in grado ora di dimostrare il seguente teorema

Teorema 36. *Sia* $\mathbf{A}$ *una matrice quadrata di ordine* n*, allora vale*

$$|\,\mathbf{A}\,| = |\,\mathbf{A}^T\,|.$$

Dimostrazione. Si può dimostrare per induzione sull'ordine della matrice.

$n = 1$ ovviamente non c'è nulla da dimostrare;

$n > 1$ assumiamo come ipotesi induttiva che l'enunciato del teorema sia vero per $n-1$ e verifichiamo che segue n. Consideriamo lo sviluppo del determinante di $\mathbf{A}$ sulla prima colonna e lo sviluppo del determinante di $\mathbf{A}^T$ sulla prima riga. In entrambi i casi abbiamo una somma di termini ognuno dei quali è il prodotto di un elemento della colonna di $\mathbf{A}$ o della riga di $\mathbf{A}^T$ – che quindi coincidono – per i corrispondenti cofattori. Ma questi sono determinanti di sottomatrici di ordine $n-1$ – si elimina sempre una riga ed una colonna – $\mathbf{A}$ e di $\mathbf{A}^T$ a due a due una trasposta dell'altra, per cui dall'ipotesi induttiva sono uguali. Il teorema segue immediatamente. ■

1.2.14 Determinante di matrici diagonali a blocchi

Teorema 37. *Sia* $\mathbf{A}$ *una matrice quadrata partizionata a blocchi come segue*

$$\mathbf{A} = \begin{bmatrix} \mathbf{B} & \mathbf{0} \\ \mathbf{0} & \mathbf{C} \end{bmatrix},$$

dove $\mathbf{B} \in \mathbb{K}^{p\times p}$ *e* $\mathbf{C} \in \mathbb{K}^{q\times q}$*, allora*

$$|\,\mathbf{A}\,| = |\,\mathbf{B}\,||\,\mathbf{C}\,|.$$

Dimostrazione. Osserviamo che fissata la matrice $\mathbf{C}$ il determinante di $\mathbf{A}$ è funzione delle colonne della matrice $\mathbf{B}$ e quindi possiamo scrivere

$$|\,\mathbf{A}\,| = \mathcal{D}_{\mathbf{C}}(\mathbf{B}_{\bullet 1}, \mathbf{B}_{\bullet 2}, \ldots, \mathbf{B}_{\bullet p}).$$

È immediato verificare che $\mathcal{D}_{\mathbf{C}}(\mathbf{B}_{\bullet 1}, \mathbf{B}_{\bullet 2}, \ldots, \mathbf{B}_{\bullet p})$ soddisfa i primi due assiomi dei determinanti e quindi per il corollario del teorema 15 si ha

$$\begin{aligned} \mathcal{D}_{\mathbf{C}}(\mathbf{B}_{\bullet 1}, \mathbf{B}_{\bullet 2}, \ldots, \mathbf{B}_{\bullet p}) &= |\,\mathbf{B}\,|\mathcal{D}_{\mathbf{C}}(\mathbf{I}_{\bullet 1}, \mathbf{I}_{\bullet 2}, \ldots, \mathbf{I}_{\bullet p}), \\ &= |\,\mathbf{B}\,|\begin{vmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{C} \end{vmatrix}. \end{aligned}$$

In maniera analoga otteniamo

$$\begin{aligned}\begin{vmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \mathbf{C}\end{vmatrix} &= \mathcal{D}_{\mathbf{I}}(\mathbf{C}_{\bullet 1}, \mathbf{C}_{\bullet 2}, \ldots, \mathbf{C}_{\bullet q})\\ &= |\,\mathbf{C}\,|\,|\,\mathbf{I}_{\bullet 1}, \mathbf{I}_{\bullet 2}, \ldots, \mathbf{I}_{\bullet q}\,|,\\ &= |\,\mathbf{C}\,|\begin{vmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\end{vmatrix},\\ &= |\,\mathbf{C}\,|,\end{aligned}$$

da cui segue subito

$$|\,\mathbf{A}\,| = |\,\mathbf{B}\,|\begin{vmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \mathbf{C}\end{vmatrix} = |\,\mathbf{B}\,|\,|\,\mathbf{C}\,|.$$

1.3 Autovalori ed autovettori

Sia $\mathbf{A} \in \mathbb{K}^{n\times n}$ e consideriamo la matrice $\mathbf{M}(\lambda)$ definita al variare dello scalare $\lambda \in \mathbb{K}$ dalla relazione

$$\mathbf{M}(\lambda) = \mathbf{A} - \lambda\mathbf{I}.$$

Per quali valori dello scalare λ la matrice $\mathbf{M}(\lambda)$ e la sua trasposta $\mathbf{M}^T(\lambda)$ sono singolari?

La condizione di singolarità si impone richiedendo che le matrici $\mathbf{M}(\lambda)$ e $\mathbf{M}^T(\lambda)$ non siano di rango massimo, cioè che il loro determinante sia nullo.

Dato che si ha $|\,\mathbf{M}(\lambda)\,| = |\,\mathbf{M}^T(\lambda)\,|$ per ogni scalare λ, è evidente che gli scalari che cerchiamo sono gli stessi per $|\,\mathbf{M}(\lambda)\,|$ ed $|\,\mathbf{M}^T(\lambda)\,|$.

Definizione 33. I valori che rendono singolare la matrice $\mathbf{M}(\lambda)$ – e quindi anche $\mathbf{M}^T(\lambda)$ – sono gli *autovalori* di $\mathbf{A}$.

Lemma 38. *Il determinante della matrice $\mathbf{M}(\lambda)$ è un polinomio in λ di ordine n.*

Dimostrazione. Dalla formula dello sviluppo del determinante:

$$|\,\mathbf{A} - \lambda\mathbf{I}\,| = \sum_{(i_1,i_2,\ldots,i_n)\in\Pi(n)} \sigma(i_1,i_2,\ldots,i_n)(A_{i_1 1} - \delta_{i_1 1}\lambda)\cdots(A_{i_n n} - \delta_{i_n n}\lambda)$$

tutti i termini della sommatoria sono polinomi in λ di grado al piú n. ∎

Dato che una matrice è singolare quando il determinante è nullo, la condizione di singolarità per $\mathbf{M}(\lambda)$ ci porta alla seguente definizione.

Definizione 34 (Polinomio caratteristico). Il polinomio

$$p_{\mathbf{A}}(\lambda) := |\,\mathbf{A} - \lambda\mathbf{I}\,|,$$

è detto polinomio caratteristico della matrice $\mathbf{A}$.

Osservazione 15. Gli zeri del polinomio caratteristico di $\mathbf{A}$ sono gli autovalori della matrice $\mathbf{A}$.

Per il teorema fondamentale dell'algebra il polinomio caratteristico si fattorizza come prodotto di binomi elevati ad opportuni esponenti m_{ai}

$$p_{\mathbf{A}}(\lambda) = (-1)^n(\lambda - \lambda_1)^{m_{a1}}(\lambda - \lambda_2)^{m_{a2}} \cdots (\lambda - \lambda_s)^{m_{as}},$$

dove $m_{a1} + m_{a2} + \cdots + m_{as} = n$ e $\lambda_i \neq \lambda_j$ se $i \neq j$.

Definizione 35. Se il polinomio caratteristico di $\mathbf{A}$ ammette $s \leq n$ zeri distinti che sono gli s autovalori, e si fattorizza come sopra, ad ogni autovalore λ_i compete una *molteplicità algebrica* m_{ai}.

Se la matrice $\mathbf{M}(\lambda)$ è singolare per un dato valore dello scalare λ, allora deve esistere almeno un vettore $\mathbf{x} \neq \mathbf{0}$ tale che

$$\mathbf{M}(\lambda)\mathbf{x} = (\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}.$$

Analogamente se la matrice $\mathbf{M}^T(\lambda)$ è singolare per un dato valore dello scalare λ, allora deve esistere almeno un vettore $\mathbf{y} \neq \mathbf{0}$ tale che

$$\mathbf{M}^T(\lambda)\mathbf{y} = (\mathbf{A}^T - \lambda\mathbf{I})\mathbf{y} = \mathbf{0}, \qquad \text{cioè, trasponendo,}$$
$$\mathbf{y}^T\mathbf{M}(\lambda) = \mathbf{y}(\mathbf{A} - \lambda\mathbf{I}) = \mathbf{0}^T.$$

In generale si ha che $\mathbf{x} \neq \mathbf{y}$, anche se si riferiscono allo stesso scalare λ che rende singolari $\mathbf{M}(\lambda)$ e $\mathbf{M}^T(\lambda)$.

Definizione 36 (Autovettore destro). Un vettore colonna $\mathbf{u} \neq \mathbf{0}$ è un autovettore *destro* della matrice $\mathbf{A}$ rispetto ad uno scalare λ se vale la relazione

$$\mathbf{A}\mathbf{u} = \lambda\mathbf{u}.$$

Definizione 37 (Autovettore sinistro). Un vettore riga $\mathbf{w}^T \neq \mathbf{0}$ è un autovettore *sinistro* della matrice $\mathbf{A}$ rispetto ad uno scalare λ se vale la relazione

$$\mathbf{w}^T\mathbf{A} = \lambda\mathbf{w}^T.$$

cioè $\mathbf{w}$ è autovettore di $\mathbf{A}^T$.

Osservazione 16. Ricordando la definizione di *kernel* di una matrice, possiamo dire che gli autovettori destri – risp. sinistri – di $\mathbf{A}$ rispetto ad un autovalore $\lambda \in \mathbb{K}^n$ sono tutti e soli gli elementi del $\ker(\mathbf{A} - \lambda\mathbf{I})$ – risp. del $\ker(\mathbf{A}^T - \lambda\mathbf{I})$.

Lemma 39. *Se $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$ sono k autovettori dello stesso autovalore λ_i, allora ogni loro combinazione lineare è ancora un autovettore rispetto allo stesso autovalore.*

Dimostrazione. Scelti k scalari qualsiasi $\beta_1, \beta_2, \ldots, \beta_k$ si ha che

$$\mathbf{A}\left(\sum_{j=1}^{k}\beta_j\mathbf{x}_j\right) = \sum_{j=1}^{k}\beta_j\mathbf{A}\mathbf{x}_j = \sum_{j=1}^{k}\beta_j\lambda_i\mathbf{x}_j = \lambda_i\left(\sum_{j=1}^{k}\beta_j\mathbf{x}_j\right).$$

Definizione 38 (Molteplicità geometrica). Definiamo come *molteplicità geometrica* dell'autovalore λ_i, indicata con m_{gi} il massimo numero di autovettori linearmente indipendenti che possiamo scegliere nello spazio di vettori $\ker(\mathbf{A} - \lambda\mathbf{I})$ associato all'autovalore λ_i.

Teorema 40. *La molteplicità algebrica m_{ai} di un autovalore λ_i è sempre maggiore o uguale della corrispondente molteplicità geometrica m_{gi}, cioè*

$$m_{ai} \geq m_{gi}.$$

Dimostrazione. Siano $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{m_{gi}}$ gli autovettori corrispondenti a λ_i, che possiamo sempre considerare *ortonormali* (altrimenti si applica Gram-Schmidt).
Se $m_{gi} < n$ aggiungiamo altri $n - m_{gi}$ vettori per completare una base ortonormale. Sia $\mathbf{T}$ la matrice le cui colonne sono questi vettori.

$$\mathbf{T} = (\mathbf{u}_1,\ \mathbf{u}_2,\ ,\mathbf{u}_3, \ldots, \mathbf{u}_{n-1},\ \mathbf{u}_n), \qquad \text{che verifica}$$
$$\mathbf{T}^T\mathbf{T} = \mathbf{I}, \qquad \text{da cui} \qquad \mathbf{T}^T = \mathbf{T}^{-1}$$
$$\mathbf{A}\mathbf{T} = \left[\,\lambda_i\mathbf{u}_1, \ldots, \lambda_i\mathbf{u}_{m_{gi}}, \mathbf{w}, \ldots, \mathbf{z}\,\right].$$

Possiamo quindi scrivere

$$\mathbf{T}^{-1}\mathbf{A}\mathbf{T} = \mathbf{T}^T\mathbf{A}\mathbf{T} = \left[\begin{array}{ccc:c} \lambda_i & & & \\ & \ddots & & \mathbf{B} \\ & & \lambda_i & \\ \hdashline & & & \\ \mathbf{0} & & & \mathbf{C} \\ & & & \end{array}\right],$$

da cui segue che

$$
\begin{aligned}
|\,\mathbf{A}-\lambda\mathbf{I}\,| &= |\,\mathbf{T}^{-1}(\mathbf{A}-\lambda\mathbf{I})\mathbf{T}\,| = |\,\mathbf{T}^{-1}\mathbf{A}\mathbf{T}-\lambda\mathbf{I}\,| \\
&= \left|\begin{array}{ccc:c} \lambda_i-\lambda & & & \\ & \ddots & & \mathbf{B} \\ & & \lambda_i-\lambda & \\ \hdashline \multicolumn{3}{c:}{\mathbf{0}} & \mathbf{C}-\lambda\mathbf{I} \end{array}\right|, \\
&= (\lambda_i-\lambda)^{m_{gi}}|\,\mathbf{C}-\lambda\mathbf{I}\,|,
\end{aligned}
$$

quindi λ_i è una radice del polinomio caretteristico di molteplicità almeno m_{gi} ma può essere superiore se $|\,\mathbf{C}-\lambda_i\mathbf{I}\,|=0$. Concludendo $m_{ai}\geq m_{gi}$. ■

Osservazione 17. Se un autovalore, per esempio l'i-esimo, ha molteplicità $m_{ai} > m_{gi}$, allora, poiché $m_{a1}+m_{a2}+\cdots+m_{as}=n$, non ci sono abbastanza autovettori linearmente indipendenti per formare una base.

Osservazione 18. Una matrice $\mathbf{A}$ è non singolare se e solo se non ammette lo zero come autovalore.

Osservazione 19. Se $\mathbf{A}$ è a valori reali e λ è un numero reale allora possiamo trovare una base di $\ker(\mathbf{A}-\lambda_i\mathbf{I})$ i cui vettori hanno componenti reali.

1.3.1 Matrici reali simmetriche e hermitiane

Ricordiamo le definizioni già introdotte nel capitolo riguardante vettori e matrici. Una matrice quadrata $\mathbf{A}$ si dice *simmetrica* se $\mathbf{A}=\mathbf{A}^T$ mentre si dice *hermitiana* se $\mathbf{A}=\mathbf{A}^H$.

Osservazione 20. Per matrici reali (tutte le componenti sono reali), i concetti di simmetria e hermitianeità coincidono, mentre differiscono per matrici complesse (almeno una componente è complessa).

Esempio 25.

$$
\mathbf{A}=\begin{bmatrix} 1 & 2 & 3 \\ 2 & 1 & 1 \\ 3 & 1 & 0 \end{bmatrix}, \qquad \mathbf{B}=\begin{bmatrix} 1 & 2 & 3 \\ -2 & 1 & 1 \\ -3 & 1 & 0 \end{bmatrix},
$$

$\mathbf{A}$ è una matrice simmetrica, mentre $\mathbf{B}$ non lo è.

Esempio 26.

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 1 & 1 \\ 3 & 1 & 0 \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} 1 & 2+\imath & 3-4\imath \\ 2-\imath & 1 & 1 \\ 3+4\imath & 1 & 0 \end{bmatrix},$$

$$\mathbf{C} = \begin{bmatrix} 1 & 2-\imath & 3+4\imath \\ 2-\imath & 1 & 1 \\ 3+4\imath & 1 & 0 \end{bmatrix},$$

$\mathbf{A}$ e $\mathbf{B}$ sono matrici hermitiane, $\mathbf{C}$ non lo è. $\mathbf{A}$ e $\mathbf{C}$ sono matrici simmetriche, $\mathbf{B}$ non lo è.

Completiamo queste due definizioni, introducendo una nuova definizione.

Definizione 39. Una matrice quadrata $\mathbf{A}$ si dice *normale* se $\mathbf{A}\mathbf{A}^H = \mathbf{A}^H\mathbf{A}$.

Esempio 27.

$$\mathbf{A} = \begin{bmatrix} 1 & 2-\imath & 0 \\ 2+\imath & 4 & 1 \\ 0 & 1 & 1 \end{bmatrix},$$

$\mathbf{A}$ è una matrice normale (è anche hermitiana).

Le proprietà di simmetria per matrici reali o di hermitianeità per matrici complesse, inducono proprietà molto "forti" su autovalori ed autovettori. Vale infatti il seguente teorema.

Teorema 41. *Una matrice reale simmetrica o hermitiana ha solo autovalori reali.*

Dimostrazione. $\mathbf{A}$ reale e simmetrica $\Rightarrow$ $\mathbf{A}$ hermitiana. Supponiamo che $\mathbf{A}$ sia hermitiana e che siano λ e $\mathbf{u} \neq \mathbf{0}$ un autovalore ed il suo autovettore. Moltiplicando a sinistra per $\mathbf{u}^H$ si ottiene

$$\mathbf{A}\mathbf{u} = \lambda\mathbf{u}, \qquad \Rightarrow \qquad \mathbf{u}^H\mathbf{A}\mathbf{u} = \lambda\mathbf{u}^H\mathbf{u} = \lambda \left\|\mathbf{u}\right\|_2^2,$$

da cui si ricava un espressione per l'autovalore λ

$$\mathbf{u}^H\mathbf{A}\mathbf{u} = \lambda \left\|\mathbf{u}\right\|_2^2, \qquad \Rightarrow \qquad \lambda = \frac{\mathbf{u}^H\mathbf{A}\mathbf{u}}{\left\|\mathbf{u}\right\|_2^2}.$$

Dalla definizione di norma e di autovettore segue che il denominatore è sicuramente un numero reale *strettamente* positivo. Anche il numeratore è un numero reale, come si vede dalla sequenza di uguaglianze

$$\begin{aligned}
\mathbf{u}^H\mathbf{A}\mathbf{u} &= (\mathbf{u}^H\mathbf{A}\mathbf{u})^T, && [\mathbf{u}^H\mathbf{A}\mathbf{u} \text{ è un numero}]\\
&= \overline{\overline{(\mathbf{u}^H\mathbf{A}\mathbf{u})^T}}, && [z = \overline{\overline{z}}]\\
&= \overline{(\mathbf{u}^H\mathbf{A}\mathbf{u})^H}, && [\overline{(\cdot)}^T = (\cdot)^H]\\
&= \overline{\mathbf{u}^H\mathbf{A}^H(\mathbf{u}^H)^H}, && [\text{trasposizione del prodotto}]\\
&= \overline{\mathbf{u}^H\mathbf{A}\mathbf{u}}, && [\mathbf{A} = \mathbf{A}^H].
\end{aligned}$$

quindi $\mathbf{u}^H\mathbf{A}\mathbf{u} \in \mathbb{R}$. Essendo λ un rapporto tra numeri reali, è a sua volta un numero reale. ■

Teorema 42. *Per ogni autovalore di una matrice reale simmetrica esiste un autovettore che ha solo componenti reali.*

Dimostrazione. Sia $\mathbf{A}$ reale e simmetrica, λ un autovalore e $\mathbf{u} = \mathbf{a} + \imath\mathbf{b}$ un suo autovettore in cui abbiamo separato le componenti reali e complesse:

$$\mathbf{A}(\mathbf{a} + \imath\mathbf{b}) = \lambda(\mathbf{a} + \imath\mathbf{b}),$$

da cui segue uguagliando le parti immaginarie e reali

$$\mathbf{A}\mathbf{a} = \lambda\mathbf{a}, \quad \mathbf{A}\mathbf{b} = \lambda\mathbf{b}.$$

Poiché si suppone sempre $\mathbf{u} \neq \mathbf{0}$, almeno uno dei vettori reali $\mathbf{a}$ o $\mathbf{b}$ è non nullo ed è autovettore di $\mathbf{A}$ rispetto all'autovalore λ. ■

Teorema 43. *Sia $\mathbf{A}$ matrice reale simmetrica o hermitiana, λ e μ due autovalori distinti e $\mathbf{u}$ e $\mathbf{v}$ due corrispondenti autovettori. Allora $\mathbf{u}$ e $\mathbf{v}$ sono ortogonali, cioè*

$$\mathbf{u} \cdot \mathbf{v} = 0.$$

Dimostrazione. Osserviamo che

$$\begin{aligned}
\mathbf{A}\mathbf{u} = \lambda\mathbf{u}, \quad &\Rightarrow \quad \mathbf{v}^H\mathbf{A}\mathbf{u} = \lambda\mathbf{v}^H\mathbf{u},\\
\mathbf{A}\mathbf{v} = \mu\mathbf{v}, \quad &\Rightarrow \quad \mathbf{u}^H\mathbf{A}\mathbf{v} = \mu\mathbf{u}^H\mathbf{v}\\
&\Rightarrow \quad \mathbf{v}^H\mathbf{A}\mathbf{u} = \mu\mathbf{v}^H\mathbf{u}.
\end{aligned}$$

Sottraendo membro a membro le due relazioni così ottenute

$$\mathbf{v}^H \mathbf{A}\mathbf{u} = \lambda \mathbf{v}^H \mathbf{u},$$
$$\mathbf{v}^H \mathbf{A}\mathbf{u} = \mu \mathbf{v}^H \mathbf{u}.$$

segue che

$$0 = (\lambda - \mu)\mathbf{v}^H \mathbf{u} \qquad \Rightarrow \qquad \mathbf{v}^H \mathbf{u} = \mathbf{u} \cdot \mathbf{v} = 0$$

perché $\lambda \neq \mu$. ■

Infine vale la seguente osservazione.

Osservazione 21. Se $\mathbf{A}$ è una matrice simmetrica o hermitiana, allora segue immediatamente dalle definizioni che ogni autovettore destro è anche autovettore sinistro. Per esempio, per $\mathbf{A}$ simmetrica

$$\begin{aligned} \mathbf{A}\mathbf{x} = \lambda \mathbf{x} \Rightarrow (\mathbf{A}\mathbf{x})^T &= \lambda \mathbf{x}^T, && \text{[trasposizione]} \\ \Rightarrow \mathbf{x}^T \mathbf{A}^T &= \lambda \mathbf{x}^T, && [\mathbf{A} = \mathbf{A}^T] \\ \Rightarrow \mathbf{x}^T \mathbf{A} &= \lambda \mathbf{x}^T. \end{aligned}$$

Infine, una proprietà fondamentale delle matrici reali simmetriche o hermitiane consiste nella possibilità di diagonalizzarle mediante opportune trasformazioni dette *trasformazioni di similitudine* che coinvolgono (ma solo in questo caso particolare) matrici *ortogonali* o *unitarie*, di cui diamo di seguito la definizione.

Definizione 40 (matrice ortogonale). Una matrice $\mathbf{A}$ è *ortogonale* se la sua trasposta coincide con la sua inversa. Cioè

$$\mathbf{A}^T = \mathbf{A}^{-1}.$$

Definizione 41 (matrice unitaria). Una matrice $\mathbf{A}$ è *unitaria* se la sua trasposta coniugata coincide con la sua inversa. Cioè

$$\mathbf{A}^H = \mathbf{A}^{-1}.$$

Osservazione 22. Nel caso la matrice $\mathbf{A}$ sia a valori reali allora il concetto di unitarietà e ortogonalità coincidono.

Teorema 44. *Sia* $\mathbf{A}$ *una matrice reale simmetrica, allora esiste una matrice* $\mathbf{U}$ *ortogonale tale che:*

$$\mathbf{U}^T\mathbf{A}\mathbf{U} = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix},$$

con $\lambda_1, \lambda_2, \ldots, \lambda_n$ *gli autovalori della matrice* $\mathbf{A}$

Dimostrazione. La dimostrazione è per induzione. Il teorema è vero per matrici 1×1. Sia λ_1 un autovalore di $\mathbf{A}$ (reale per il teorema 41) ed $\mathbf{w}_1$ un corrispondente autovettore di norma 1 (che possiamo assumere a valori reali per la osservazione 19). Possiamo completare $\mathbf{w}_1$ ad una base ortonormale $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_n$ per la osservazione 5. Sia $\mathbf{U}$ la matrice le cui colonne sono i vettori $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_n$. Allora avremo $\mathbf{W}^T\mathbf{W} = \mathbf{I}$, infatti

$$(\mathbf{W}^T\mathbf{W})_{ij} = \sum_{k=1}^{n} W_{ki}\, W_{kj} = \mathbf{W}_{\bullet i} \cdot \mathbf{W}_{\bullet j} = \delta_{ij}.$$

Inoltre dal fatto che $\mathbf{A}\mathbf{W}_{\bullet 1} = \lambda_1 \mathbf{W}_{\bullet 1}$ avremo

$$\mathbf{A}\mathbf{W} = \mathbf{W} \left[\begin{array}{c:c} \lambda_1 & \boldsymbol{\eta} \\ \hdashline \mathbf{0} & \mathbf{B} \end{array}\right],$$

dove $\boldsymbol{\eta}$ e $\mathbf{B}$ sono rispettivamente un vettore riga e una matrice quadrata non ancora specificate. Moltiplicando a sinistra per $\mathbf{W}^T$ otteniamo

$$\mathbf{W}^T\mathbf{A}\mathbf{W} = \left[\begin{array}{c:c} \lambda_1 & \boldsymbol{\eta}^T \\ \hdashline \mathbf{0} & \mathbf{B} \end{array}\right], \tag{1.20}$$

e dal fatto che $\mathbf{A}$ è simmetrica segue che $\mathbf{W}^T\mathbf{A}\mathbf{W}$ è simmetrica. Quindi la matrice a blocchi a destra in (1.20) è simmetrica e quindi $\boldsymbol{\eta} = \mathbf{0}$ e $\mathbf{B}$ è simmetrica. Applicando

l'induzione esisterà una matrice $\mathbf{S}$ ortogonale tale che

$$\mathbf{S}^T\mathbf{B}\mathbf{S} = \begin{bmatrix} \lambda_2 & & & \\ & \lambda_3 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix},$$

per cui avremo

$$\left[\begin{array}{c|c} 1 & \mathbf{0}^T \\ \hline \mathbf{0} & \mathbf{S}^T \end{array}\right] \mathbf{W}^T\mathbf{A}\mathbf{W} \left[\begin{array}{c|c} 1 & \mathbf{0}^T \\ \hline \mathbf{0} & \mathbf{S} \end{array}\right] = \left[\begin{array}{c|c} 1 & \mathbf{0}^T \\ \hline \mathbf{0} & \mathbf{S}^T \end{array}\right] \left[\begin{array}{c|c} \lambda_1 & \mathbf{0}^T \\ \hline \mathbf{0} & \mathbf{B} \end{array}\right] \left[\begin{array}{c|c} 1 & \mathbf{0}^T \\ \hline \mathbf{0} & \mathbf{S} \end{array}\right],$$

$$= \left[\begin{array}{c|c} \lambda_1 & \mathbf{0}^T \\ \hline \mathbf{0} & \mathbf{S}^T\mathbf{B}\mathbf{S} \end{array}\right], = \left[\begin{array}{c|ccc} \lambda_1 & \multicolumn{3}{c}{\mathbf{0}^T} \\ \hline & \lambda_2 & & \\ & & \lambda_3 & \\ \mathbf{0} & & & \ddots \\ & & & \lambda_n \end{array}\right],$$

$$= \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix}.$$

Quindi ponendo

$$\mathbf{U} = \mathbf{W} \left[\begin{array}{c|c} 1 & \mathbf{0}^T \\ \hline \mathbf{0} & \mathbf{S} \end{array}\right],$$

abbiamo che $\mathbf{U}$ è la matrice ortogonale cercata. ■

Teorema 45. *Sia* $\mathbf{A}$ *una matrice hermitiana, allora esiste una matrice unitaria* $\mathbf{U}$ *tale che:*

$$\mathbf{U}^H\mathbf{A}\mathbf{U} = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix},$$

con $\lambda_1, \lambda_2, \ldots, \lambda_n$ *gli autovalori della matrice* $\mathbf{A}$

Dimostrazione. La dimostrazione è praticamente identica a quella del teorema precedente, purché si sostituiscano i termini "simmetrico" e "ortogonale" con "hermitiano" e "unitario" e tutte le trasposizioni di matrici reali – che coinvolgono la notazione $(\cdot)^T$ – siano reinterpretate come trasposizioni con coniugazione – notazione $(\cdot)^H$. ■

1.3.2 Spettro, raggio spettrale e localizzazione degli autovalori sul piano complesso

Definizione 42 (Spettro). L'insieme degli autovalori di una matrice $\mathbf{A}$ si chiama *spettro* della matrice e si indica usualmente col simbolo $\sigma(\mathbf{A})$.

Definizione 43 (Raggio spettrale). Data una matrice quadrata $\mathbf{A}$ con autovalori $\lambda_1, \lambda_2, \ldots, \lambda_n$ il numero

$$\varrho(\mathbf{A}) = \max\{|\lambda_i| \ : \ i = 1, 2, \ldots, n\},$$

è detto *raggio spettrale* della matrice $\mathbf{A}$.

Esempio 28. Determinare il raggio spettrale della seguente matrice

$$\mathbf{A} = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 1 & 0 \\ 1 & 2 & 1 \end{bmatrix}$$

Calcoliamo gli autovalori come zeri del polinomio caratteristico

$$\begin{aligned} p_{\mathbf{A}}(\lambda) &= |\,\mathbf{A} - \lambda\mathbf{I}\,| = \ldots \\ &= (1-\lambda)(1-\lambda+i)(1-\lambda-i), \end{aligned}$$

Determiniamo il raggio spettrale dalla definizione

$$\varrho(\mathbf{A}) = \max\{|1|, |1+i|, |1-i|\} = \sqrt{2}.$$

Teorema 46 (dei cerchi di Gerschgorin). *Sia $\mathbf{A} \in \mathbb{K}^{n\times n}$ e l'insieme dei numeri complessi*

$$S = \bigcup_{i=1}^{n} \left\{ z \,:\, |A_{ii} - z| \leq \sum_{j=1}^{n}{}^{(i)} |A_{ij}| \right\};$$

allora ogni autovalore di $\mathbf{A}$ sta in S.

Dimostrazione. Sia λ, $\mathbf{u}$ una coppia autovalore, autovettore di $\mathbf{A}$, e k l'indice della componente di modulo massimo di $\mathbf{u}$

$$|u_k| = \max_{i=1,ldots s,n} |u_i|.$$

Calcoliamo la relazione $\mathbf{A}\mathbf{u} - \lambda\mathbf{u} = \mathbf{0}$ alla $k-$esima componente

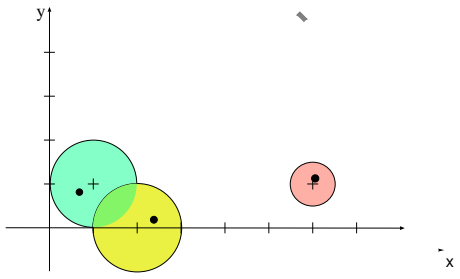
$$\sum_{j=1}^{n} A_{kj}u_j - \lambda u_k = 0, \qquad \Rightarrow \qquad (A_{kk} - \lambda)u_k = -\sum_{j=1}^{n}{}^{(k)} A_{kj}u_j,$$

Prendiamo il modulo da entrambe le parti e dividiamo per $|u_k| > 0$:

$$|A_{kk} - \lambda| \leq \sum_{j=1}^{n}{}^{(k)} |A_{kj}| \left|\frac{u_j}{u_k}\right| \leq \sum_{j=1}^{n}{}^{(k)} |A_{kj}|, \quad \Rightarrow \lambda \in S.$$

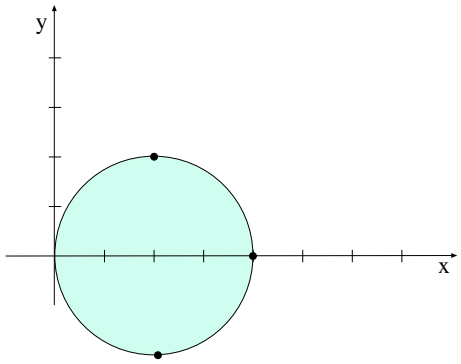
Esempio 29.

$$\mathbf{A} = \begin{bmatrix} 1+i & \frac{1}{2} & \frac{1}{2} \\ 1 & 2 & 0 \\ \frac{1}{4} & -\frac{1}{4} & 6+i \end{bmatrix}, \qquad \sigma(\mathbf{A}) = \begin{bmatrix} \lambda_1 &=& 0.69 + 0.82i, \\ \lambda_2 &=& 2.29 + 0.18i, \\ \lambda_3 &=& 6.02 + 1.01i \end{bmatrix}$$



Esempio 30. Possiamo avere anche situazioni “insolite”. Per esempio,

$$\mathbf{A} = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 2 & 2 \\ 0 & -2 & 2 \end{bmatrix}, \qquad \sigma(\mathbf{A}) = \begin{bmatrix} \lambda_1 & = & 4, \\ \lambda_2 & = & 2(1+i), \\ \lambda_3 & = & 2(1-i) \end{bmatrix}$$



1.3.3 Matrici a diagonale dominante

Definizione 44 (Dominanza diagonale). Una matrice quadrata $\mathbf{A}$ si dice a *diagonale dominante* se

$$|A_{ii}| \geq \sum_{j=1}^{n}{}^{(i)} |A_{ij}| \quad i = 1, 2, \ldots, n$$

e la disuguaglianza stretta vale per *almeno* un indice i. Se la disuguaglianza stretta vale *per ogni indice* i allora la matrice A si dice a *diagonale strettamente dominante*.

Teorema 47. *Se una matrice* $\mathbf{A}$ *è a diagonale strettamente dominante allora è non singolare.*

Dimostrazione. Basta osservare che lo zero non può appartenere all’unione dei cerchi di Gerschgorin, e quindi la matrice non ha l’autovalore nullo. ■

1.3.4 Matrici definite positive

Definizione 45 (matrice definita positiva). Sia $\mathbf{A}$ una matrice quadrata tale che per ogni vettore $\mathbf{u} \in \mathbb{K}^n$ valga

① $\mathbf{u}^T\mathbf{A}\mathbf{u} \geq 0$

② $\mathbf{u}^T\mathbf{A}\mathbf{u} = 0 \quad \Rightarrow \quad \mathbf{u} = 0$

allora diremo che $\mathbf{A}$ è *definita positiva*. Se la condizione ② viene a mancare allora diremo che $\mathbf{A}$ è *semidefinita positiva*.

Definizione 46 (matrice simmetrica e definita positiva). Se la matrice $\mathbf{A}$ definita positiva è anche simmetrica allora diremo che $\mathbf{A}$ è *simmetrica e definita positiva*, che indicheremo con l'abbreviazione SPD [24].

Teorema 48. *Se $\mathbf{A}$ è definita positiva allora ha solo autovalori positivi.*

Dimostrazione. Sia λ un autovalore di $\mathbf{A}$ allora esisterà un autovettore $\mathbf{u}$ tale che

$$\mathbf{A}\,\mathbf{u} = \lambda\mathbf{u},$$

moltiplicando per $\mathbf{u}^T$ abbiamo

$$\mathbf{u}^T\mathbf{A}\,\mathbf{u} = \lambda\mathbf{u}^T\mathbf{u} = \lambda\,\|\mathbf{u}\|_2^2\,,$$

da cui

$$\lambda = \frac{\mathbf{u}^T\mathbf{A}\,\mathbf{u}}{\|\mathbf{u}\|_2^2} > 0.$$

Teorema 49. *Se $\mathbf{A}$ è SPD allora $|\,\mathbf{A}\,| > 0$.*

Dimostrazione. $\mathbf{A}$ reale simmetrica $\Rightarrow$ esiste una matrice ortogonale reale $\mathbf{U}$ tale che

$$\mathbf{A} = \mathbf{U}^T\boldsymbol{\Lambda}\mathbf{U},$$

con $\boldsymbol{\Lambda} = diag(\lambda_1, \lambda_2, \ldots, \lambda_n)$, e $\lambda_i > 0$ autovalori di $\mathbf{A}$ SPD. Ricordiamo che

$$\begin{aligned}
|\,\mathbf{A}\,| &= |\,\mathbf{U}^{-1}\,||\,\boldsymbol{\Lambda}\,||\,\mathbf{U}\,|, && \text{[Binet]}\\
&= |\,\mathbf{U}^{-1}\mathbf{U}\,||\,\boldsymbol{\Lambda}\,|, && \text{[Binet]}\\
&= |\,\boldsymbol{\Lambda}\,|, && |\,\mathbf{I}\,| = 1\\
&= \lambda_1\lambda_2\cdots\lambda_n > 0
\end{aligned}$$

[24] Dall'inglese **S**ymmetric **P**ositive **D**efinite

Teorema 50. *Sia* $\mathbf{M} \in \mathbb{K}^{(n+m)\times(n+m)}$ *SPD della forma*

$$\mathbf{M} = \left[\begin{array}{c:c} \mathbf{A} & \mathbf{B}^T \\ \hdashline \mathbf{B} & \mathbf{C} \end{array}\right],$$

con $\mathbf{A} \in \mathbb{K}^{n\times n}$ *e* $\mathbf{C} \in \mathbb{K}^{m\times m}$ *allora le matrici* $\mathbf{A}$ *e* $\mathbf{C}$ *sono SPD.*

Dimostrazione. Consideriamo tutti i vettori della forma

$$\boldsymbol{\omega} = \left[\begin{array}{c} \mathbf{u} \\ \hdashline \mathbf{0} \end{array}\right].$$

Dalla sequenza di relazioni

$$0 \leq \boldsymbol{\omega}^T \mathbf{M} \boldsymbol{\omega} = \left[\begin{array}{c:c} \mathbf{u}^T & \mathbf{0}^T \end{array}\right] \left[\begin{array}{c:c} \mathbf{A} & \mathbf{B}^T \\ \hdashline \mathbf{B} & \mathbf{C} \end{array}\right] \left[\begin{array}{c} \mathbf{u} \\ \hdashline \mathbf{0} \end{array}\right] = \mathbf{u}^T \mathbf{A} \mathbf{u}.$$

segue che $\mathbf{u}^T\mathbf{A}\mathbf{u} \geq 0$ per ogni $\mathbf{u} \in \mathbb{K}^n$. Quando $\mathbf{u}^T\mathbf{A}\mathbf{u} = 0$ si ha anche $\boldsymbol{\omega}^T\mathbf{M}\boldsymbol{\omega} = 0$, e quindi $\boldsymbol{\omega} = \mathbf{0}$ perché $\mathbf{M}$ è SPD. Ma questo implica che $\mathbf{u} = \mathbf{0}$, di conseguenza $\mathbf{A}$ è SPD. In modo analogo si procede per la matrice $\mathbf{C}$, prendendo vettori della forma $\boldsymbol{\omega}^T = (\mathbf{0}, \mathbf{u})^T$. ■

Teorema 51 (di Sylvester). [25] *Sia* $\mathbf{M} \in \mathbb{R}^{n\times n}$ *e consideriamo la seguente partizione della matrice*

$$\mathbf{M} = \left[\begin{array}{c:c} \mathbf{A}^{(k)} & \mathbf{B}^T \\ \hdashline \mathbf{B} & \mathbf{C} \end{array}\right],$$

dove $\mathbf{A}^{(k)} \in \mathbb{R}^{k\times k}$ *e* $\mathbf{C} \in \mathbb{R}^{(n-k)\times(n-k)}$. *Allora sono equivalenti*

① $\mathbf{M}$ *è SPD,*

[25] James Joseph Sylvester 1814–1897

② $|\,\mathbf{A}^{(k)}\,| > 0 \quad per \quad k = 1, 2, \ldots, n.$

Dimostrazione. ①⇒② Per il teorema precedente ogni blocco $\mathbf{A}^{(k)}$ è SPD e quindi $|\,\mathbf{A}^{(k)}\,| > 0$
②⇒① Si dimostra per induzione come segue.

- $n = 1$ la matrice $\mathbf{M}$ è uno scalare ed il teorema è banalmente verificato.
- $n > 1$ Supponiamo il teorema vero per matrici di dimensione $n-1$ *(ipotesi induttiva)*.

Partizioniamo la matrice $\mathbf{M}$ nel seguente modo

$$\mathbf{M} = \left[\begin{array}{c:c} \mathbf{A}^{(n-1)} & \mathbf{b} \\ \hdashline \mathbf{b}^T & A_{nn} \end{array}\right],$$

e definiamo

$$\mathbf{L} = \left[\begin{array}{c:c} \mathbf{I} & -(\mathbf{A}^{(n-1)})^{-T}\mathbf{b} \\ \hdashline \mathbf{0} & 1 \end{array}\right], \quad \mathbf{L}^T = \left[\begin{array}{c:c} \mathbf{I} & \mathbf{0} \\ \hdashline -\mathbf{b}^T(\mathbf{A}^{(n-1)})^{-1} & 1 \end{array}\right].$$

Osserviamo che

$$\mathbf{L}^T\mathbf{M}\mathbf{L} = \left[\begin{array}{c:c} \mathbf{A}^{(n-1)} & \mathbf{0} \\ \hdashline \mathbf{0} & A_{nn} - \mathbf{b}^T\mathbf{A}^{(n-1)}b \end{array}\right].$$

Tenendo conto che $|\,\mathbf{L}\,| = |\,\mathbf{L}^T\,| = 1$ si ha che

$$\begin{aligned} |\,\mathbf{M}\,| &= |\,\mathbf{L}^{-T}\mathbf{L}^T\mathbf{M}\mathbf{L}\mathbf{L}^{-1}\,|, \\ &= |\,\mathbf{L}^{-T}\,||\,\mathbf{L}^T\mathbf{M}\mathbf{L}\,||\,\mathbf{L}^{-1}\,|, \\ &= \left|\left[\begin{array}{c:c} \mathbf{A}^{(n-1)} & \mathbf{0} \\ \hdashline \mathbf{0} & A_{nn} - \mathbf{b}^T\mathbf{A}^{(n-1)}\mathbf{b} \end{array}\right]\right|, \\ &= |\,\mathbf{A}^{(n-1)}\,|(A_{nn} - \mathbf{b}^T\mathbf{A}^{(n-1)}\mathbf{b}), \end{aligned}$$

e poiché $|\,\mathbf{M}\,|, \left|\,\mathbf{A}^{(n-1)}\,\right| > 0$ per ipotesi induttiva, allora segue che

$$A_{nn} - \mathbf{b}^T\mathbf{A}^{(n-1)}\mathbf{b} > 0.$$

Consideriamo ora un vettore qualunque $\mathbf{z}$, ponendo $\boldsymbol{\omega} = \mathbf{L}^{-1}\mathbf{z}$ e partizionando $\boldsymbol{\omega}$ come segue

$$\boldsymbol{\omega} = \begin{bmatrix} \mathbf{u} \\ \alpha \end{bmatrix},$$

dove $\mathbf{u}$ è un vettore di $n-1$ componenti ed α è uno scalare. Allora avremo che

$$\begin{aligned} \mathbf{z}^T\mathbf{M}\mathbf{z} &= \boldsymbol{\omega}\mathbf{L}^T\mathbf{M}\mathbf{L}\boldsymbol{\omega}, \\ &= [\,\mathbf{u}^T, \alpha\,] \left[\begin{array}{c:c} \mathbf{A}^{(n-1)} & \mathbf{0} \\ \hdashline \mathbf{0} & A_{nn} - \mathbf{b}^T\mathbf{A}^{(n-1)}\mathbf{b} \end{array}\right] \begin{bmatrix} \mathbf{u} \\ \alpha \end{bmatrix}, \\ &= \mathbf{u}^T\mathbf{A}^{(n-1)}\mathbf{u} + \alpha^2(A_{nn} - \mathbf{b}^T\mathbf{A}^{(n-1)}\mathbf{b}). \end{aligned}$$

Per ipotesi induttiva la matrice $\mathbf{A}^{(n-1)}$ è SPD, quindi

$$\mathbf{z}^T\mathbf{M}\mathbf{z} \geq 0.$$

Sia ora $\mathbf{z}^T\mathbf{M}\mathbf{z} = 0$ allora questo implica che $\alpha = 0$ ed $\mathbf{u} = \mathbf{0}$ quindi $\mathbf{z} = \mathbf{L}^{-1}\mathbf{0} = \mathbf{0}$. Di conseguenza $\mathbf{M}$ è SPD. ■

1.4 Norme di matrici

Vogliamo mostrare che la nozione di norma introdotta per i vettori si può estendere al caso delle matrici. Dato che l'idea di "lunghezza di una matrice" non è intuitiva, svilupperemo l'argomento in maniera formale. Iniziamo introducendo alcune proprietà interessanti delle norme vettoriali.

1.4.1 Alcune proprietà delle norme vettoriali

Teorema 52 (Continuità delle norme). *La norma vettoriale è una applicazione (uniformemente) continua dallo spazio dei vettori $\mathbb{K}^n$ in $\mathbb{R}$.*

Dimostrazione. L'uniforme continuità delle norme segue immediatamente dalla relazione

$$\Big|\, \|\mathbf{x}\| - \|\mathbf{y}\| \,\Big| \leq \|\mathbf{x} - \mathbf{y}\|$$

Teorema 53 (Equivalenza delle norme). *Se $\|\cdot\|'$ e $\|\cdot\|''$ sono due norme in $\mathbb{K}^n$, allora esistono due costanti $\alpha,\ \beta > 0$ tali che,*

$$\alpha \|\mathbf{x}\|'' \leq \|\mathbf{x}\|' \leq \beta \|\mathbf{x}\|'' \qquad \forall \mathbf{x} \in \mathbb{K}^n.$$

Dimostrazione. Se $\mathbf{x} = \mathbf{0}$ il teorema è banalmente verificato. Se $\mathbf{x} \neq \mathbf{0}$ si procede dimostrando che l'enunciato è vero per $\|\cdot\|'' = \|\cdot\|_\infty$ ed il caso generale segue immediatamente per confronto.

Consideriamo l'insieme $S = \{\mathbf{y} \in \mathbb{K}^n : \|\mathbf{y}\|_\infty = 1\}$. Si osservi che S è chiuso e limitato (per esempio tutte le componenti di ogni vettore $\mathbf{y} \in S$ verificano la condizione $|x_i| \leq 1$), quindi è un compatto in $\mathbb{K}^n$. La funzione norma $\|\cdot\|'$ è continua e quindi assume in S il suo minimo (finito) strettamente positivo ed il suo massimo che indicheremo rispettivamente con α e β. Per ogni vettore $\mathbf{x} \in \mathbb{K}^n$ si ha che

$$\mathbf{y} = \frac{\mathbf{x}}{\|\mathbf{x}\|_\infty} \in S$$

e quindi vale

$$0 < \alpha = \min_{\mathbf{y} \in S} \|\mathbf{y}\|' \leq \left\| \frac{\mathbf{x}}{\|\mathbf{x}\|_\infty} \right\|' \leq \max_{\mathbf{y} \in S} \|\mathbf{y}\|' = \beta\,,$$

da cui si ottiene $\alpha\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|' \leq \beta\|\mathbf{x}\|_\infty$. ∎

Osservazione 23. Per ogni vettore $\mathbf{x} \in \mathbb{K}^n$ valgono le diseguaglianze

$$
\begin{aligned}
&(i) \qquad \|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_2 \leq \sqrt{n}\|\mathbf{x}\|_\infty \\
&(ii) \qquad \|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_1 \leq \sqrt{n}\|\mathbf{x}\|_\infty \\
&(iii) \qquad \|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_1 \leq n\|\mathbf{x}\|_\infty
\end{aligned}
$$

Dimostrazione. Le disuguaglianze del punto (i) si ottengono di fatto ripetendo l'argomento utilizzato nella dimostrazione del teorema precedente; le disuguaglianze del punto (ii) si ottengono o per maggiorazione diretta o sfruttando alcune proprietà fondamentali delle norme vettoriali e dei prodotti scalari; infine, le disuguaglianze del punto (iii) si ottengono combinando le diseguaglianze dei primi due punti.

(i) Si prenda $\|\cdot\|' = \|\cdot\|_2$ e si osservi che ogni vettore $\mathbf{y} \in S$ deve avere almeno una componente di modulo uno, diciamo quella corrispondente all'indice k con $1 \leq k \leq n$, e tutte le altre per $i \neq k$ di modulo $|y_i| \leq |y_k| = 1$. Si ha che

$$
\|\mathbf{y}\|_2{}^2 = 1 + \sum_{i=1}^{n}{}^{(k)} |y_i|^2 \qquad \forall\, \mathbf{y} \in S,
$$

da cui segue che

$$
\begin{aligned}
\alpha &= \min_{\mathbf{y}\in S} \|\mathbf{y}\|_2 = 1 \\
\beta &= \max_{\mathbf{y}\in S} \|\mathbf{y}\|_2 = \sqrt{n}
\end{aligned}
$$

(ii) La prima diseguaglianza del punto (ii) si ottiene notando che per $\mathbf{x} \in \mathbb{K}^n$

$$
\|\mathbf{x}\|_2{}^2 = \sum_{i=1}^{n} |x_i|^2 \leq \sum_{i=1}^{n} |x_i|^2 + \sum_{i\neq j} |x_i|\,|x_j| = \left[\sum_{i=1}^{n} |x_i|\right]^2 = \|\mathbf{x}\|_1{}^2
$$

mentre la seconda si ottiene applicando la diseguaglianza di Cauchy-Schwarz

$$
|(\mathbf{x}, \mathbf{y})| \leq \|\mathbf{x}\|_2 \|\mathbf{y}\|_2
$$

al generico vettore $\mathbf{x} \in \mathbb{K}^n$ ed al vettore ausiliario $\mathbf{y}$ definito da

$$
y_i = \begin{cases} \dfrac{x_i}{\overline{x_i}} & \text{se } x_i \neq 0 \\ 0 & \text{se } x_i = 0 \end{cases}
$$

e considerando che vale

$$|(\mathbf{x},\mathbf{y})| = \left|\sum_{i=1}^{n} \overline{x_i} y_i\right| = \sum_{i=1}^{n} |x_i| = \|\mathbf{x}\|_1 \quad \text{e} \quad \|\mathbf{y}\|_2 = \Big[\,|(\mathbf{y},\mathbf{y})|\Big]^{\frac{1}{2}} = \left[\sum_{i=1}^{n} |y_i|^2\right]^{\frac{1}{2}} \leq \sqrt{n}$$

(iii) le diseguaglianze del punto (iii) si ottengono combinando facilmente le prime due.

■

1.4.2 Cosa è la norma di una matrice?

Sia data una matrice $\mathbf{A} \in \mathbb{K}^{m\times n}$ e un vettore colonna $\mathbf{x} \in \mathbb{K}^n$. Il prodotto $\mathbf{Ax}$ è un vettore colonna di dimensione m. Consideriamo inoltre due norme vettoriali, che indicheremo con $\|\cdot\|_a$ e $\|\cdot\|_b$, definite rispettivamente in $\mathbb{K}^m$ e $\mathbb{K}^n$. Cerchiamo la piú piccola costante K, ammesso che esista, che soddisfa la diseguaglianza

$$\|\mathbf{Ax}\|_a \leq K \|\mathbf{x}\|_b , \qquad \forall \mathbf{x} \in \mathbb{K}^n$$

Per $\mathbf{x} = \mathbf{0}$ la diseguaglianza precedente è verificata banalmente per qualsiasi costante K. Per questa ragione possiamo considerare solo vettori $\mathbf{x} \neq \mathbf{0}$, per i quali vale sicuramente l'espressione

$$\|\mathbf{Ax}\|_a = \frac{\|\mathbf{Ax}\|_a}{\|\mathbf{x}\|_b} \|\mathbf{x}\|_b \leq \left[\sup_{\mathbf{y}\neq\mathbf{0}} \frac{\|\mathbf{Ay}\|_a}{\|\mathbf{y}\|_b}\right] \|\mathbf{x}\|_b ,$$

che ci permette di identificare K con l'estremo superiore tra parentesi quadre. Nel seguito dimostreremo che nelle condizioni in cui ci siamo posti questo estremo superiore esiste sempre finito. Per ora osserviamo che questa quantità dipende soltanto dalla matrice A e dalla scelta delle norme vettoriali fatta all'inizio, come del resto è ragionevole aspettarsi. Questa quantità è quindi una caratteristica della matrice $\mathbf{A}$ e che si comporta come una norma, nel senso che soddisfa le tre proprietà fondamentali che intervengono nella definizione assiomatica delle norme.

Teorema 54 (di esistenza ed unicità). *Per ogni matrice* $\mathbf{A} \in \mathbb{K}^{m\times n}$ *vale*

$$\sup_{\mathbf{y}\neq\mathbf{0}} \frac{\|\mathbf{Ay}\|_a}{\|\mathbf{y}\|_b} = \max_{\|\mathbf{y}\|_b=1} \|\mathbf{Ay}\|_a .$$

Dimostrazione. Dato che vale

$$\frac{\|\mathbf{Ay}\|_a}{\|\mathbf{y}\|_b} = \left\|\mathbf{A}\left(\frac{\mathbf{y}}{\|\mathbf{y}\|_b}\right)\right\|_a = \|\mathbf{Az}\|_a \quad \text{con} \quad \mathbf{z} = \frac{\mathbf{y}}{\|\mathbf{y}\|_b} \quad \text{e} \quad \|\mathbf{z}\|_b = 1,$$

allora deve valere anche

$$\sup_{\mathbf{y}\neq\mathbf{0}} \frac{\|\mathbf{A}\mathbf{y}\|_a}{\|\mathbf{y}\|_b} = \sup_{\|\mathbf{z}\|_b=1} \|\mathbf{A}\mathbf{z}\|_a .$$

Ragionando come nella dimostrazione del teorema di equivalenze delle norme, introduciamo ancora l'insieme $S = \{\mathbf{z} \in \mathbb{K}^n : \|\mathbf{z}\|_\infty = 1\}$, che abbiamo detto essere un compatto in $\mathbb{K}^n$. La funzione che associa ad ogni vettore $\mathbf{z} \in S$ la quantità $\|\mathbf{A}\mathbf{z}\|_a$ è ovviamente continua (composizione di funzioni continue) e quindi assume in S il suo massimo. Sia $\mathbf{z}_M \in S$ il vettore tale che

$$\|\mathbf{A}\mathbf{z}_M\|_a = \max_{\|\mathbf{z}\|_b=1} \|\mathbf{A}\mathbf{z}\|_a .$$

Essendo l'estremo superiore un maggiorante e non potendo essere strettamente maggiore di altri maggioranti (per esempio il max appena trovato) ne consegue che

$$\max_{\|\mathbf{z}\|_b=1} \|\mathbf{A}\mathbf{z}\|_a = \|\mathbf{A}\mathbf{z}_M\|_a \leq \sup_{\|\mathbf{z}\|_b=1} \|\mathbf{A}\mathbf{z}\|_a \leq \max_{\|\mathbf{z}\|_b=1} \|\mathbf{A}\mathbf{z}\|_a .$$

Teorema 55 (Norma di matrice). *La quantità di cui si è appena dimostrata l'esistenza e l'unicità*

$$\|\mathbf{A}\| = \sup_{\mathbf{y}\neq\mathbf{0}} \frac{\|\mathbf{A}\mathbf{y}\|_a}{\|\mathbf{y}\|_b}$$

soddisfa le tre proprietà assiomatiche delle norme. Questo ci giustifica nell'usare il simbolo $\|\mathbf{A}\|$ *che indicherà la norma della matrice* $\mathbf{A}$.

Dimostrazione. Proveremo che la funzione che associa ad ogni matrice $\mathbf{A} \in \mathbb{K}^n$ il valore reale non negativo indicato con $\|\mathbf{A}\|$ soddisfa le tre proprietà assiomatiche introdotte nella definizione delle norme.

1. $\|\mathbf{A}\|$ ovviamente non può essere negativa. Dobbiamo verificare dunque che è nulla se e solo se $\mathbf{A} = \mathbf{0}$. Se consideriamo la matrice nulla $\mathbf{A} = \mathbf{0}$ allora si ha che $\mathbf{A}\mathbf{x} = \mathbf{0}$ per ogni $\mathbf{x} \in \mathbb{K}^n$ e quindi

   $$\frac{\|\mathbf{A}\mathbf{x}\|_a}{\|\mathbf{x}\|_b} = 0 \qquad \forall\, \mathbf{x} \neq \mathbf{0},$$

   ma ovviamente $\|\mathbf{A}\| = 0$. Dimostriamo che $\|\mathbf{A}\| = 0$ implica $\mathbf{A} = \mathbf{0}$ ragionando per negazione, cioè supponiamo che $\mathbf{A} \neq \mathbf{0}$ e verifichiamo che $\|\mathbf{A}\| > 0$. Sia infatti $\mathbf{e}_i$

l'i-esimo vettore della base canonica. Quindi se $\mathbf{A} \neq \mathbf{0}$ allora, $\mathbf{A}\mathbf{e}_i \neq \mathbf{0}$ per almeno un indice i e quindi

$$\frac{\|\mathbf{A}\mathbf{e}_i\|_a}{\|\mathbf{e}_i\|_b} > 0.$$

2. Siano $\mathbf{A}, \mathbf{B} \in \mathbb{K}^{m\times n}$

$$\frac{\|(\mathbf{A}+\mathbf{B})\mathbf{x}\|_a}{\|\mathbf{x}\|_b} \leq \frac{\|\mathbf{A}\mathbf{x}\|_a}{\|\mathbf{x}\|_b} + \frac{\|\mathbf{B}\mathbf{x}\|_a}{\|\mathbf{x}\|_b} \leq \sup_{\mathbf{y}\neq\mathbf{0}} \frac{\|\mathbf{A}\mathbf{y}\|_a}{\|\mathbf{y}\|_b} + \sup_{\mathbf{y}\neq\mathbf{0}} \frac{\|\mathbf{B}\mathbf{y}\|_a}{\|\mathbf{y}\|_b} \leq \|\mathbf{A}\| + \|\mathbf{B}\|,$$

poiché questa diseguaglianza vale per ogni $\mathbf{x} \in \mathbb{K}^n$ allora possiamo scrivere

$$\|\mathbf{A}+\mathbf{B}\| = \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\|(\mathbf{A}+\mathbf{B})\mathbf{x}\|_a}{\|\mathbf{x}\|_b} \leq \|\mathbf{A}\| + \|\mathbf{B}\|.$$

3. Sia $\lambda \in \mathbb{K}$ uno scalare arbitrario,

$$\begin{aligned}\|\lambda\mathbf{A}\| &= \sup_{\mathbf{y}\neq\mathbf{0}} \frac{\|\lambda\mathbf{A}\mathbf{y})\|_a}{\|\mathbf{y}\|_b} = \sup_{\mathbf{y}\neq\mathbf{0}} \frac{|\lambda|\,\|\mathbf{A}\mathbf{y}\|_a}{\|\mathbf{y}\|_b} = |\lambda| \sup_{\mathbf{y}\neq\mathbf{0}} \frac{\|\mathbf{A}\mathbf{y}\|_a}{\|\mathbf{y}\|_b},\\ &= |\lambda|\,\|\mathbf{A}\|.\end{aligned}$$

Osservazione 24. Consideriamo ora $\mathbf{A}, \mathbf{B} \in \mathbb{K}^{m\times m}$ e $\|\cdot\|_a = \|\cdot\|_b$, allora possiamo considerare la norma del prodotto $\mathbf{A}\mathbf{B}$ definita come segue

$$\|\mathbf{A}\mathbf{B}\| = \sup_{\mathbf{y}\neq\mathbf{0}} \frac{\|\mathbf{A}\mathbf{B}\mathbf{y}\|}{\|\mathbf{y}\|}, \tag{1.21}$$

da (1.21) possiamo dedurre

$$\frac{\|\mathbf{A}\mathbf{B}\mathbf{y}\|}{\|\mathbf{y}\|} \leq \frac{\|\mathbf{A}\|\,\|\mathbf{B}\mathbf{y}\|}{\|\mathbf{y}\|} = \|\mathbf{A}\| \frac{\|\mathbf{B}\mathbf{y}\|}{\|\mathbf{y}\|} \leq \|\mathbf{A}\|\,\|\mathbf{B}\|,$$

poiché questa diseguaglianza vale per ogni $\mathbf{y} \in \mathbb{K}^n$, ricaviamo che

$$\|\mathbf{A}\mathbf{B}\| \leq \|\mathbf{A}\|\,\|\mathbf{B}\|.$$

L'ultima osservazione suggerisce di introdurre la seguente definizione per le norme di matrice.

Definizione 47 (Definizione assiomatica di norma). Per ogni α scalare e $\mathbf{A}, \mathbf{B} \in \mathbb{K}^{m\times n}$ valgono

1. $\|\mathbf{A}\| \geq 0$, $\|\mathbf{A}\| = 0$ e solo se $\mathbf{A} = \mathbf{0}$.
2. $\|\alpha\mathbf{A}\| = |\alpha| \, \|\mathbf{A}\|$.
3. $\|\mathbf{A} + \mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\|$.
4. Se $m = n$ verrà imposta l'ulteriore proprietà $\|\mathbf{AB}\| \leq \|\mathbf{A}\| \, \|\mathbf{B}\|$.

1.4.3 Norme compatibili

Definizione 48. Una norma matriciale $\|\cdot\|$ è *compatibile* con le norme vettoriali $\|\cdot\|_a$ e $\|\cdot\|_b$ se per ogni vettore $\mathbf{x}$ e matrice $\mathbf{M}$ vale la relazione

$$\|\mathbf{Mx}\|_a \leq \|\mathbf{M}\| \, \|\mathbf{x}\|_b \, .$$

Casi importanti seguito) si hanno quando le norme $\|\cdot\|_a$ e $\|\cdot\|_b$ sono dello stesso tipo, ed in particolare quando ci riferiamo alle norme $\|\cdot\|_1$, $\|\cdot\|_2$ e $\|\cdot\|_\infty$. Le corrispondenti norme delle matrici vengono indicate con gli stessi simboli e valgono:

Teorema 56. *La norma matriciale* $\|\cdot\|_\infty : \mathbb{K}^{m\times n} \mapsto \mathbb{R}$ *definita da*

$$\sup_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{Ax}\|_\infty}{\|\mathbf{x}\|_\infty}, \tag{1.22}$$

vale

$$\|\mathbf{A}\|_\infty = \max_{i=1}^{m} \left(\sum_{j=1}^{n} |A_{ij}| \right) .$$

Dimostrazione. Osserviamo che se $\mathbf{A} \in \mathbb{K}^{m\times n}$ vale

$$\|\mathbf{Ax}\|_\infty = \max_{i=1}^{m} \left| \sum_{j=1}^{n} A_{ij} x_j \right| \leq \max_{i=1}^{m} \sum_{j=1}^{n} |A_{ij} x_j| \leq \max_{i=1}^{m} \left(\max_{j=n} |x_j| \right) \sum_{j=1}^{n} |A_{ij}| \, ,$$

$$\leq \|\mathbf{x}\|_\infty \max_{i=1}^{m} \sum_{j=1}^{n} |A_{ij}| \, ,$$

quindi

$$\frac{\|\mathbf{Ax}\|_\infty}{\|\mathbf{x}\|_\infty} \leq \max_{i=1}^{n} \sum_{j=1}^{m} |A_{ij}|\,,$$

sia ora k la riga per cui

$$\max_{i=1}^{m} \sum_{j=1}^{n} |A_{ij}| = \sum_{j=1}^{n} |A_{kj}|\,,$$

e sia $\mathbf{x}$ il vettore definito da

$$x_j = \begin{cases} +1 & \text{se } A_{kj} > 0 \\ -1 & \text{se } A_{kj} \leq 0 \end{cases},$$

allora avremo $\|\mathbf{x}\|_\infty = 1$ e

$$(\mathbf{Ax})_k = \sum_{j=1}^{n} A_{kj} x_j = \sum_{j=1}^{n} |A_{kj}| = \max_{i=1}^{m} \sum_{j=1}^{n} |A_{ij}|\,,$$

$$\Downarrow$$

$$\|\mathbf{Ax}\|_\infty \geq \max_{i=1}^{m} \sum_{j=1}^{n} |A_{ij}|\,,$$

e questo implica la (1.22). ■

In modo analogo si dimostra il seguente teorema:

Teorema 57.

$$\sup_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{Ax}\|_1}{\|\mathbf{x}\|_1},$$

vale

$$\|\mathbf{A}\|_1 = \max_{j=1}^{m} \left(\sum_{i=1}^{m} |A_{ij}| \right).$$

Definizione 49. La norma $\|\cdot\|_2 : \mathbb{K}^{m\times n} \mapsto \mathbb{R}$ è definita come segue

$$\|\mathbf{A}\|_2 = \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{Ax}\|_2}{\|\mathbf{x}\|_2}$$

Teorema 58. *Vale la seguente uguaglianza*

$$\|\mathbf{A}\|_2 = \sqrt{\varrho(\mathbf{A}^H\mathbf{A})}.$$

Dimostrazione. Osserviamo che

$$\|\mathbf{A}\|_2^2 = \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{A}\mathbf{x}\|_2^2}{\|\mathbf{x}\|_2^2} = \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\mathbf{x}^H\mathbf{A}^H\mathbf{A}\mathbf{x}}{\mathbf{x}^H\mathbf{x}} \tag{1.23}$$

La matrice $\mathbf{A}^H\mathbf{A}$ è una matrice hermitiana e quindi per il teorema 45 esiste una matrice unitaria $\mathbf{U}$ che la diagonalizza

$$\mathbf{U}^H\mathbf{A}^H\mathbf{A}\mathbf{U} = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix}, \tag{1.24}$$

dove $\lambda_1, \lambda_2, \ldots, \lambda_n$ sono gli autovalori della matrice $\mathbf{A}^H\mathbf{A}$. Dalla (1.23) segue

$$\begin{aligned} \|\mathbf{A}\|_2^2 &= \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\mathbf{x}^H\mathbf{U}\mathbf{U}^H\mathbf{A}^H\mathbf{A}\mathbf{U}\mathbf{U}^H\mathbf{x}}{\mathbf{x}^H\mathbf{U}\mathbf{U}^H\mathbf{x}} \\ &= \sup_{\mathbf{y}\neq\mathbf{0}} \frac{\mathbf{y}\mathbf{U}^H\mathbf{A}^H\mathbf{A}\mathbf{U}\mathbf{y}}{\mathbf{y}^H\mathbf{y}} \end{aligned}$$

dalla (1.24)

$$\begin{aligned} \|\mathbf{A}\|_2^2 &= \sup_{\mathbf{y}\neq\mathbf{0}} \frac{\sum_{i=1}^{n} \lambda_i y_i^2}{\sum_{i=1}^{n} y_i^2} \\ &= \max\{\lambda_1, \lambda_2, \ldots, \lambda_n\} \\ &= \varrho(\mathbf{A}^H\mathbf{A}) \end{aligned}$$

Osservazione 25. E' da notare che nel caso $n = m$ le matrici $\mathbf{A}^H\mathbf{A}$ e $\mathbf{A}\mathbf{A}^H$ sono simili, infatti

$$\mathbf{A}^{-H}(\mathbf{A}^H\mathbf{A})\mathbf{A}^H = \mathbf{A}\mathbf{A}^H$$

e quindi

$$\|\mathbf{A}\|_2 = \sqrt{\varrho(\mathbf{A}^H\mathbf{A})} = \sqrt{\varrho(\mathbf{A}\mathbf{A}^H)}. \tag{1.25}$$

Nel caso $n \neq m$ le matrici $\mathbf{A}\mathbf{A}^H$ e $\mathbf{A}^H\mathbf{A}$ non possono essere simili in quanto $\mathbf{A}\mathbf{A}^H \in \mathbb{K}^{n\times n}$ e $\mathbf{A}^H\mathbf{A} \in \mathbb{K}^{m\times m}$. Si può comunque provare che la (1.25) vale anche in questo caso.

CAPITOLO

DUE

RISOLUZIONE DI SISTEMI LINEARI: METODI DIRETTI

2.1 Eliminazione di Gauss

In metodo di eliminazione di Gauss[1] applicato ad un sistema lineare della forma

$$\begin{cases} A_{11}x_1 + A_{12}x_2 + \cdots + A_{1n}x_n = b_1 \\ A_{21}x_1 + A_{22}x_2 + \cdots + A_{2n}x_n = b_2 \\ \vdots \\ A_{n1}x_1 + A_{n2}x_2 + \cdots + A_{nn}x_n = b_n \end{cases}$$

lo trasforma in un sistema equivalente (cioè con la stessa soluzione) della forma

$$\begin{cases} C_{11}x_1 + C_{12}x_2 + \cdots + C_{1n}x_n = d_1 \\ C_{22}x_2 + \cdots + C_{2n}x_n = d_2 \\ \vdots \\ C_{nn}x_n = d_n \end{cases}$$

detta *forma triangolare*. Un sistema in forma triangolare si può risolvere immediatamente tramite *sostituzioni all'indietro*, cioè partendo dall'ultima incognita, il cui valore, si osservi, è già noto.

$$\begin{aligned} x_n &= \frac{d_n}{C_{nn}} \\ x_{n-1} &= \frac{d_{n-1} - C_{n-1n}x_n}{C_{n-1n-1}} \\ x_{n-2} &= \frac{d_{n-2} - C_{n-2n}x_n - C_{n-2n-1}x_{n-1}}{C_{n-2n-2}} \\ &\vdots \\ x_1 &= \frac{d_1 - C_{1n}x_n - C_{1n-1}x_{n-1} - \cdots - C_{12}x_2}{C_{11}} \end{aligned} \tag{2.1}$$

Esempio 31. La soluzione del sistema in forma triangolare:

$$\begin{cases} x + y + z = 1 \\ 4y + 2z = 0 \\ z = 2 \end{cases}$$

[1] Johann Carl Friedrich Gauss 1777–1855

(nel quale abbiamo posto $x = x_1$, $y = x_2$ e $z = x_3$), si ottiene immediatamente sostituendo all'indietro le variabili come mostrato in (2.1),

$$\begin{aligned} z &= \frac{2}{1} = 2, \\ y &= \frac{0 - 2z}{4} = \frac{0 - 2 \cdot 2}{4} = -1, \\ x &= \frac{1 - 1z - 1y}{1} = \frac{1 - 1(-1) - 1 \cdot 2}{1} = 0. \end{aligned}$$

L'algoritmo di eliminazione di Gauss procede formalmente esprimendo la variabile che si vuole eliminare come combinazione lineare delle variabili non ancora eliminate. A tal scopo, si utilizza una equazione del sistema, che successivamente non sarà piú presa in considerazione. L'eliminazione ha luogo per sostituzione della variabile da eliminare nelle rimanenti equazioni. L'algoritmo può essere espresso efficacemente, come vedremo, in termini matriciali poiché, nella pratica, si procede costruendo opportune combinazioni lineari delle equazioni del sistema, quindi di righe della matrice dei coefficienti.

Per fissare le idee, illustriamo il procedimento con un esempio dettagliato.

Esempio 32. Sia dato il seguente sistema lineare

$$\begin{cases} x + y + z + w = 1 & [1] \\ -x - y + 4z = 0 & [2] \\ x - z - w = 2 & [3] \\ x + w = 0 & [4] \end{cases}$$

in cui abbiamo numerato le equazioni a destra in parentesi quadra per facilità di riferimento. Trasformeremo questo sistema in forma triangolare eliminando di volta in volta le incognite x, y, z [2] per mezzo di opportune combinazioni lineari delle equazioni.

Per eliminare l'incognita x dalle equazioni $[2]$, $[3]$ e $[4]$ operiamo le seguenti trasformazioni:

- sostituiamo all'equazione $[2]$ l'equazione $[2'] \leftarrow [2] + [1]$;
- sostituiamo all'equazione $[3]$ l'equazione $[3'] \leftarrow [3] - [1]$;
- sostituiamo all'equazione $[4]$ l'equazione $[4'] \leftarrow [4] - [1]$;

[2] Perché non si elimina l'incognita w?

Si ottiene

$$\begin{cases} x+y+z+w = 1 & [1] \\ 5z+w = 1 & [2'] \leftarrow [2]+[1] \\ -y-2z-2w = 1 & [3'] \leftarrow [3]-[1] \\ -y-z = -1 & [4'] \leftarrow [4]-[1] \end{cases}$$

Osserviamo che l'equazione $[2']$ non contiene la variabile y mentre la equazione $[3']$ la contiene; conviene quindi scambiare l'equazione $[2']$ con la $[3']$, prima di procedere all'eliminazione successiva

$$\begin{cases} x+y+z+w = 1 & [1] \\ -y-2z-2w = 1 & [3'] \\ 5z+w = 1 & [2'] \\ -y-z = -1 & [4'] \end{cases}$$

Per eliminare l'incognita y dall'equazione $[4']$ operiamo la seguente trasformazione:

- sostituiamo all'equazione $[4']$ l'equazione $[4''] = \leftarrow [4'] - [3']$;

$$\begin{cases} x+y+z+w = 1 & [1] \\ -y-2z-2w = 1 & [3'] \\ 5z+w = 1 & [2'] \\ z+2w = -2 & [4''] \leftarrow [4']-[3'] \end{cases}$$

Per eliminare l'incognita z dall'equazione $[4'']$, operiamo la seguente trasformazione:

- sostituiamo all'equazione $[4'']$ l'equazione $[4'''] = \leftarrow [4''] - \frac{1}{5}[2']$;

$$\begin{cases} x+y+z+w = 1 & [1] \\ -y-2z-2w = 1 & [3'] \\ 5z+w = 1 & [2'] \\ \frac{9}{5}w = -\frac{11}{5} & [4'''] \leftarrow [4'']-\frac{1}{5}[2'] \end{cases}$$

Il sistema è ora in forma triangolare e tramite il procedimento all'indietro (2.1) possiamo calcolare la soluzione:

$$
\begin{aligned}
w &= \frac{-\frac{11}{5}}{\frac{9}{5}} = -\frac{11}{9} \\
z &= \frac{1 - 1\left(-\frac{11}{9}\right)}{5} = \frac{4}{9} \\
y &= \frac{1 + 2\left(-\frac{11}{9}\right) + 2\frac{4}{9}}{-1} = \frac{5}{9} \\
x &= \frac{1 - 1\left(-\frac{11}{9}\right) - 1\frac{4}{9} - 1\frac{5}{9}}{1} = \frac{11}{9}
\end{aligned}
$$

Osservazione 26. Per trasformare un sistema lineare di forma qualsiasi in forma triangolare sono sufficienti solo due tipi di operazioni:

— sommare ad una equazione un'altra equazione moltiplicata per un opportuno scalare;

— scambiare due equazioni.

Si noti che la soluzione di un sistema non cambia se scambiamo due equazioni oppure se sostituiamo ad una equazione la stessa equazione sommata ad un'altra equazione [3].

L' algoritmo di Gauss si può quindi descrivere come segue:

Algorithm *Metodo di eliminazione di Gauss*
Input: Dato il sistema lineare $\mathbf{A}^{(1)}\mathbf{x} = \mathbf{b}^{(1)}$

$$
\mathbf{A}^{(1)} = \begin{bmatrix} A_{11}^{(1)} & A_{12}^{(1)} & \cdots & A_{1n}^{(1)} \\ A_{21}^{(1)} & A_{22}^{(1)} & \ddots & \vdots \\ \vdots & \ddots & \ddots & A_{n-1n}^{(1)} \\ A_{n1}^{(1)} & \cdots & A_{nn-1}^{(1)} & A_{nn}^{(1)} \end{bmatrix}, \qquad \mathbf{x} = \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{bmatrix}, \qquad \mathbf{b}^{(1)} = \begin{bmatrix} b_0^{(1)} \\ b_1^{(1)} \\ \vdots \\ b_n^{(1)} \end{bmatrix},
$$

[3] Perché succede questo? Il lettore verifichi che è una conseguenza della linearità.

1. **for** $i \leftarrow 1$ **to** n-1
2. **do** ($*$ Sia k il primo intero con $k \geq i$ per cui $A_{ki}^{(i)} \neq 0$. $*$)
3. **if** $k \neq i$
4. **then** ($*$ scambio la $i-$esima equazione con la $k-$esima $*$)
5. ($*$ Costruisco il sistema $\mathbf{A}^{(i+1)}\mathbf{x} = \mathbf{b}^{(i+1)}$ equivalente al sistema $\mathbf{A}^{(i)}\mathbf{x} = \mathbf{b}^{(i)}$ come segue: Alla equazione $k-$esima con $k = i+1, \ldots, n$ sottraggo l'equazione $i-$esima moltiplicata per $A_{ki}^{(i)}/A_{ii}^{(i)}$ che porta alla seguenti formule: $*$)
6. **for** $k \leftarrow i+1$ **to** n
7. **do** $\beta \leftarrow A_{ki}^{(i)}/A_{ii}^{(i)}$
8. $b_k^{(i+1)} \leftarrow b_k^{(i)} - \beta b_i^{(i)}$.
9. **for** $j \leftarrow i+1$ **to** n
10. **do** $A_{kj}^{(i+1)} \leftarrow A_{kj}^{(i)} - \beta A_{ij}^{(i)}$
11. ($*$ Alla fine di queste operazioni otterrò il sistema equivalente $\mathbf{A}^{(n)}\mathbf{x} = \mathbf{b}^{(n)}$ dove

$$\mathbf{A}^{(n)} = \begin{bmatrix} A_{11}^{(1)} & A_{12}^{(1)} & A_{13}^{(1)} & \cdots & A_{1n}^{(1)} \\ 0 & A_{22}^{(2)} & A_{23}^{(2)} & \cdots & A_{2n}^{(2)} \\ 0 & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & & A_{n-1n-1}^{(n-1)} & A_{n-1n}^{(n-1)} \\ 0 & 0 & \cdots & 0 & A_{nn}^{(n)} \end{bmatrix}, \qquad \mathbf{b}^{(n)} = \begin{bmatrix} b_1^{(1)} \\ b_2^{(2)} \\ \vdots \\ b_{n-1}^{(n-1)} \\ b_n^{(n)} \end{bmatrix}.$$

A questo punto il sistema è in forma triangolare ed è facile risolverlo con le seguenti formule: $*$)

12. $x_n \leftarrow b_n/A_{nn}$
13. **for** $k \leftarrow n-1$ **downto** 1
14. **do** $\beta \leftarrow b_k$
15. **for** $j \leftarrow k+1$ **to** n
16. **do** $\beta \leftarrow \beta - A_{kj}x_j$
17. $x_k \leftarrow \beta/A_{kk}$
18. **return** $\mathbf{x}$

Osservazione 27. L'operazione di scambio di equazioni al fine di trovare $A_{ki}^{(i)} \neq 0$ nella linea 8 dell'algoritmo precedente può essere modificato considerando il seguente criterio alternativo

- Sia k tale che

$$\left|A_{ki}^{(i)}\right| \geq \left|A_{ji}^{(i)}\right|, \qquad \forall j \geq i.$$

  Se $k \neq i$ allora scambio la riga $i-$esima con la riga $k-$esima della matrice $\mathbf{A}^{(i)}$ e la componente $i-$esima con la componente $k-$esima del vettore $\mathbf{b}^{(k)}$.

Questo procedimento di scambio si chiama *pivoting*. Lo scopo del pivoting è di migliorare l'accuratezza della soluzione calcolata. Infatti si può mostrare che la divisione per un numero molto piccolo (in modulo) può essere estremamente inaccurata.

L' algoritmo di Gauss si può quindi descrivere anche come segue:

Algorithm *Metodo di eliminazione di Gauss (in loco)*
Input: La matrice dei coefficienti $\mathbf{A}$ e il vettore dei termini noti $\mathbf{b}$

```
1.  (* Eliminazione *)
2.  for i ← 1 to n-1
3.      do (* Pivoting: determinare k tale che |A_ki| ≥ |A_ji| con j ≥ i. *)
4.         k ← i
5.         for j ← i + 1 to n
6.             do if |A_ji| > |A_ki| then k ← j
7.         (* e scambiare la riga i−esima equazione con la k−esima. *)
8.         if k ≠ i
9.           then b_i ↔ b_k
10.                for j ← i to n
11.                    do A_ij ↔ A_kj
12.        (* Eliminazione *)
13.        for k ← i + 1 to n
14.            do β ← A_ki / A_ii
15.               b_k ← b_k − β b_i
16.               for j ← i + 1 to n
17.                   do A_kj ← A_kj − β A_ij
18. (* Sostituzione all'indietro *)
19. x_n ← b_n / A_nn
20. for k ← n − 1 downto 1
21.     do β ← b_k
22.        for j ← k + 1 to n
```

23. **do** $\beta \leftarrow \beta - A_{kj}x_j$
24. $x_k \leftarrow \beta / A_{kk}$
25. **return** $\mathbf{x}$

2.1.1 Forma matriciale del metodo di Gauss

Il metodo di Gauss per la risoluzione di sistemi lineari può essere interpretato utilizzando il formalismo delle matrici. Il sistema originario espresso in forma matriciale come $\mathbf{Ax} = \mathbf{b}$ si trasforma in un sistema della forma $\mathbf{Ux} = \mathbf{c}$ con la stessa soluzione $\mathbf{x}$, in cui $\mathbf{U}$ è una matrice triangolare superiore.

La trasformazione dal sistema $\mathbf{Ax} = \mathbf{b}$ nel sistema $\mathbf{Ux} = \mathbf{c}$ si ottiene tramite passaggi successivi che richiedono la moltiplicazione a sinistra per opportune matrici elementari dette *matrici di Frobenius* e per matrici di scambio di righe.

$$
\begin{aligned}
&\mathbf{T}_1\mathbf{A}\mathbf{x} = \mathbf{T}_1\mathbf{b}, && \Rightarrow && \mathbf{A}^{(2)}\mathbf{x} = \mathbf{b}^{(2)}\\
&\mathbf{T}_2\mathbf{A}^{(2)}\mathbf{x} = \mathbf{T}_2\mathbf{b}^{(2)}, && \Rightarrow && \mathbf{A}^{(3)}\mathbf{x} = \mathbf{b}^{(3)}\\
&\mathbf{T}_3\mathbf{A}^{(3)}\mathbf{x} = \mathbf{T}_3\mathbf{b}^{(3)}, && \Rightarrow && \mathbf{A}^{(4)}\mathbf{x} = \mathbf{b}^{(4)}\\
& && \vdots &&\\
&\mathbf{T}_{n-1}\mathbf{A}^{(n-1)}\mathbf{x} = \mathbf{T}_{n-1}\mathbf{b}^{(n-1)}, && \Rightarrow && \mathbf{A}^{(n)}\mathbf{x} = \mathbf{b}^{(n)}
\end{aligned}
$$

dove $\mathbf{U} = \mathbf{A}^{(n)}$ e $\mathbf{c} = \mathbf{b}^{(n)}$.

Definizione 50. Dato un vettore della forma $\mathbf{v} = [\,0, \ldots, 0, v_{k+1}, \ldots, v_n\,]^T$ cioè tale che

$$v_i = 0, \qquad i = 1, 2, \ldots, k$$

chiameremo matrice di Frobenius (Ferdinand Georg Frobenius 1849–1917) una matrice $\mathbf{F}$ della forma

$$
\mathbf{F} = \mathbf{I} + \mathbf{v}\mathbf{e}_k^T = \left[\begin{array}{ccc:c:ccc}
1 & & & 0 & 0 & \cdots & 0\\
& \ddots & & \vdots & \vdots & & \vdots\\
& & 1 & 0 & 0 & \cdots & 0\\
\hdashline
0 & \cdots & 0 & 1 & 0 & \cdots & 0\\
\hdashline
0 & \cdots & 0 & v_{k+1} & 1 & & \\
\vdots & & \vdots & \vdots & & \ddots & \\
0 & \cdots & 0 & v_n & & & 1
\end{array}\right].
$$

Osservazione 28. Data la matrice $\mathbf{A}$,

$$\mathbf{A} = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ \vdots & \vdots & & \vdots \\ A_{k1} & A_{k2} & \cdots & A_{kn} \\ A_{k+11} & A_{k+12} & \cdots & A_{k+1n} \\ \vdots & \vdots & & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nn} \end{bmatrix},$$

il prodotto $\mathbf{FA}$ è

$$\mathbf{FA} = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ \vdots & \vdots & & \vdots \\ A_{k1} & A_{k2} & \cdots & A_{kn} \\ A_{k+11} + v_{k+1}A_{k1} & A_{k+12} + v_{k+1}A_{k2} & \cdots & A_{k+1n} + v_{k+1}A_{kn} \\ \vdots & \vdots & & \vdots \\ A_{n1} + v_n A_{k1} & A_{n2} + v_n A_{k2} & \cdots & A_{nn} + v_n A_{kn} \end{bmatrix}.$$

L'effetto della moltiplicazione consiste quindi nel sostituire alla riga i-esima, della matrice $\mathbf{A}$, con $i \geq k+1$, la riga i-esima sommata alla riga k-esima moltiplicata per lo scalare v_i. Con la stessa simbologia utilizzata nella presentazione dell'esempio della sezione precedente, stiamo quindi modificando la matrice $\mathbf{A}$ come segue:

- sostituiamo all'equazione $[k+1]$ l'equazione $[(k+1)'] \leftarrow [(k+1)] + v_{k+1}[k]$;
- sostituiamo all'equazione $[k+2]$ l'equazione $[(k+2)'] \leftarrow [(k+2)] + v_{k+2}[k]$;
- ...
- sostituiamo all'equazione $[n]$ l'equazione $[n'] \leftarrow [n] + v_n[k]$;

Osservazione 29. L'inversa della matrice di Frobenius

$$\mathbf{F} = \mathbf{I} + \mathbf{v}\mathbf{e}_k^T$$

è la matrice

$$\mathbf{F}^{-1} = \mathbf{I} - \mathbf{v}\mathbf{e}_k^T.$$

Infatti

$$\left(\mathbf{I}+\mathbf{v}\mathbf{e}_k^T\right)\left(\mathbf{I}-\mathbf{v}\mathbf{e}_k^T\right)=\mathbf{I}+\mathbf{v}\mathbf{e}_k^T-\mathbf{v}\mathbf{e}_k^T+\mathbf{v}(\mathbf{e}_k^T\mathbf{v})\mathbf{e}_k^T=\mathbf{I}+v_k\mathbf{v}\mathbf{e}_k^T,$$

ed osservando che $\mathbf{e}_k^T\mathbf{v}=v_k=0$ ne consegue che

$$\left(\mathbf{I}+\mathbf{v}\mathbf{e}_k^T\right)\left(\mathbf{I}-\mathbf{v}\mathbf{e}_k^T\right)=\mathbf{I}.$$

Le matrici di Frobenius hanno la seguente notevole proprietà.

Lemma 59. *Siano* $\mathbf{F}_i=\mathbf{I}+\mathbf{v}_i\mathbf{e}_i^T$ *matrici di Frobenius, con* $\mathbf{v}_i=(0,\ldots,0,v_{i+1},v_{i+2},\ldots,v_n)^T$. *Allora per ogni* $2\geq k\leq n-1$[4] *si ha che*

$$\mathbf{F}_1\mathbf{F}_2\cdots\mathbf{F}_k=\mathbf{I}+\mathbf{v}_1\mathbf{e}_1^T+\mathbf{v}_2\mathbf{e}_2^T+\cdots+\mathbf{v}_k\mathbf{e}_k^T$$

Dimostrazione. La dimostrazione procede per induzione sull'indice k.

Al primo passo – $k=2$ – del ragionamento induttivo si ottiene

$$\begin{aligned}\mathbf{F}_1\mathbf{F}_2&=(\mathbf{I}+\mathbf{v}_1\mathbf{e}_1^T)(\mathbf{I}+\mathbf{v}_2\mathbf{e}_2^T),\\&=\mathbf{I}+\mathbf{v}_1\mathbf{e}_1^T+\mathbf{v}_2\mathbf{e}_2^T+\mathbf{v}_1(\mathbf{e}_1^T\mathbf{v}_2)\mathbf{e}_2^T,\\&=\mathbf{I}+\mathbf{v}_1\mathbf{e}_1^T+\mathbf{v}_2\mathbf{e}_2^T,\end{aligned}$$

poiché $\mathbf{e}_1^T\mathbf{v}_2=\mathbf{v}_2|_1=0$.

Supponiamo ora che il lemma sia vero per un indice generico $k-1$, cioè

$$\mathbf{F}_1\mathbf{F}_2\cdots\mathbf{F}_{k-1}=\mathbf{I}+\mathbf{v}_1\mathbf{e}_1^T+\mathbf{v}_2\mathbf{e}_2^T+\cdots+\mathbf{v}_{k-1}\mathbf{e}_{k-1}^T.$$

Allora possiamo scrivere

$$\begin{aligned}\left(\mathbf{F}_1\mathbf{F}_2\cdots\mathbf{F}_{k-1}\right)\mathbf{F}_k&=\left(\mathbf{I}+\mathbf{v}_1\mathbf{e}_1^T+\mathbf{v}_2\mathbf{e}_2^T+\cdots+\mathbf{v}_{k-1}\mathbf{e}_{k-1}^T\right)\left(\mathbf{I}+\mathbf{v}_k\mathbf{e}_k^T\right)\\&=\mathbf{I}+\mathbf{v}_1\mathbf{e}_1^T+\mathbf{v}_2\mathbf{e}_2^T+\cdots+\mathbf{v}_{k-1}\mathbf{e}_{k-1}^T+\mathbf{v}_k\mathbf{e}_k^T\\&\quad+\mathbf{v}_1\underbrace{(\mathbf{e}_1^T\mathbf{v}_k)}_{=0}\mathbf{e}_k^T+\cdots+\mathbf{v}_{k-1}\underbrace{(\mathbf{e}_{k-1}\mathbf{v}_k)}_{=0}\mathbf{e}_k^T,\\&=\mathbf{I}+\mathbf{v}_1\mathbf{e}_1^T+\mathbf{v}_2\mathbf{e}_2^T+\cdots+\mathbf{v}_{k-1}\mathbf{e}_{k-1}^T+\mathbf{v}_k\mathbf{e}_k^T,\end{aligned}$$

poiché tutti i termini "misti" sono nulli, essendo

$$\mathbf{e}_i^T\mathbf{v}_k=(\mathbf{v}_k)_i=0,\qquad i=1,2,\ldots,k-1.$$

[4] Cosa succederebbe se considerassimo anche indici $k\geq n$?

Definizione 51. Chiameremo matrice di scambio una matrice $\mathbf{S}_{ij}$ della forma:

$$\mathbf{S}_{ij} = \mathbf{I} - (\mathbf{e}_i - \mathbf{e}_j)(\mathbf{e}_i - \mathbf{e}_j)^T = \begin{bmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & & 0 & & 1 & \\ & & & & 1 & & \\ & & & & & \ddots & \\ & & & & & & 1 \\ & & & 1 & & 0 & \\ & & & & & & 1 \\ & & & & & & & \ddots \\ & & & & & & & & 1 \end{bmatrix}.$$

Osservazione 30. Data la matrice $\mathbf{A}$ dove $\mathbf{A}_{i\bullet}$ è la riga $i-$esima della matrice $\mathbf{A}$, l'effetto di $\mathbf{S}_{ij}\mathbf{A}$ è il seguente:

$$\mathbf{A} = \begin{bmatrix} \vdots \\ \mathbf{A}_{i\bullet} \\ \vdots \\ \mathbf{A}_{j\bullet} \\ \vdots \end{bmatrix}, \qquad \mathbf{SA} = \begin{bmatrix} \vdots \\ \mathbf{A}_{j\bullet} \\ \vdots \\ \mathbf{A}_{i\bullet} \\ \vdots \end{bmatrix},$$

cioè scambia la riga $i-$esima con la riga $j-$esima.

Osservazione 31. Scambiando due volte le stesse righe di una matrice riotteniamo la matrice iniziale. Quindi, se $\mathbf{S}_{ij}$ è una matrice di scambio allora $\mathbf{S}_{ij}\mathbf{S}_{ij} = \mathbf{I}$ cioè $\mathbf{S}_{ij}$ coincide con la sua inversa.

Osservazione 32. Il prodotto di due matrici di scambio è una matrice di permutazione, che modifica l'ordinamento delle righe di una matrice. Le matrici di permutazione sono matrici ortogonali, per cui la loro inversa coincide con la trasposta.

Le due operazioni fondamentali richieste dal procedimento di eliminazione di Gauss – somma ad una equazione di un'altra equazione moltiplicata per uno scalare e scambio di due equazioni – si possono esprimere tramite applicazioni successive di opportune matrici di Frobenius e matrici di scambio elementari.

2.1.2 Primo passo del metodo di Gauss

Dato il sistema $\mathbf{A}\mathbf{x} = \mathbf{b}$ poniamo $\mathbf{A}^{(1)} = \mathbf{A}$ e $\mathbf{b}^{(1)} = \mathbf{b}$,

$$\mathbf{A}^{(1)} = \begin{bmatrix} A_{11}^{(1)} & A_{12}^{(1)} & \cdots & A_{1n}^{(1)} \\ A_{21}^{(1)} & A_{22}^{(1)} & \ddots & \vdots \\ \vdots & \ddots & \ddots & A_{n-1n}^{(1)} \\ A_{n1}^{(1)} & \cdots & A_{nn-1}^{(1)} & A_{nn}^{(1)} \end{bmatrix}.$$

Supponendo $A_{11}^{(1)} \neq 0$ definiamo

$$v_i^{(1)} = \begin{cases} 0 & \text{se } i = 1 \\ \dfrac{A_{i1}^{(1)}}{A_{11}^{(1)}} & \text{se } i = 2, 3, \ldots, n \end{cases}, \qquad \mathbf{v}^{(1)} = \begin{bmatrix} 0 \\ \mathbf{v}_2^{(1)} \\ \mathbf{v}_3^{(1)} \\ \vdots \\ \mathbf{v}_n^{(1)} \end{bmatrix},$$

$$\mathbf{L}_1 = \mathbf{I} - \mathbf{v}^{(1)}\mathbf{e}_1^T = \begin{bmatrix} 1 & 0 & \cdots & & 0 \\ -v_2^{(1)} & 1 & & & \vdots \\ \vdots & & \ddots & & 0 \\ -v_{n-1}^{(1)} & & & 1 & 0 \\ -v_n^{(1)} & 0 & \cdots & 0 & 1 \end{bmatrix}.$$

Moltiplicando il sistema $\mathbf{A}^{(1)}\mathbf{x} = \mathbf{b}^{(1)}$ a sinistra per la matrice di Frobenius $\mathbf{L}_1$ otteniamo il sistema trasformato $\mathbf{A}^{(2)}\mathbf{x} = \mathbf{b}^{(2)}$ dove,

$$\mathbf{A}^{(2)} = \mathbf{L}_1\mathbf{A}^{(1)} = \left[\begin{array}{c|ccc} A_{11}^{(1)} & A_{12}^{(1)} & A_{13}^{(1)} & \cdots & A_{1n}^{(1)} \\ \hline 0 & A_{22}^{(2)} & A_{23}^{(2)} & \cdots & A_{2n}^{(2)} \\ 0 & A_{32}^{(2)} & A_{33}^{(2)} & \cdots & A_{3n}^{(2)} \\ \vdots & \vdots & & & \vdots \\ 0 & A_{n2}^{(2)} & A_{n3}^{(2)} & \cdots & A_{nn}^{(2)} \end{array}\right], \qquad \mathbf{b}^{(2)} = \mathbf{L}_1\mathbf{b}^{(1)} = \left[\begin{array}{c} b_1^{(1)} \\ \hline b_2^{(2)} \\ b_3^{(2)} \\ \vdots \\ b_n^{(2)} \end{array}\right],$$

ed inoltre si ha

$$\begin{aligned} A_{ij}^{(2)} &= A_{ij}^{(1)} - v_i^{(1)} A_{1j}^{(1)}, \qquad && i, j = 2, 3, \ldots, n \\ b_i^{(2)} &= b_i^{(1)} - v_i^{(1)} b_1^{(1)} && i = 2, 3, \ldots, n. \end{aligned}$$

2.1.3 $k-$esimo passo del metodo di Gauss

Supponiamo di aver effettuato $k-1$ passi dell'algoritmo di Gauss (senza permutazione di righe). Siamo quindi nella situazione $\mathbf{A}^{(k)} = \mathbf{L}_{k-1}\cdots\mathbf{L}_2\mathbf{L}_1\mathbf{A}^{(1)}$ e $\mathbf{b}^{(k)} = \mathbf{L}_{k-1}\cdots\mathbf{L}_2\mathbf{L}_1\mathbf{b}^{(1)}$ dove

$$\mathbf{A}^{(k)} = \left[\begin{array}{ccc:ccc} A_{11}^{(1)} & \cdots & A_{1k-1}^{(1)} & A_{1k}^{(1)} & \cdots & A_{1n}^{(1)} \\ & \ddots & \vdots & \vdots & & \vdots \\ 0 & & A_{k-1k-1}^{(k-1)} & A_{k-1k}^{(k-1)} & \cdots & A_{k-1n}^{(k-1)} \\ \hdashline 0 & \cdots & 0 & A_{kk}^{(k)} & \cdots & A_{kn}^{(k)} \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & A_{nk}^{(k)} & \cdots & A_{nn}^{(k)} \end{array}\right], \qquad \mathbf{b}^{(k)} = \left[\begin{array}{c} b_1^{(1)} \\ \vdots \\ b_{k-1}^{(k-1)} \\ \hdashline b_k^{(k)} \\ \vdots \\ b_n^{(k)} \end{array}\right].$$

Supponendo $A_{kk}^{(k)} \neq 0$ e ponendo

$$v_i^{(k)} = \begin{cases} 0 & \text{se } i = 1, 2, \ldots, k \\ \dfrac{A_{ik}^{(k)}}{A_{kk}^{(k)}} & \text{se } i = k+1, k+2, \ldots, n \end{cases}, \qquad \mathbf{v}^{(k)} = \left[\begin{array}{c} 0 \\ \vdots \\ 0 \\ \mathbf{v}_{k+1}^{(k)} \\ \vdots \\ \mathbf{v}_n^{(k)} \end{array}\right],$$

$$\mathbf{L}_k = \mathbf{I} - \mathbf{v}^k\mathbf{e}_k^T = \left[\begin{array}{ccc:c:ccc} 1 & & & 0 & 0 & \cdots & 0 \\ & \ddots & & \vdots & \vdots & & \vdots \\ & & 1 & 0 & 0 & \cdots & 0 \\ \hdashline 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ \hdashline 0 & \cdots & 0 & -v_{k+1}^{(k)} & 1 & & \\ \vdots & & \vdots & \vdots & & \ddots & \\ 0 & \cdots & 0 & -v_n^{(k)} & & & 1 \end{array}\right],$$

otteniamo

$$\mathbf{L}_k\mathbf{A}^{(k)} = \left[\begin{array}{ccc:ccc} A_{11}^{(1)} & \cdots & A_{1k-1}^{(1)} & A_{1k}^{(1)} & A_{1k+1}^{(1)} & \cdots & A_{1n}^{(1)} \\ & \ddots & \vdots & \vdots & \vdots & & \vdots \\ 0 & & A_{k-1k-1}^{(k-1)} & A_{k-1k}^{(k-1)} & A_{k-1k+1}^{(k-1)} & \cdots & A_{k-1n}^{(k-1)} \\ \hdashline 0 & \cdots & 0 & A_{kk}^{(k)} & A_{kk+1}^{(k)} & \cdots & A_{kn}^{(k)} \\ 0 & \cdots & 0 & 0 & A_{k+1k+1}^{(k+1)} & \cdots & A_{k+1n}^{(k+1)} \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & 0 & A_{nk+1}^{(k+1)} & \cdots & A_{nn}^{(k+1)} \end{array}\right], \qquad \mathbf{L}_k\mathbf{b}^{(k)} = \left[\begin{array}{c} b_1^{(1)} \\ \vdots \\ b_{k-1}^{(k)} \\ \hdashline b_k^{(k)} \\ \hdashline b_{k+1}^{(k+1)} \\ \vdots \\ b_n^{(k+1)} \end{array}\right],$$

dove

$$\begin{aligned} A_{ij}^{(k+1)} &= A_{ij}^{(k)} - v_i^{(k)} A_{kj}^{(k)}, \qquad & i,j = k+1, k+2, \ldots, n \\ b_i^{(k+1)} &= b_i^{(k)} - v_i^{(k)} b_i^{(k)} & i = k+1, k+2, \ldots, n. \end{aligned}$$

Come si vede l'algoritmo di Gauss ripete lo stesso procedimento ad ogni passo a matrici sempre più piccole fino ad arrivare alla triangolarizzazione della matrice originale.

Osservazione 33. Consideriamo ora una matrice $\mathbf{A}$ di dimensione $n \times n$ e supponiamo che ogni passo dell'algoritmo di Gauss sia applicabile, cioè $A_{kk}^{(k)} \neq 0$ per ogni k. Dopo $n-1$ passi dell'algoritmo di Gauss otteniamo:

$$\mathbf{L}_{n-1}\mathbf{L}_{n-2}\cdots\mathbf{L}_1\mathbf{A} = \mathbf{U},$$

chiamando

$$\mathbf{L} = \mathbf{L}_1^{-1}\mathbf{L}_2^{-1}\mathbf{L}_3^{-1}\cdots\mathbf{L}_{n-1}^{-1},$$

allora

$$\mathbf{A} = \mathbf{L}\mathbf{U}.$$

Inoltre

① Le matrici $\mathbf{L}_k = \mathbf{I} - \mathbf{v}_k\mathbf{e}_k^T$ sono matrici di Frobenius.

② Le matrici $\mathbf{L}_k^{-1} = \mathbf{I} + \mathbf{v}_k\mathbf{e}_k^T$ sono ancora matrici di Frobenius.

③ Applicando il lemma 59

$$\begin{aligned}\mathbf{L} &= \mathbf{L}_1^{-1}\mathbf{L}_2^{-1}\cdots\mathbf{L}_{n-1}^{-1},\\ &= (\mathbf{I}+\mathbf{v}_1\mathbf{e}_1^T)(\mathbf{I}+\mathbf{v}_2\mathbf{e}_2^T)\cdots(\mathbf{I}+\mathbf{v}_{n-1}\mathbf{e}_{n-1}^T),\\ &= \mathbf{I}+\mathbf{v}_1\mathbf{e}_1^T+\mathbf{v}_2\mathbf{e}_2^T+\cdots\mathbf{v}_{n-2}\mathbf{e}_{n-2}^T+\mathbf{v}_{n-1}\mathbf{e}_{n-1}^T.\end{aligned}$$

Si vede che $\mathbf{L}$ è una matrice triangolare inferiore.

La matrice $\mathbf{L}$ è una matrice triangolare inferiore, quindi la matrice $\mathbf{A}$ è stata decomposta nel prodotto di una matrice triangolare inferiore $\mathbf{L}$ e una matrice triangolare superiore $\mathbf{U}$. Tale decomposizione prende il nome di *decomposizione* o *fattorizzazione LU*.

Osservazione 34. Dalla precedente osservazione vediamo che la decomposizione LU della matrice $\mathbf{A}$ può essere scritta come segue

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0\\ v_2^{(1)} & 1 & 0 & \cdots & 0\\ v_3^{(1)} & v_3^{(2)} & \ddots & \ddots & \vdots\\ \vdots & \vdots & & 1 & 0\\ v_n^{(1)} & v_n^{(2)} & \cdots & v_n^{(n-1)} & 1 \end{bmatrix}, \qquad v_i^{(j)} = \frac{A_{ij}^{(j)}}{A_{jj}^{(j)}}$$

$$\mathbf{U} = \begin{bmatrix} A_{11}^{(1)} & A_{12}^{(1)} & A_{13}^{(1)} & \cdots & A_{1n}^{(1)}\\ 0 & A_{22}^{(2)} & A_{23}^{(2)} & \cdots & A_{2n}^{(2)}\\ 0 & 0 & \ddots & \ddots & \vdots\\ \vdots & \vdots & & A_{n-1n-1}^{(n-1)} & A_{n-1n}^{(n-1)}\\ 0 & 0 & \cdots & 0 & A_{nn}^{(n)} \end{bmatrix},$$

è quindi chiaro che si possono memorizzare sia la matrice $\mathbf{L}$ che la matrice $\mathbf{U}$ in una unica matrice (la diagonale di $\mathbf{L}$ non viene memorizzata). Analizzando il metodo di Gauss si vede che le matrici $\mathbf{L}$ e $\mathbf{U}$ possono essere memorizzate nella matrice originale $\mathbf{A}$. Questo permette di scrivere il seguente algoritmo che calcola la decomposizione LU di una matrice $\mathbf{A}$ e salva il risultato nella matrice stessa.

Algorithm *Decomposizione LU*

Input: La matrice $\mathbf{A}$
Output: La matrice $\mathbf{A}$ contenente i fattori della decomposizione LU

```
1.  for i ← 1 to n-1
2.      do for k ← i + 1 to n
3.          do β ← A_ki / A_ii
4.             for j ← i + 1 to n
5.                 do A_kj ← A_kj − β A_ij
6.             A_ki ← β
```

2.1.4 Algoritmo di Gauss in presenza di pivoting

Lemma 60. *Sia $\mathbf{F} = \mathbf{I} - \mathbf{v}\mathbf{e}_k^T$ una matrice di Frobenius e $\mathbf{S}_{ij}$ una matrice di scambio, se $i, j > k$ allora la matrice $\mathbf{F}\mathbf{S}_{ij}$ si può scrivere come $\mathbf{S}_{ij}\mathbf{F}'$ dove $\mathbf{F}'$ è ancora una matrice di Frobenius e vale*

$$\mathbf{F}' = \mathbf{I} - \mathbf{w}\mathbf{e}_k^T, \qquad \mathbf{w} = \mathbf{S}_{ij}\mathbf{v}.$$

Dimostrazione. Il risultato è conseguenza immediata dei seguenti passaggi

$$\begin{aligned}
\mathbf{F}\mathbf{S}_{ij} &= \left(\mathbf{I} - \mathbf{v}\mathbf{e}_k^T\right)\mathbf{S}_{ij}, \\
&= \mathbf{S}_{ij} - \mathbf{v}\mathbf{e}_k^T\mathbf{S}_{ij}, \\
&= \mathbf{S}_{ij} - \mathbf{v}\mathbf{e}_k^T\left(\mathbf{I} - (\mathbf{e}_i - \mathbf{e}_j)(\mathbf{e}_i - \mathbf{e}_j)^T\right), \\
&= \mathbf{S}_{ij} - \mathbf{v}\mathbf{e}_k^T + \mathbf{v}\mathbf{e}_k^T(\mathbf{e}_i - \mathbf{e}_j)(\mathbf{e}_i - \mathbf{e}_j)^T, \\
&= \mathbf{S}_{ij} - \mathbf{v}\mathbf{e}_k^T,
\end{aligned}
\qquad
\begin{aligned}
\mathbf{S}_{ij}\mathbf{F}' &= \mathbf{S}_{ij}\left(\mathbf{I} - \mathbf{w}\mathbf{e}_k^T\right), \\
&= \mathbf{S}_{ij} - \mathbf{S}_{ij}\mathbf{w}\mathbf{e}_k^T, \\
&= \mathbf{S}_{ij} - \mathbf{S}_{ij}\mathbf{S}_{ij}\mathbf{v}\mathbf{e}_k^T, \\
&= \mathbf{S}_{ij} - \mathbf{v}\mathbf{e}_k^T,
\end{aligned}$$

dove si è anche sfruttato il fatto che il vettore $\mathbf{e}_k$ è ortogonale al vettore $\mathbf{e}_i - \mathbf{e}_j$. ■

L'algoritmo di Gauss con pivoting può quindi essere messo in questa forma:

$$\mathbf{L}_{n-1}\mathbf{P}_{n-1}\mathbf{L}_{n-2}\mathbf{P}_{n-2}\cdots\mathbf{L}_2\mathbf{P}_2\mathbf{L}_1\mathbf{P}_1\mathbf{A} = \mathbf{U}, \tag{2.2}$$

dove $\mathbf{P}_k$ è una matrice di scambio o la matrice identità a seconda che sia necessario o meno scambiare delle righe al k esimo passo dell'algoritmo di Gauss. Osserviamo che $\mathbf{P}_k = \mathbf{S}_{ij}$ per $i, j \geq k$ e quindi applicando il lemma 60 otteniamo $\mathbf{P}_k\mathbf{L}_{k-1} = \mathbf{L}'_{k-1}\mathbf{P}_k$. Possiamo quindi spostare i prodotti per le matrici di scambio a destra nella (2.2) ottenendo

$$\left(\mathbf{L}'_{n-1}\mathbf{L}'_{n-2}\cdots\mathbf{L}'_2\mathbf{L}'_1\right)\left(\mathbf{P}_{n-1}\mathbf{P}_{n-2}\cdots\mathbf{P}_2\mathbf{P}_1\right)\mathbf{A} = \mathbf{U},$$

dove la matrice $\mathbf{U}$ è triangolare superiore ed abbiamo introdotto le $n-1$ matrici di Frobenius per $k = 1, 2, \ldots, n-1$

$$\mathbf{L}'_k = \mathbf{I} - \mathbf{w}_k \mathbf{e}_k^T, \qquad \mathbf{w}_k = \mathbf{P}_k \cdots \mathbf{P}_2 \mathbf{P}_1 \mathbf{v}_k.$$

Ponendo dunque

$$\mathbf{L} = (\mathbf{L}'_1)^{-1}(\mathbf{L}'_2)^{-1}(\mathbf{L}'_3)^{-1} \cdots (\mathbf{L}'_{n-1})^{-1}, \qquad \mathbf{P} = \mathbf{P}_{n-1}\mathbf{P}_{n-2} \cdots \mathbf{P}_2 \mathbf{P}_1,$$

otteniamo

$$\mathbf{PA} = \mathbf{LU},$$

cioè l'algoritmo di eliminazione consiste nella decomposizione LU della matrice $\mathbf{PA}$, che risulta applicando alla matrice $\mathbf{A}$ la sequenza di scambi di righe della permutazione specificata dalla matrice $\mathbf{P}$.

Osservazione 35. Si è illustrato il funzionamento dell'algoritmo di fattorizzazione con pivoting considerando la possibilità di scambi di righe, detto *pivot parziale*, che corrisponde sul sistema lineare a scambiare tra loro le equazioni, cioè a numerarle in ordine diverso.

Si può introdurre un pivoting piú generale, detto *pivot totale*, che ammetta sia scambi di righe che di colonne. Lo scambio di colonne corrisponde a numerare diversamente le variabili del sistema lineare.

L'argomento sviluppato a proposito della fattorizzazione *LU* con pivot parziale si generalizza facilmente al caso del pivot totale. e si ottiene una decomposizione che formalmente si scrive così:

$$\mathbf{PAQ} = \mathbf{LU}. \tag{2.3}$$

In (2.3) $\mathbf{P}$ è la matrice di permutazione delle righe, ottenuta come prodotto delle matrici di scambio elementari delle righe e QQ è la matrice di permutazione delle colonne ottenuta come prodotto delle matrici elementari di scambio delle colonne.

Algorithm *Decomposizione LU con pivoting*
Input: La matrice $\mathbf{A}$
Output: La matrice $\mathbf{A}$ contenente la decomposizione LU e il vettore $\mathbf{p}$ contenente la permutazione
1. ($*$ Inizializza il vettore della permutazione $*$)

2. **for** $i \leftarrow 1$ **to** n
3. **do** $p_i \leftarrow i$
4. **for** $i \leftarrow 1$ **to** n-1
5. **do** $i_p \leftarrow p_i$
6. ($*$ Pivoting: trovo k tale che $\left|A_{k_p i}\right| \geq |A_{ji}|$ con $j \geq i$. $*$)
7. $k_p \leftarrow i_p$
8. **for** $j \leftarrow i+1$ **to** n
9. **do** $j_p \leftarrow p_j$
10. **if** $|A_{j_p i}| > \left|A_{k_p i}\right|$
11. **then** $k \leftarrow j$
12. $k_p \leftarrow j_p$
13. ($*$ scambio la riga $i-$esima equazione con la $k-$esima. $*$)
14. $p_k \leftrightarrow p_i$
15. ($*$ Eliminazione $*$)
16. **for** $k \leftarrow i+1$ **to** n
17. **do** $k_p \leftarrow p_k$
18. $\beta \leftarrow A_{k_p i}/A_{i_p i}$
19. **for** $j \leftarrow i+1$ **to** n
20. **do** $A_{k_p j} \leftarrow A_{k_p j} - \beta A_{i_p j}$
21. $A_{k_p i} \leftarrow \beta$
22. **return** $\mathbf{A}$, $\mathbf{p}$

Osservazione 36 (Pivot numerico). Se l'elemento pivotale è non nullo, in teoria non sarebbe necessario operare uno scambio di righe. Tuttavia, se l'elemento pivotale è molto piccolo in valore assoluto, l'algoritmo di fattorizzazione tenderà a produrre dei complementi di Schur molto grandi, e questo potrebbe originare problemi di stabilità numerica. Nella pratica si controlla questo effetto scegliendo comunque l'elemento pivotale in valore assoluto maggiore nella colonna – *pivot parziale* – o nella sottomatrice – *pivot totale* – formate dagli elementi che devono ancora essere fattorizzati, quando l'elemento pivotale corrente ha un valore inferiore ad una soglia prefissata. E' possibile dimostrare che con questa tecnica, detta *pivot numerico*, si garantisce la stabilità numerica dell'algoritmo di fattorizzazione nel senso sopra esposto.

Osservazione 37 (Pivot simbolico). Esiste un'altra ragione molto importante per operare una scelta di pivot diversa da quella legata al valore dell'elemento pivotale, che accenniamo brevemente per completezza. Nella pratica si ha spesso a che fare con

matrici *sparse*, cioè con matrici che hanno un numero di elementi non nulli dell'ordine della dimensione della matrice, e non del suo quadrato. E' importante mantenere la sparsità della matrice durante il processo di fattorizzazione. Infatti, il meccanismo del complemento di Schur tende a produrre non-zeri, e se il numero di non-zeri che vengono prodotti diventa enorme la memoria del calcolatore potrebbe non essere sufficiente per memorizzare i fattori $\mathbf{L}$ ed $\mathbf{U}$. Un semplice scambio di righe, che equivale a scambiare le equazioni del sistema, permette di controllare la formazione dei non-zeri. Mostriamo questo effetto con il seguente esempio.

Esempio 33. Consideriamo la fattorizzazione LU della matrice $\mathbf{A} \in \mathbb{R}^{10\times 10}$

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

e della matrice che si ottiene prendendo righe e colonne nell'ordine inverso, cioè

$$\mathbf{PAP} = \begin{bmatrix} \mathbf{1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{1} \\ 0 & \mathbf{1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{1} \\ 0 & 0 & \mathbf{1} & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{1} \\ 0 & 0 & 0 & \mathbf{1} & 0 & 0 & 0 & 0 & 0 & \mathbf{1} \\ 0 & 0 & 0 & 0 & \mathbf{1} & 0 & 0 & 0 & 0 & \mathbf{1} \\ 0 & 0 & 0 & 0 & 0 & \mathbf{1} & 0 & 0 & 0 & \mathbf{1} \\ 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{1} & 0 & 0 & \mathbf{1} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{1} & 0 & \mathbf{1} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{1} & \mathbf{1} \\ \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} \end{bmatrix}$$

Abbiamo fissato per gli elementi di $\mathbf{A}$ diversi da zero il valore 1, ma l'argomento che desideriamo illustrare riguarda solo il processo di formazione dei non-zeri durante la fattorizzazione, noto col termine inglese di *fill-in*, e non il valore assunto da essi assunto nei fattori di Gauss $\mathbf{L}$ e $\mathbf{U}$.

Nella pagina successiva mostriamo la struttura dei non-zeri della matrice $\mathbf{A}$ e della matrice permutata simmetricamente, $\mathbf{PAP}$, e la struttura dei non-zeri dei loro rispettivi fattori triangolare superiore ed inferiore prodotti dall'algoritmo di Gauss, dove supponiamo che non sia necessario fare un pivot numerico.

Si noti che nel caso della matrice $\mathbf{A}$ presa nella forma originale i fattori di Gauss sono *pieni*, quindi abbiamo un riempimento totale.

Nella forma permutata invece, non si producono non-zeri, e ricordando che la diagonale di $\mathbf{L}$ è formata tutta da 1, e quindi può non essere memorizzata, non si richiede ulteriore memoria al calcolatore oltre a quella necessaria per memorizzare gli elementi non-zero della matrice.

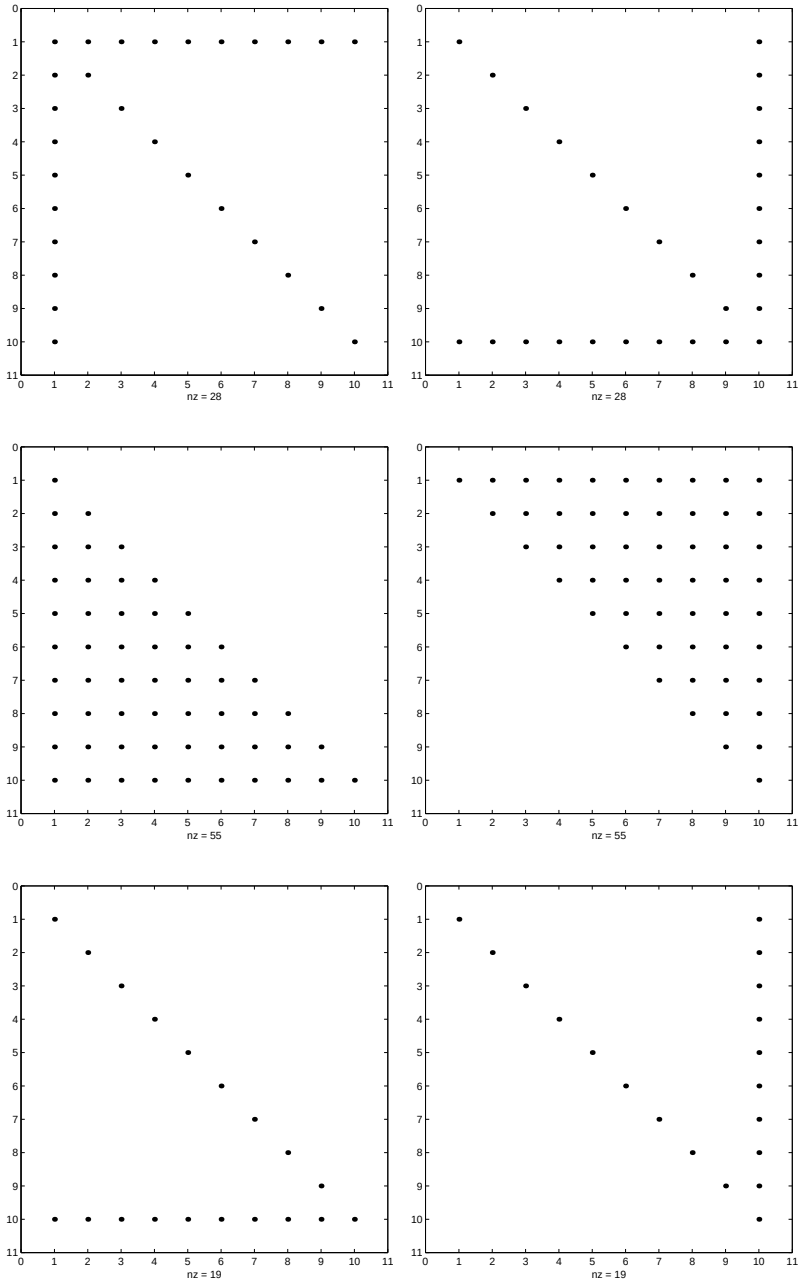


Figura 2.1: In alto mostriamo la struttura dei non zeri di $\mathbf{A}$ e di $\mathbf{PAP}$, in mezzo quella dei fattori di $\mathbf{A}$ ed in basso quella dei fattori di $\mathbf{PAP}$.

2.2 Fattorizzazione di Cholesky

Sia $\mathbf{A} \in \mathbb{R}^{n\times n}$ non-singolare e $\mathbf{b} \in \mathbb{R}^n$ e vogliamo determinare $\mathbf{x} \in \mathbb{R}^n$ tale che

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

Sappiamo che si può procedere per mezzo della fattorizzazione di Gauss $\mathbf{A} = \mathbf{L}\mathbf{U}$ – dove supponiamo per semplicità di esposizione che non sia necessario fare pivoting – risolvendo i due sistemi lineari con matrici triangolare inferiore $\mathbf{L}$ e superiore $\mathbf{U}$

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

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

Il costo computazionale di questo procedimento è non solo il costo della soluzione dei due sistemi, ma anche e soprattutto il costo necessario per determinare $\mathbf{L}$ e $\mathbf{U}$. Inoltre il calcolatore deve avere una capacità di memoria sufficiente per memorizzare entrambi i fattori. Ciò non produce normalmente problemi se la matrice $\mathbf{A}$ è "piena", dato che si sfrutta lo stesso spazio di memoria già occupato per la matrice. Se la matrice è "sparsa", a causa del processo di formazione di non-zeri di cui si è discusso nella sezione precedente, la memoria del calcolatore potrebbe non essere sufficiente per memorizzare entrambi i fattori.

Per queste ragioni è interessante capire quando è possibile una fattorizzazione del tipo

$$\mathbf{A} = \mathbf{R}^T\mathbf{R}$$

con $\mathbf{R}$ matrice triangolare superiore, che chiameremo *fattorizzazione di Cholesky*.

Poichè si richiede di calcolare un solo fattore invece che due, è sensato aspettarsi sia un costo computazionale che un'occupazione di memoria forse non dimezzati ma sicuramente inferiori[5].

Il seguente teorema caratterizza le matrici per cui esiste la fattorizzazione di Cholesky.

Teorema 61. *Le matrici quadrate non-singolari $\mathbf{A}$ per cui è possibile una fattorizzazione della forma $\mathbf{A} = \mathbf{R}^T\mathbf{R}$ con $\mathbf{R}$ matrice triangolare superiore sono* tutte e sole *le matrici* simmetriche e definite positive*, (*SPD*).*

Dimostrazione. Mostriamo innanzitutto la *necessità*, cioè che la fattorizzazione di Cholesky è possibile solo per matrici simmetriche e definite positive. Mostreremo in seguito la *sufficienza*, cioè che data una matrice che ha tali proprietà è sempre possibile calcolare il suo *fattore di Cholesky* $\mathbf{R}$.

[5] Il gioco vale la candela!

Necessità

Sia $\mathbf{A}, \mathbf{R} \in \mathbb{R}^{n\times n}$ con $\mathbf{A}$ non singolare ed $\mathbf{R}$ triangolare superiore tale che valga la decomposizione $\mathbf{A} = \mathbf{R}^T\mathbf{R}$.

- **Simmetria**

$$\mathbf{A}^T = \left(\mathbf{R}^T\mathbf{R}\right)^T = \mathbf{R}^T\mathbf{R} = \mathbf{A}.$$

- **Positività**

  Se $\mathbf{A}$ è non-singolare, allora anche $\mathbf{R}$ è non-singolare perché dal teorema di Binet segue $|\,\mathbf{A}\,| = |\,\mathbf{R}^T\mathbf{R}\,| = |\,\mathbf{R}^T\,|\,|\,\mathbf{R}\,| = |\,\mathbf{R}\,|^2$.

  Inoltre, se $\mathbf{R}$ è non-singolare, si ha $\mathbf{R}\mathbf{x} \neq \mathbf{0}$ per ogni vettore $\mathbf{x} \neq \mathbf{0}$. Quindi,

$$\mathbf{x}^T\mathbf{A}\mathbf{x} = \mathbf{x}^T\mathbf{R}^T\mathbf{R}\mathbf{x} = \|\mathbf{R}\mathbf{x}\|_2^2 > 0$$

Sufficienza

La dimostrazione procede per induzione sull'ordine della matrice, ma è anche in parte costruttiva. Illustreremo un algorimo di riduzione, che permette di ricondurre il problema della fattorizzazione di una matrice *SPD* di ordine n a quello di una matrice *SPD* di ordine $n-1$. Per induzione, si troverà che tutte le matrici *SPD* sono fattorizzabili.

Per facilitare la costruzione dell'algoritmo di riduzione, riscriviamo la matrice simmetrica $\mathbf{A} \in \mathbb{R}^{n\times n}$ in una forma equivalente a blocchi

$$\mathbf{A} = \left[\begin{array}{c:ccc} A_{11} & A_{12} & \cdots & A_{1n} \\ \hdashline A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nn} \end{array}\right] = \left[\begin{array}{c:c} \alpha & \mathbf{a}^T \\ \hdashline \mathbf{a} & \tilde{\mathbf{A}} \end{array}\right],$$

introducendo la sottomatrice $\tilde{\mathbf{A}} \in \mathbb{R}^{(n-1)\times(n-1)}$, il vettore $\mathbf{a} \in \mathbb{R}^{n-1}$ e lo scalare $\alpha \in \mathbb{R}^n$

$$\tilde{\mathbf{A}} = \begin{bmatrix} A_{22} & \cdots & A_{2n} \\ \vdots & \ddots & \vdots \\ A_{n2} & \cdots & A_{nn} \end{bmatrix}, \qquad \mathbf{a} = \begin{bmatrix} A_{21} \\ \vdots \\ A_{n1} \end{bmatrix}, \qquad \alpha = A_{11}.$$

Scriviamo la matrice $\mathbf{R}$ a blocchi in modo simile ad $\mathbf{A}$

$$\mathbf{R} = \left[\begin{array}{c:c} \rho & \mathbf{r}^T \\ \hdashline \mathbf{0} & \mathbf{R}' \end{array}\right],$$

e sviluppiamo il prodotto $\mathbf{R}^T\mathbf{R}$

$$\mathbf{R}^T\mathbf{R} = \begin{bmatrix} \rho & \mathbf{0}^T \\ \mathbf{r} & \mathbf{R}'^T \end{bmatrix} \begin{bmatrix} \rho & \mathbf{r}^T \\ \mathbf{0} & \mathbf{R}' \end{bmatrix} = \begin{bmatrix} \rho^2 & \rho\mathbf{r}^T \\ \rho\mathbf{r} & \mathbf{r}\mathbf{r}^T + \mathbf{R}'^T\mathbf{R}' \end{bmatrix}.$$

La fattorizzazione $\mathbf{A} = \mathbf{R}^T\mathbf{R}$ si può esprimere uguagliando i blocchi nelle posizioni corrispondenti di $\mathbf{R}^T\mathbf{R}$ ed $\mathbf{A}$

$$\begin{bmatrix} \alpha & \mathbf{a}^T \\ \mathbf{a} & \tilde{\mathbf{A}} \end{bmatrix} = \begin{bmatrix} \rho^2 & \rho\mathbf{r}^T \\ \rho\mathbf{r} & \mathbf{r}\mathbf{r}^T + \mathbf{R}'^T\mathbf{R}' \end{bmatrix},$$

per cui si ottengono immediatamente le relazioni seguenti

$$\begin{aligned} &\alpha = \rho^2, \\ &\mathbf{a} = \rho\mathbf{r}, \qquad (\text{o equivalentemente } \mathbf{a}^T = \rho\mathbf{r}^T), \\ &\tilde{\mathbf{A}} = \mathbf{R}'^T\mathbf{R}' + \mathbf{r}\mathbf{r}^T. \end{aligned} \tag{2.4}$$

Le prime due relazioni forniscono una espressione diretta per determinare ρ ed $\mathbf{r}$, cioè

$$\rho = \sqrt{\alpha}, \qquad \text{e} \qquad \mathbf{r} = \frac{1}{\sqrt{\alpha}}\mathbf{a},$$

dove supponiamo per il momento[6] che α sia un numero reale strettamente positivo. L'ultima relazione in (2.4), riscritta come

$$\mathbf{R}'^T\mathbf{R}' = \tilde{\mathbf{A}} - \frac{1}{\alpha}\mathbf{a}\mathbf{a}^T,$$

suggerisce che la matrice $\tilde{\mathbf{R}}$ sia a sua volta il fattore triangolare superiore – ammesso ovviamente che esista – della fattorizzazione di Cholesky della matrice "ridotta" $\mathbf{A}' \in \mathbb{R}^{(n-1)\times(n-1)}$

$$\mathbf{A}' = \tilde{\mathbf{A}} - \frac{1}{\alpha}\mathbf{a}\mathbf{a}^T.$$

[6] Se α fosse strettamente negativo, il ragionamento che segue avrebbe ancora senso ammettendo di calcolare la radice quadrata in campo complesso. Questo tuttavia non produrrebbe una algoritmo numerico efficiente. Se α fosse nullo, il ragionamento che segue non sarebbe possibile. Il problema, in realtà, non si pone, perché dimostreremo tra un attimo che l'ipotesi *SPD* assicura la positività stretta di questo scalare.

Procedendo in questo modo potremmo scrivere il fattore triangolare superiore $\mathbf{R}$ di Cholesky come

$$\mathbf{R} = \begin{bmatrix} \sqrt{\alpha} & \frac{1}{\sqrt{\alpha}}\mathbf{a}^T \\ \mathbf{0} & \mathbf{R}' \end{bmatrix},$$

in cui la prima riga sarebbe completamente determinata e resterebbe da calcolare il fattore triangolare superiore $\mathbf{R}'$ della matrice $\mathbf{A}'$ di ordine $n-1$. Di fatto, avremmo ridotto di uno la dimensione del problema ed è evidente che se si potesse ripetere questo procedimento n volte, arriveremmo alla fine ad una matrice di ordine 1 la cui fattorizzazione sarebbe banalmente

$$[\, A_{11} \,] = [\, \sqrt{A_{11}} \,][\, \sqrt{A_{11}} \,].$$

Mostreremo che l'ipotesi che $\mathbf{A}$ sia *SPD* garantisce che tutto il procedimento ipotizzato è in realtà ben definito. Innanzitutto, osserviamo che valgono le seguenti proprietà.

(a) **Se $\mathbf{A}$ è SPD, allora $\boldsymbol{\alpha > 0}$.**
Il risultato si ottiene applicando la definizione di matrice definita positiva ad $\mathbf{e}_1$, primo vettore della base canonica:

$$\alpha = A_{11} = \mathbf{e}_1^T \mathbf{A} \mathbf{e}_1 > 0.$$

(b) **Se $\mathbf{A}$ è SPD, allora $\mathbf{A}'$ è SPD**.
La simmetria di $\mathbf{A}'$ è evidente dalla sua definizione. Vogliamo mostrare inoltre che per ogni vettore $\mathbf{y} \in \mathbb{R}^{n-1}$ non nullo si ha

$$\mathbf{y}^T \mathbf{A}' \mathbf{y} > 0 \qquad \text{cioè} \qquad \mathbf{y}^T \left(\tilde{\mathbf{A}} - \alpha^{-1} \mathbf{a}^T \mathbf{a} \right) \mathbf{y} > 0,$$

nell'ipotesi che $\mathbf{A}$ sia definita positiva. Dato un vettore generico $\mathbf{y} \in \mathbb{R}^{n-1}$, introduciamo un vettore ausiliario $\mathbf{z} = [\, \eta \quad \mathbf{y}^T \,]^T \in \mathbb{R}^n$, dove η è a sua volta un qualsiasi numero reale non nullo. La positività di $\mathbf{A}$ implica che $\mathbf{z}^T \mathbf{A} \mathbf{z} > 0$, cioè

$$[\, \eta \quad \mathbf{y}^T \,] \begin{bmatrix} \alpha & \mathbf{a}^T \\ \mathbf{a} & \mathbf{A}' \end{bmatrix} \begin{bmatrix} \eta \\ \mathbf{y} \end{bmatrix} = \alpha\, \eta^2 + 2\mathbf{a}^T \mathbf{y}\, \eta + \mathbf{y}^T \tilde{\mathbf{A}} \mathbf{y} > 0.$$

La positività di questo trinomio di secondo grado in η implica che il discriminante del trinomio sia negativo, per cui sicuramente vale

$$\Delta = \left(\mathbf{a}^T \mathbf{y} \right)^2 - \alpha \mathbf{y}^T \tilde{\mathbf{A}} \mathbf{y} < 0,$$

e questa è proprio la condizione che dobbiamo verificare!

Sfruttando le due proprietà *(a)* e *(b)* possiamo concludere la dimostrazione procedendo per induzione sull'ordine della matrice. Assumiamo che $\mathbf{A} \in \mathbb{R}^{n\times n}$ sia *SPD*, e che si abbia

- per $n = 1$ una matrice *SPD* si fattorizza come già riportato prima, e cioè
$$[A_{11}] = [\sqrt{A_{11}}][\sqrt{A_{11}}].$$
- per ogni matrice $\mathbf{A}'$ *SPD* di ordine $n-1$ esiste una matrice triangolare superiore $\mathbf{R}$ che la fattorizza nella forma di Cholesky $\mathbf{A}' = \mathbf{R}'\mathbf{R}'^T$.

Allora, le proprietà *(a)* e *(b)* implicano che se $\mathbf{A}$ è *SPD* è possibile operare la riduzione, e dalle ipotesi induttive segue subito che la matrice è fattorizzabile in forma di Cholesky. ■

2.2.1 Algoritmo di calcolo

Nel presente paragrafo mostreremo come procede logicamente la fattorizzazione con il metodo di Cholesky. Esaminiamo in dettaglio il primo ed il secondo step, lo step generico k ed il passo finale.

Step 1

Il teorema appena dimostrato ci informa che possiamo scrivere anche una decomposizione del tipo

$$\mathbf{A} = \begin{bmatrix} \alpha & \mathbf{a}^T \\ \mathbf{a} & \tilde{\mathbf{A}} \end{bmatrix} = \begin{bmatrix} \sqrt{\alpha} & \mathbf{0} \\ \frac{1}{\sqrt{\alpha}}\mathbf{a} & \mathbf{I}_{n-1} \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & \mathbf{A}' \end{bmatrix} \begin{bmatrix} \sqrt{\alpha} & \frac{1}{\sqrt{\alpha}}\mathbf{a}^T \\ \mathbf{0} & \mathbf{I}_{n-1} \end{bmatrix} = \mathbf{R}_1^T \mathbf{A}_1 \mathbf{R}_1$$

dove la definizione delle matrici $\mathbf{A}_1$ e $\mathbf{R}_1$ è ovvia ed inoltre abbiamo introdotto la matrice "ridotta" di ordine $(n-1)$

$$\mathbf{A}' = \tilde{\mathbf{A}} - \frac{1}{\alpha}\mathbf{a}\mathbf{a}^T,$$

che sappiamo essere *SPD*, perché la matrice originale $\mathbf{A}$ è *SPD*. La matrice $\mathbf{A}'$ è il *complemento di Schur* della matrice $\tilde{\mathbf{A}}$ rispetto all'elemento pivotale α.

Step 2

Poiché la matrice ridotta $\mathbf{A}'$ è ancora *SPD*, possiamo ripetere su di essa lo step precedente. Iniziamo riscrivendo la matrice A' a blocchi

$$\mathbf{A}' = \begin{bmatrix} \alpha' & \mathbf{a}'^T \\ \mathbf{a}' & \tilde{\mathbf{A}}' \end{bmatrix}$$

ed operiamo la riduzione come nel primo step decomponendola come segue,

$$\begin{bmatrix} \alpha' & \mathbf{a}'^T \\ \mathbf{a} & \tilde{\mathbf{A}}' \end{bmatrix} = \begin{bmatrix} \sqrt{\alpha'} & \mathbf{0} \\ \frac{1}{\sqrt{\alpha'}}\mathbf{a}' & \mathbf{I}_{n-2} \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & \mathbf{A}'' \end{bmatrix} \begin{bmatrix} \sqrt{\alpha'} & \frac{1}{\sqrt{\alpha'}}\mathbf{a}'^T \\ \mathbf{0} & \mathbf{I}_{n-2} \end{bmatrix}.$$

La matrice "ridotta" $\mathbf{A}'' \in \mathbb{R}^{(n-2)\times(n-2)}$

$$\mathbf{A}'' = \tilde{\mathbf{A}}' - \frac{1}{\alpha'}\mathbf{a}'\mathbf{a}'^T,$$

che si calcola come *complemento di Schur* della matrice $\tilde{\mathbf{A}}'$ rispetto all'elemento pivotale α', sarà ancora *SPD*, perché lo era la matrice $\mathbf{A}'$. Di fatto, abbiamo così decomposto la matrice A_1 ottenuta al primo step

$$\begin{aligned} \mathbf{A}_1 &= \left[\begin{array}{c|cc} 1 & \mathbf{0}^T \\ \hline \mathbf{0} & \alpha' & \mathbf{a}'^T \\ & \mathbf{a}' & \tilde{\mathbf{A}}' \end{array}\right] = \left[\begin{array}{c|cc} 1 & \mathbf{0}^T \\ \hline \mathbf{0} & \sqrt{\alpha'} & \mathbf{0} \\ & \frac{1}{\sqrt{\alpha'}}\mathbf{a}' & \mathbf{I}_{n-2} \end{array}\right] \left[\begin{array}{c|cc} 1 & \mathbf{0}^T \\ \hline \mathbf{0} & 1 & \mathbf{0}^T \\ & \mathbf{0} & \mathbf{A}'' \end{array}\right] \left[\begin{array}{c|cc} 1 & \mathbf{0}^T \\ \hline \mathbf{0} & \sqrt{\alpha'} & \frac{1}{\sqrt{\alpha'}}\mathbf{a}'^T \\ & \mathbf{0} & \mathbf{I}_{n-2} \end{array}\right], \\ &= R_2^T A_2 R_2, \end{aligned}$$

dove le matrici $\mathbf{R}_2$ e $\mathbf{A}_2$ sono introdotte in maniera ovvia. Dopo due step di fattorizzazione, la matrice $\mathbf{A}$ è stata così decomposta

$$\mathbf{A} = \mathbf{R}_1^T \mathbf{A}_1 \mathbf{R}_1 = \mathbf{R}_1^T \mathbf{R}_2^T \mathbf{A}_2 \mathbf{R}_2 \mathbf{R}_1.$$

Step k generico

Dopo $k-1$ step abbiamo decomposto la matrice $\mathbf{A}$ nel prodotto dei fattori

$$\mathbf{A} = \mathbf{R}_1^T \mathbf{R}_2^T \ldots \mathbf{R}_{k-1}^T \mathbf{A}_{k-1} \mathbf{R}_{k-1} \ldots \mathbf{R}_2 \mathbf{R}_1,$$

dove

$$\mathbf{A}_{k-1} = \begin{bmatrix} \mathbf{I}_{k-1} & \mathbf{0}^T \\ \mathbf{0} & \mathbf{A}^{(k-1)} \end{bmatrix}.$$

Dal teorema generale segue che il blocco di matrice $\mathbf{A}^{(k-1)} \in \mathbb{R}^{(n-k+1)\times(n-k+1)}$

$$\mathbf{A}^{(k-1)} = \begin{bmatrix} \alpha^{(k-1)} & \mathbf{a}^{(k-1)T} \\ \mathbf{a}^{(k-1)} & \tilde{\mathbf{A}}^{(k-1)} \end{bmatrix}$$

è una matrice *SPD*, e quindi si può ancora procedere nella decomposizione.

Decomponiamo a blocchi la matrice $\mathbf{A}_{k-1}$

$$\begin{aligned}
\mathbf{A}_{k-1} &= \left[\begin{array}{c:cc} \mathbf{I}_{k-1} & \multicolumn{2}{c}{\mathbf{0}^T} \\ \hdashline \mathbf{0} & \alpha^{(k-1)} & \mathbf{a}^{(k-1)T} \\ & \mathbf{a}^{(k-1)} & \tilde{\mathbf{A}}^{(k-1)} \end{array}\right], \\
&= \left[\begin{array}{c:cc} I_{k-1} & \multicolumn{2}{c}{\mathbf{0}^T} \\ \hdashline \mathbf{0} & \sqrt{\alpha^{(k-1)}} & 0 \\ & \dfrac{1}{\sqrt{\alpha^{(k-1)}}}\mathbf{a}^{(k-1)} & \mathbf{I}_{n-k} \end{array}\right] \left[\begin{array}{c:cc} I_{k-1} & \multicolumn{2}{c}{\mathbf{0}^T} \\ \hdashline \mathbf{0} & 1 & \mathbf{0}^T \\ & \mathbf{0} & \mathbf{A}^{(k)} \end{array}\right] \left[\begin{array}{c:cc} I_{k-1} & \multicolumn{2}{c}{\mathbf{0}^T} \\ \hdashline \mathbf{0} & \sqrt{\alpha^{(k-1)}} & \dfrac{1}{\sqrt{\alpha^{(k-1)}}}\mathbf{a}^{(k-1)T} \\ & \mathbf{0} & \mathbf{I}_{n-k} \end{array}\right], \\
&= \mathbf{R}_k^T \mathbf{A}_k \mathbf{R}_k,
\end{aligned}$$

introducendo le matrici R_2, A_2, dove il blocco interno $\mathbf{A}^{(k)}$ è il complemento di Schur della matrice $\tilde{\mathbf{A}}^{(k-1)}$ rispetto all'elemento pivotale $\alpha^{(k-1)}$, ed è ancora una matrice *SPD*.

Dopo k step abbiamo decomposto la matrice $\mathbf{A}$ nella forma seguente,

$$\mathbf{A} = \mathbf{R}_1^T \mathbf{R}_2^T \dots \mathbf{R}_k^T \mathbf{A}_k \mathbf{R}_k \dots \mathbf{R}_2 \mathbf{R}_1.$$

Step n

Procedendo in questo modo è evidente che dopo n passi otterremo la seguente decomposizione della matrice $\mathbf{A}$ originale,

$$\mathbf{A} = \mathbf{R}_1^T \mathbf{R}_2^T \dots \mathbf{R}_n^T \mathbf{I}_n \mathbf{R}_n \mathbf{R}_{n-1} \dots \mathbf{R}_1.$$

Le matrici $\mathbf{R}_i^T$, $i = 1, 2 \dots n$ non solo sono triangolari inferiori ma sono anche matrici di Frobenius. Piú precisamente, la matrice $\mathbf{R}_i^T$ è matrice di Frobenius di indice i e dato che il prodotto di

queste matrici (e delle trasposte) appare nell'ordine giusto, esso è una matrice triangolare inferiore che ha nella colonna i-esima l'i-esima colonna di $\mathbf{R}_i^T$. Possiamo introdurre la $\mathbf{R}^T$ – e quindi la $\mathbf{R}$ – come

$$\mathbf{R}^T = \mathbf{R}_1^T \mathbf{R}_2^T \dots \mathbf{R}_n^T,$$

e concludere che ad ogni passo dell'algoritmo stiamo effettivamente calcolando una colonna (o una riga) del fattore di Cholesky (o del suo trasposto).

Osservazione 38. Si noti che la simmetria della matrice $\mathbf{A}$ permette di "risparmiare" nel calcolo degli elementi dei vari complementi di Schur, perché anche questi risulteranno simmetrici come conseguenza del teorema generale. Nel metodo di fattorizzazione di Cholesky si opera quindi sempre sulla diagonale e sulla metà superiore (o inferiore) della matrice, essendo il resto degli elementi noto per trasposizione.

2.2.2 Connessioni con la fattorizzazione di Gauss

Osservazione 39. La fattorizzazione di Cholesky è una variante "furba" per matrici *SPD* della fattorizzazione di Gauss. Infatti, decomponiamo la matrice $\mathbf{A}$ a blocchi come segue

$$\begin{bmatrix} \alpha\beta & \mathbf{r}^T \\ \mathbf{c} & \mathbf{A}' \end{bmatrix} = \begin{bmatrix} \alpha & \mathbf{0} \\ \frac{1}{\beta}\mathbf{c} & \mathbf{L}' \end{bmatrix} \begin{bmatrix} \beta & \frac{1}{\alpha}\mathbf{r}^T \\ \mathbf{c} & \mathbf{U}' \end{bmatrix},$$

dove ora $\alpha\beta$ è l'elemento pivotale e $\mathbf{r}$ e $\mathbf{c}$ rappresentano la prima riga e colonna di $\mathbf{A}$ a partire dal secondo elemento, e supponiamo che esistano due matrici triangolari inferiore e superiore $\mathbf{L}'$ e $\mathbf{U}'$ tali che

$$\mathbf{A}' = \mathbf{L}'\mathbf{U}' + \frac{1}{\alpha\beta}\mathbf{c}\mathbf{r}^T.$$

- se $\alpha = 1$ e $\mathrm{diag}(\mathbf{L}') = 1$, allora abbiamo la fattorizzazione di Gauss (senza pivot) $\mathbf{A} = \mathbf{L}\mathbf{U}$;
- se $\alpha = \beta$, $\mathbf{L}' = \mathbf{U}'^T$, e $\mathbf{c} = \mathbf{r}$ allora abbiamo la fattorizzazione di Cholesky $\mathbf{A} = \mathbf{R}^T\mathbf{R}$ con $\mathbf{R} = \mathbf{U}$.

Nel secondo caso la matrice deve essere obbligatoriamente simmetrica.

In entrambi i casi la matrice $\mathbf{A}''$ che esce durante il procedimento è *il complemento di Schur* del blocco di elementi corrispondenti alle righe e colonne $i, j = 2, \ldots, n$ di $\mathbf{A}$ rispetto all'elemento pivotale.

Osservazione 40. Dato che il teorema generale garantisce che l'elemento pivotale è sempre strettamente positivo, si potrebbe pensare che non è necessario operare alcun pivoting. Valgono tuttavia le stesse considerazioni fatte a proposito del pivot numerico e del pivot simbolico per l'algoritmo di Gauss. Nel caso di matrici sparse è sempre consigliabile operare un pivot simbolico per controllare la formazione dei non-zeri.

CAPITOLO

TREE

METODI ITERATIVI PER SISTEMI LINEARI

3.1 Metodi Iterativi per Sistemi Lineari

Sia $\mathbf{A} \in \mathbb{R}^{n\times n}$ non-singolare e $\mathbf{b} \in \mathbb{R}^n$.

Definizione 52. Indichiamo con il termine *residuo* la funzione vettoriale $\mathbf{r}(\cdot) : \mathbb{R}^n \to \mathbb{R}^n$ definita come

$$\mathbf{r}(\mathbf{x}) = \mathbf{b} - \mathbf{A}\mathbf{x}$$

Problema: dati $\mathbf{A} \in \mathbb{R}^{n\times n}$ non-singolare e $\mathbf{b} \in \mathbb{R}^n$, determinare $\mathbf{x} \in \mathbb{R}^n$ tale che

$$\mathbf{r}(\mathbf{x}) = \mathbf{b} - \mathbf{A}\mathbf{x} = \mathbf{0}.$$

Osservazione 41. Evidentemente, il vettore $\mathbf{x} \in \mathbb{R}^n$ che annulla il residuo, cioè tale che $\mathbf{r}(\mathbf{x}) = \mathbf{0}$, è anche soluzione del sistema lineare $\mathbf{A}\mathbf{x} = \mathbf{b}$.

Strategia Iterativa: costruiamo, partendo da una soluzione "tentativa" $\mathbf{x}^0$ una successione di soluzioni $\mathbf{x}^k$ attraverso uno *schema di iterazione* tipo

$$\mathbf{x}^{k+1} = \mathbf{G}(\mathbf{x}^k)$$

chiedendo che sotto opportune ipotesi si abbia "convergenza", cioè

$$\mathbf{x}^* = \lim_{k\mapsto\infty} \mathbf{x}^k, \qquad \mathbf{r}(\mathbf{x}^*) = \mathbf{0}$$

Questioni da definire

- come costruiamo gli schemi di iterazione $\mathbf{x}^{k+1} = \mathbf{G}(\mathbf{x}^k)$;
- come scegliamo la soluzione iniziale $\mathbf{x}^0$;
- come determiniamo l'accuratezza della soluzione approssimata;
- come arrestiamo le iterazioni, non potendo fare infiniti passi;
- come valutiamo l'efficienza di uno schema iterativo.

3.1.1 Costruzione degli schemi di iterazione mediante splitting

Scriviamo $\mathbf{A} \in \mathbb{R}^{n\times n}$ come

$$\mathbf{A} = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & A_{n-1n} \\ A_{n1} & \cdots & A_{nn-1} & A_{nn} \end{bmatrix},$$

ed introduciamo le seguenti matrici

$$\mathbf{D} = \begin{bmatrix} A_{11} & 0 & \cdots & 0 \\ 0 & A_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & A_{nn} \end{bmatrix},$$

$$-\mathbf{L} = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ A_{21} & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ A_{n1} & \cdots & A_{nn-1} & 0 \end{bmatrix}, \qquad -\mathbf{U} = \begin{bmatrix} 0 & A_{12} & \cdots & A_{1n} \\ 0 & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & A_{n-1n} \\ 0 & \cdots & 0 & 0 \end{bmatrix}.$$

Con le matrici $\mathbf{D}$, $\mathbf{U}$ ed $\mathbf{L}$ decomponiamo la matrice $\mathbf{A}$

$$\mathbf{A} = \mathbf{D} - \mathbf{L} - \mathbf{U}.$$

Utilizzando questa *decomposizione* o *splitting* di $\mathbf{A}$ il sistema $\mathbf{Ax} = \mathbf{b}$ può essere riscritto nei diversi modi

$$\begin{aligned} (i) \qquad & \mathbf{Dx} &&= (\mathbf{L} + \mathbf{U})\mathbf{x} + \mathbf{b}, \\ (ii) \qquad & (\mathbf{D} - \mathbf{L})\mathbf{x} &&= \mathbf{Ux} + \mathbf{b}, \\ (iii) \qquad & (\mathbf{D} - \mathbf{U})\mathbf{x} &&= \mathbf{Lx} + \mathbf{b}. \end{aligned}$$

Si noti che in *(i)* la matrice $\mathbf{D}$ a sinistra è diagonale, in *(ii)* la $(\mathbf{D} - \mathbf{L})$ è triangolare inferiore ed in *(iii)* la $(\mathbf{D} - \mathbf{U})$ è triangolare superiore.

Metodo di Jacobi

La formulazione del sistema $\mathbf{A}\mathbf{x} = \mathbf{b}$ in *(i)*

$$\mathbf{D}\mathbf{x} = (\mathbf{L} + \mathbf{U})\mathbf{x} + \mathbf{b}$$

suggerisce uno schema di iterazione della forma

$$\boxed{\mathbf{D}\mathbf{x}^{k+1} = (\mathbf{L} + \mathbf{U})\mathbf{x}^{k} + \mathbf{b}}$$

noto come *metodo iterativo di Jacobi*[1] .

Metodo di Gauss-Seidel

Analogamente la riformulazione in *(ii)*

$$(\mathbf{D} - \mathbf{L})\mathbf{x} = \mathbf{U}\mathbf{x} + \mathbf{b}$$

suggerisce un secondo schema di iterazione della forma

$$\boxed{(\mathbf{D} - \mathbf{L})\mathbf{x}^{k+1} = \mathbf{U}\mathbf{x}^{k} + \mathbf{b}}$$

noto come *metodo iterativo di Gauss-Seidel*[2] .

Osservazione 42. A questo punto con un poco di immaginazione si potrebbe anche pensare di utilizzare la riformulazione *(iii)*

$$(\mathbf{D} - \mathbf{U})\mathbf{x} = \mathbf{L}\mathbf{x} + \mathbf{b}$$

per definire un terzo schema iterativo come

$$\boxed{(\mathbf{D} - \mathbf{U})\mathbf{x}^{k+1} = \mathbf{L}\mathbf{x}^{k} + \mathbf{b}}\ .$$

Tuttavia si noti che quest'ultimo essere riscritto come lo schema di Gauss-Seidel, cioè lo schema ottenuto da *(ii)*, sul problema lineare in cui equazioni ed incognite sono state prese in ordine inverso. Si tratta quindi uno schema di Gauss-Seidel *all'indietro*.

[1] Carl Gustav Jacob Jacobi 1804–1851

[2] Johann Carl Friedrich Gauss 1777–1855, Philipp Ludwig von Seidel 1821–1896

Metodo SOR

Separiamo la parte diagonale di $\mathbf{A}$, cioè $\mathbf{D}$, in due termini "pesati" con un coefficiente ω reale e non negativo, da determinare,

$$\mathbf{D} = \frac{1}{\omega}\mathbf{D} + \left(1 - \frac{1}{\omega}\right)\mathbf{D}.$$

Introduciamo questi due termini al posto della matrice $\mathbf{D}$ nella decomposizione utilizzata per il metodo di Gauss-Seidel

$$\left(\frac{1}{\omega}\mathbf{D} - \mathbf{L}\right)\mathbf{x} = \left[-\left(1 - \frac{1}{\omega}\right)\mathbf{D} + \mathbf{U}\right]\mathbf{x} + \mathbf{b}.$$

Si ottiene quindi un nuovo schema iterativo

$$\boxed{\left(\frac{1}{\omega}\mathbf{D} - \mathbf{L}\right)\mathbf{x}^{k+1} = \left(\frac{\omega - 1}{\omega}\mathbf{D} + \mathbf{U}\right)\mathbf{x}^{k} + \mathbf{b}}$$

noto come **S***uccessive* **O***ver* **R***elaxation method*, da cui l'acronimo **SOR**. Si noti che il metodo **SOR** dipende da un parametro ω che deve essere individuato correttamente per garantire convergenza ed efficienza.

3.1.2 Generalizzazione

Scriviamo $\mathbf{A}$, matrice quadrata non singolare di ordine n, come differenza di due matrici $\mathbf{A} = \mathbf{P} - \mathbf{Q}$ in cui $\mathbf{P}$ è non singolare e *facile da invertire*.

$$(\mathbf{P} - \mathbf{Q})\mathbf{x} = \mathbf{b}, \qquad \Rightarrow \qquad \mathbf{P}\mathbf{x} = \mathbf{Q}\mathbf{x} + \mathbf{b},$$

Da questa decomposizione si ricava in maniera naturale lo schema di iterazione

$$\boxed{\mathbf{P}\mathbf{x}^{k+1} = \mathbf{Q}\mathbf{x}^{k} + \mathbf{b}, \qquad \Rightarrow \qquad \mathbf{x}^{k+1} = \mathbf{P}^{-1}\mathbf{Q}\mathbf{x}^{k} + \mathbf{P}^{-1}\mathbf{b},}$$

Definizione 53. La matrice $\mathbf{P}^{-1}\mathbf{Q}$, che possiamo anche scrivere come segue

$$\mathbf{P}^{-1}\mathbf{Q} = \mathbf{P}^{-1}\left(\mathbf{P} - \mathbf{A}\right) = \mathbf{I} - \mathbf{P}^{-1}\mathbf{A}$$

prende il nome di *matrice di iterazione*.

Ovviamente, è indifferente scrivere lo schema di iterazione come segue

$$\mathbf{x}^{k+1} = (\mathbf{I} - \mathbf{P}^{-1}\mathbf{A})\mathbf{x}^k + \mathbf{P}^{-1}\mathbf{b},$$

Osservazione 43. Tutti i metodi precedenti introdotti finora, *Jacobi*, *Gauss-Seidel*, *SOR*, sono casi particolari del metodo generale.

3.1.3 Quadro riassuntivo

	Jacobi	Gauss-Seidel	SOR
$\mathbf{P}$	$\mathbf{D}$	$\mathbf{D} - \mathbf{L}$	$\frac{\mathbf{D}}{\omega} - \mathbf{L}$
$\mathbf{Q}$	$\mathbf{L} + \mathbf{U}$	$\mathbf{U}$	$\frac{1-\omega}{\omega}\mathbf{D} + \mathbf{U}$
$\mathbf{I} - \mathbf{P}^{-1}\mathbf{A}$	$\mathbf{D}^{-1}(\mathbf{L} + \mathbf{U})$	$(\mathbf{D} - \mathbf{L})^{-1}\mathbf{U}$	$\left(\frac{\mathbf{D}}{\omega} - \mathbf{L}\right)^{-1}\left(\frac{1-\omega}{\omega}\mathbf{D} + \mathbf{U}\right)$

3.1.4 Convergenza degli schemi iterativi

Definizione 54. Uno schema iterativo della forma

$$\mathbf{x}^{k+1} = \mathbf{G}(\mathbf{x}^k)$$

si dice convergente se *esiste* un vettore x^* ed un norma vettoriale $\|\cdot\|$ tali che la successione delle iterate costruita a partire da un **qualsiasi** vettore iniziale $\mathbf{x}^0$ converge a x^* nella norma considerata

$$\lim_{k \mapsto \infty} \left\| \mathbf{x}^k - \mathbf{x}^* \right\| = \mathbf{0}.$$

Osservazione 44. Si noti che $\mathbf{x}^*$ deve soddisfare la condizione

$$\mathbf{x}^* = \mathbf{G}(\mathbf{x}^*)$$

Osservazione 45. Uno schema convergente nel senso della definizione precedente ha convergenza *globale*. Per problemi non-lineari uno schema iterativo può essere convergente solo per un sottoinsieme di vettori di $\mathbb{R}^n$. In tal caso si parla di convergenza *locale*. Per problemi lineari, si assume sempre che la convergenza sia globale.

Questione: *sotto quali condizioni* $\mathbf{x}^k$ *converge alla soluzione* x^**?*

Definizione 55. Definiamo errore $\boldsymbol{\varepsilon}^k$ la differenza tra la soluzione *esatta* $\mathbf{x}^*$ e la soluzione *approssimata* $\mathbf{x}^k$; cioè

$$\boldsymbol{\varepsilon}^k = \mathbf{x}^* - \mathbf{x}^k,$$

quindi la condizione che lo schema iterativo sia convergente si può esprimere anche chiedendo che l'errore sia infinitesimo in norma

$$\lim_{k \mapsto \infty} \left\| \boldsymbol{\varepsilon}^k \right\| = \mathbf{0}.$$

Come si trasforma l'errore secondo lo schema iterativo generale?

Osserviamo che dalle due espressioni che definiscono $\mathbf{x}^*$ e $\mathbf{x}^{k+1}$

$$\mathbf{P}\mathbf{x}^* = \mathbf{Q}\mathbf{x}^* + \mathbf{b},$$
$$\mathbf{P}\mathbf{x}^{k+1} = \mathbf{Q}\mathbf{x}^k + \mathbf{b},$$

si ottiene per differenza

$$\mathbf{P}\left(\mathbf{x}^* - \mathbf{x}^{k+1}\right) = \mathbf{Q}\left(\mathbf{x}^* - \mathbf{x}^k\right),$$

da cui segue iterando

$$\boldsymbol{\varepsilon}^{k+1} = \left(\mathbf{P}^{-1}\mathbf{Q}\right)\boldsymbol{\varepsilon}^k = \left(\mathbf{P}^{-1}\mathbf{Q}\right)^2 \boldsymbol{\varepsilon}^{k-1} = \ldots = \left(\mathbf{P}^{-1}\mathbf{Q}\right)^{k+1} \boldsymbol{\varepsilon}^0.$$

Osservazione 46. La convergenza si caratterizza chiedendo che l'errore sia infinitesimo in una qualche norma vettoriale.

Tuttavia, l'errore così definito implica la conoscenza della soluzione esatta $\mathbf{x}^*$. Invece si noti che

$$\mathbf{A}\boldsymbol{\varepsilon}^k = \mathbf{A}\mathbf{x}^* - \mathbf{A}\mathbf{x}^k = \mathbf{b} - \mathbf{A}\mathbf{x}^k = \mathbf{r}^k,$$

da cui segue immediatamente che

$$\mathbf{x}^k \mapsto \mathbf{x}^* \qquad \Rightarrow \qquad \mathbf{r}^k \mapsto 0.$$

Quindi studiare la convergenza di un metodo iterativo è equivalente a studiare quando il residuo tende a zero.

Come si trasforma il residuo secondo lo schema iterativo generale?

Osserviamo che si può scrivere

$$\mathbf{A} = \mathbf{P} - \mathbf{Q} = \left(\mathbf{I} - \mathbf{Q}\mathbf{P}^{-1}\right)\mathbf{P}$$

da cui segue con ovvi passaggi che

$$\begin{aligned}
\mathbf{A}\mathbf{P}^{-1}\mathbf{Q} &= \left(\mathbf{I} - \mathbf{Q}\mathbf{P}^{-1}\right)\mathbf{P}\ \ \mathbf{P}^{-1}\mathbf{Q}, \\
&= \mathbf{Q} - \mathbf{Q}\mathbf{P}^{-1}\mathbf{Q}, \\
&= \mathbf{Q}\mathbf{P}^{-1}\mathbf{P} - \mathbf{Q}\mathbf{P}^{-1}\mathbf{Q}, \\
&= \mathbf{Q}\mathbf{P}^{-1}\left(\mathbf{P} - \mathbf{Q}\right), \\
&= \mathbf{Q}\mathbf{P}^{-1}\mathbf{A}.
\end{aligned}$$

Infine, dalla relazione $\mathbf{A}\mathbf{P}^{-1}\mathbf{Q} = \mathbf{Q}\mathbf{P}^{-1}\mathbf{A}$ segue subito che

$$\begin{aligned}
\mathbf{r}^{k+1} &= \mathbf{b} - \mathbf{A}\mathbf{x}^{k+1}, \\
&= \mathbf{b} - \mathbf{A}\mathbf{P}^{-1}\mathbf{Q}\mathbf{x}^{k} - \mathbf{A}\mathbf{P}^{-1}\mathbf{b}, \\
&= \mathbf{b} - \mathbf{Q}\mathbf{P}^{-1}\mathbf{A}\mathbf{x}^{k} - \left(\mathbf{P} - \mathbf{Q}\right)\mathbf{P}^{-1}\mathbf{b}, \\
&= \mathbf{b} - \mathbf{Q}\mathbf{P}^{-1}\mathbf{A}\mathbf{x}^{k} - \mathbf{b} + \mathbf{Q}\mathbf{P}^{-1}\mathbf{b}, \\
&= \mathbf{Q}\mathbf{P}^{-1}(\mathbf{b} - \mathbf{A}\mathbf{x}^{k}), \\
&= \mathbf{Q}\mathbf{P}^{-1}\mathbf{r}^{k}.
\end{aligned}$$

Teorema 62. *Sia $\mathbf{A}$ una qualsiasi matrice quadrata. Allora le seguenti condizioni sono equivalenti:*

① $\lim_{k \mapsto \infty} \mathbf{A}^k = \mathbf{0}$.

② $\lim_{k \mapsto \infty} \mathbf{A}^k\mathbf{v} = \mathbf{0}$ *per ogni vettore* $\mathbf{v}$.

③ $\varrho(\mathbf{A}) < 1$.

④ $\|\mathbf{A}\| < 1$ *per almeno una norma matriciale.*

Teorema 63. *Dato un sistema lineare $\mathbf{A}\mathbf{x}^* = \mathbf{b}$ ed una decomposizione $\mathbf{A} = \mathbf{P} - \mathbf{Q}$ con $\mathbf{P}$ non singolare, allora il metodo iterativo*

$$\mathbf{P}\mathbf{x}^{k+1} = \mathbf{Q}\mathbf{x}^{k} + \mathbf{b},$$

converge alla soluzione $\mathbf{x}^*$ *se e solo se*

(i)

$$\boxed{\varrho(\mathbf{Q}\mathbf{P}^{-1}) < 1}$$

oppure ***equivalentemente*** *se e solo se*

(ii)

$$\boxed{\varrho(\mathbf{P}^{-1}\mathbf{Q}) < 1}$$

Dimostrazione. Dimostriamo separatamente le due tesi del teorema.

(i) Il residuo $k-$esimo può essere scritto come

$$\mathbf{r}^k = (\mathbf{Q}\mathbf{P}^{-1})^k \mathbf{r}^0,$$

per cui dal teorema 1 segue che

$$\varrho(\mathbf{Q}\mathbf{P}^{-1}) < 1 \qquad \Rightarrow \qquad \lim_{k\mapsto\infty} \mathbf{r}^k = \mathbf{0}.$$

Viceversa, se il metodo converge, si ha sempre dal teorema 62 che

$$\text{per ogni } \mathbf{r}^0, \quad \lim_{k\mapsto\infty} \mathbf{r}^k = \lim_{k\mapsto\infty} (\mathbf{Q}\mathbf{P}^{-1})^k \mathbf{r}^0 = \mathbf{0} \;\Rightarrow$$
$$\Rightarrow \qquad \varrho(\mathbf{Q}\mathbf{P}^{-1}) < 1.$$

(ii) Si dimostra analogamente sfruttando il modo in cui si trasforma l'errore durante il procedimento iterativo. Poiché

$$\boldsymbol{\varepsilon}^k = (\mathbf{P}^{-1}\mathbf{Q})^k \boldsymbol{\varepsilon}^0,$$

dal teorema 62 segue che

$$\varrho(\mathbf{P}^{-1}\mathbf{Q}) < 1 \qquad \Rightarrow \qquad \lim_{k\mapsto\infty} \boldsymbol{\varepsilon}^k = \mathbf{0}.$$

Viceversa, se il metodo converge, si ha sempre dal teorema 1 che

$$\text{per ogni } \varepsilon^0, \quad \lim_{k\mapsto\infty} \varepsilon^k = \lim_{k\mapsto\infty} (\mathbf{P}^{-1}\mathbf{Q})^k \varepsilon^0 = \mathbf{0} \;\Rightarrow$$
$$\Rightarrow \qquad \varrho(\mathbf{P}^{-1}\mathbf{Q}) < 1.$$

Corollario 64. *Il metodo di Jacobi converge alla soluzione se e solo se*

$$\varrho\left(\mathbf{D}^{-1}(\mathbf{L}+\mathbf{U})\right) < 1.$$

Corollario 65. *Il metodo di Gauss-Seidel converge se e solo se*

$$\varrho\left((\mathbf{D}-\mathbf{L})^{-1}\mathbf{U}\right) < 1.$$

Corollario 66. *Il metodo SOR converge se e solo se*

$$\varrho\left(\left(\frac{\mathbf{D}}{\omega}-\mathbf{L}\right)^{-1}\left(\frac{1-\omega}{\omega}\mathbf{D}+\mathbf{U}\right)\right) < 1.$$

Teorema 67. *Il raggio spettrale della matrice di iterazione del metodo SOR soddisfa*

$$\varrho\left(\left(\frac{1-\omega}{\omega}\mathbf{D}+\mathbf{U}\right)\left(\frac{\mathbf{D}}{\omega}-\mathbf{L}\right)^{-1}\right) \geq |1-\omega|\,.$$

Osservazione 47. Questo risultato dice che scegliere $\omega \in (0,2)$ è condizione *necessaria* per la convergenza del metodo SOR, indipendentemente dalla matrice $\mathbf{A} = \mathbf{D} - \mathbf{L} - \mathbf{U}$, dato che

$$\omega \notin (0,2) \Rightarrow |1-\omega| \geq 1$$

.

Dimostrazione. Calcoliamo il determinante della matrice di iterazione del metodo SOR.

$$\begin{aligned} det\left(\left[\frac{\mathbf{D}}{\omega}-\mathbf{L}\right]^{-1}\left[\frac{1-\omega}{\omega}\mathbf{D}+\mathbf{U}\right]\right) & \\ = \quad det\left(\frac{\mathbf{D}}{\omega}-\mathbf{L}\right)^{-1} & \; det\left(\frac{1-\omega}{\omega}\mathbf{D}+\mathbf{U}\right) \\ = \quad \frac{\prod_{i=1}^{n}\frac{1-\omega}{\omega}D_{ii}}{\prod_{i=1}^{n}\frac{1}{\omega}D_{ii}} & \\ = \quad (1-\omega)^n. & \end{aligned}$$

Osserviamo ora che se M è una qualsiasi matrice con autovalori λ_i e raggio spettrale $\varrho(\mathrm{M})$, allora

$$\varrho(\mathrm{M})^n \geq \prod_{i=1}^{n} |\lambda_i| \geq |det\,(\mathrm{M})|\,.$$

Prendendo M uguale alla matrice di iterazione del metodo SOR si ha

$$\begin{aligned}
&\varrho\left(\left(\frac{\mathbf{D}}{\omega}-\mathbf{L}\right)^{-1}\left(\frac{1-\omega}{\omega}\mathbf{D}+\mathbf{U}\right)\right)\\
&\qquad\geq \quad \left|det\left(\left(\frac{\mathbf{D}}{\omega}-\mathbf{L}\right)^{-1}\left(\frac{1-\omega}{\omega}\mathbf{D}+\mathbf{U}\right)\right)\right|^{\frac{1}{n}}\\
&\qquad= \quad |1-\omega|\,.
\end{aligned}$$

3.1.5 Controllo della Convergenza

Se $\boldsymbol{\varepsilon}^{k-1} \neq \mathbf{0}$, la quantità $\left\|\boldsymbol{\varepsilon}^k\right\| / \left\|\boldsymbol{\varepsilon}^{k-1}\right\|$ esprime la riduzione dell'errore al k-esimo passo. La media geometrica delle riduzioni dell'errore sui primi k passi

$$\sigma_k = \sqrt[k]{\frac{\left\|\boldsymbol{\varepsilon}^1\right\|}{\left\|\boldsymbol{\varepsilon}^0\right\|}\frac{\left\|\boldsymbol{\varepsilon}^2\right\|}{\left\|\boldsymbol{\varepsilon}^1\right\|}\cdots\frac{\left\|\boldsymbol{\varepsilon}^k\right\|}{\left\|\boldsymbol{\varepsilon}^{k-1}\right\|}} = \sqrt[k]{\frac{\left\|\boldsymbol{\varepsilon}^k\right\|}{\left\|\boldsymbol{\varepsilon}^0\right\|}}$$

esprime la *riduzione media per passo* dell'errore.

Dalle proprietà delle norme naturali segue che

$$\boldsymbol{\varepsilon}^k = (\mathbf{P}^{-1}\mathbf{Q})^k\boldsymbol{\varepsilon}^0 \qquad \Rightarrow \qquad \left\|\boldsymbol{\varepsilon}^k\right\| \leq \left\|(\mathbf{P}^{-1}\mathbf{Q})^k\right\|\left\|\boldsymbol{\varepsilon}^0\right\|$$

e quindi si ottiene la relazione

$$\sigma_k \leq \left(\left\|(\mathbf{P}^{-1}\mathbf{Q})^k\right\|\right)^{\frac{1}{k}}.$$

Quindi σ_k stima la velocità di convergenza sulle prime k iterazioni ed è controllato da una quantità che dipende dalla matrice di iterazione dello schema. Come possiamo stimare questa quantità?

Stima della velocità di convergenza

Teorema 68. *Sia* $\mathrm{M} \in \mathbb{K}^{m\times n}$ *e* $\|\cdot\|$ *una* qualunque *norma indotta. Allora*

$$\lim_{k\mapsto\infty} \left\|(\mathrm{M})^k\right\|^{\frac{1}{k}} = \varrho(\mathrm{M}).$$

Osservazione 48. La quantità $\varrho(\mathrm{M})$ non dipende dalla norma matriciale utilizzata e nemmeno dall'indice di iterazione k. Possiamo assumerla come una *stima della velocità di convergenza* sui nostri metodi iterativi ponendo $\mathrm{M} = \mathbf{P}^{-1}\mathbf{Q}$, cioè uguale alla matrice di iterazione.

Definizione 56 (Tasso asintotico di convergenza). Si definisce *tasso asintotico di convergenza* di uno schema iterativo la costante

$$\mathcal{C} = -\log_{10} \varrho(\mathbf{P}^{-1}\mathbf{Q}).$$

La costante $\mathcal{C}$ è all'incirca il numero di iterazioni richieste per ridurre l'errore di un fattore $\frac{1}{10}$, cioè grossomodo per avere ottenere una cifra decimale in più. Infatti

$$\left(\varrho(\mathbf{P}^{-1}\mathbf{Q})\right)^k \approx \frac{1}{10} \qquad \Rightarrow \qquad k \approx -\frac{1}{\log_{10} \varrho(\mathbf{P}^{-1}\mathbf{Q})}.$$

Criteri di Arresto

Sia fissata una tolleranza η sull'errore. I criteri più comunemente usati sfruttano un controllo sulla variazione assoluta e relativa delle iterate al generico passo k

$$\textbf{IF} \quad \left\|\mathbf{x}^k - \mathbf{x}^{k-1}\right\| \leq \eta \quad \textbf{THEN STOP}$$
$$\textbf{IF} \quad \left\|\mathbf{x}^k - \mathbf{x}^{k-1}\right\| \leq \eta \left\|\mathbf{x}^k\right\| \quad \textbf{THEN STOP}.$$

Tuttavia è importante rendersi conto che queste condizioni non garantiscono che la soluzione sia stata approssimata con accuratezza η

Infatti, si ha che

$$\begin{aligned}\mathbf{x}^k - \mathbf{x}^{k-1} &= \mathbf{x}^k - \mathbf{x}^* + \mathbf{x}^* - \mathbf{x}^{k-1} \\ &= -\boldsymbol{\varepsilon}^k + \boldsymbol{\varepsilon}^{k-1} \\ &= -\mathbf{P}^{-1}\mathbf{Q}\boldsymbol{\varepsilon}^{k-1} + \boldsymbol{\varepsilon}^{k-1} \\ &= \left(\mathbf{I} - \mathbf{P}^{-1}\mathbf{Q}\right)\boldsymbol{\varepsilon}^{k-1}\end{aligned}$$

per cui possiamo esprimere l'errore al passo $k-1$ in funzione della variazione assoluta delle iterate al passo k

$$\boldsymbol{\varepsilon}^{k-1} = \left(\mathbf{I} - \mathbf{P}^{-1}\mathbf{Q}\right)^{-1}\left(\mathbf{x}^k - \mathbf{x}^{k-1}\right).$$

Dalla relazione

$$\boldsymbol{\varepsilon}^{k-1} = \left(\mathbf{I} - \mathbf{P}^{-1}\mathbf{Q}\right)^{-1}\left(\mathbf{x}^k - \mathbf{x}^{k-1}\right), \qquad \text{(ripetiamo)}$$

passando alle norme se $\left\|\mathbf{P}^{-1}\mathbf{Q}\right\| \leq 1$ si ha

$$\begin{aligned}\left\|\boldsymbol{\varepsilon}^{k-1}\right\| &\leq \left\|(\mathbf{I} - \mathbf{P}^{-1}\mathbf{Q})^{-1}\right\| \left\|\mathbf{x}^k - \mathbf{x}^{k-1}\right\| \\ &\leq \frac{\left\|\mathbf{x}^k - \mathbf{x}^{k-1}\right\|}{1 - \|\mathbf{P}^{-1}\mathbf{Q}\|}\end{aligned}$$

per cui se $\left\|\mathbf{P}^{-1}\mathbf{Q}\right\| \approx 1$ l'errore in norma $\left\|\boldsymbol{\varepsilon}^{k-1}\right\|$ può essere ancora grande.

CAPITOLO

QUATTRO

ZERI DI FUNZIONI

4.1 Introduzione

Un problema che si incontra spesso nelle applicazioni è il seguente

$$\text{trovare } x \text{ tale che } f(x) = 0, \tag{4.1}$$

dove $f : \mathbb{R} \to \mathbb{R}$ è una funzione continua a valori reali di variabile reale. Un numero reale α che soddisfi $f(\alpha) = 0$ è detto *radice*. Se la funzione $f(x)$ è lineare, cioè si scrive come

$$f(x) = ax + b,$$

dove $a \neq 0$ e b sono due costanti assegnate, esiste un' unica radice che vale $\alpha = -b/a$. Se la funzione $f(x)$ è della forma seguente

$$f(x) = ax^n - b,$$

nel caso in cui a b abbiano lo stesso segno o n sia dispari, una radice reale si trova immediatamente e vale $\alpha = \sqrt[n]{b/a}$. Nei problemi che si incontrano in pratica la funzione $f(x)$ è spesso non lineare e non è possibile trovare una formula che permetta di calcolare immediatamente almeno una radice. Di conseguenza è necessario sviluppare schemi numerici per l' approssimazione di queste radici. Gli schemi che approssimano le soluzioni di un problema del tipo (4.1) rientrano in gran parte nelle seguenti due categorie:

1. gli schemi iterativi a due punti;

2. gli schemi iterativi ad un punto.

Gli schemi a due punti calcolano ad ogni passo gli estremi di un intervallo che contiene una radice, come per esempio nel caso dei metodi di bisezione (o dicotomico) e delle corde (o false posizioni). La lunghezza di questi intervalli in genere decresce e permette di stimare la radice con l'accuratezza desiderata.

Gli schemi ad un punto, invece, generano una successione di numeri che generalmente convergono alla radice. Esempi di questi schemi sono il metodo delle secanti, di Newton-Raphson [1] ed il metodo delle iterazioni a punto fisso.

In alcuni casi particolari, dipendenti dalla forma della funzione $f(x)$, ad esempio se $f(x)$ è un polinomio, esistono tecniche speciali che non rientrano nelle precedenti due categorie.

4.2 Metodo di bisezione (o dicotomico)

Il metodo dicotomico è una diretta applicazione del seguente teorema che ricordiamo senza dimostrazione.

Teorema 69. *Sia $f : [a, b] \to \mathbb{R}$ una funzione continua in $[a, b]$ e tale che $f(a)f(b) < 0$; allora esiste almeno un punto $\alpha \in (a, b)$ tale che $f(\alpha) = 0$.*

Se $f(x)$ è continua ed a_0 e b_0 sono due punti tali che $f(a_0)f(b_0) < 0$, allora $f(a_0)$ ed $f(b_0)$ hanno segno discorde. Il teorema 69 implica l'esistenza di almeno una radice α nell'intervallo (a_0, b_0). Consideriamo ora il punto medio del suddetto intervallo,

$$c_0 = \frac{a_0 + b_0}{2}.$$

Si può realizzare solo una delle seguenti tre situazioni:

1. $f(c_0) = 0$; in tal caso c è una radice.
2. $f(c_0)f(a_0) < 0$; in tal caso esiste almeno una radice nell' intervallo (a_0, c_0).
3. $f(c_0)f(b_0) < 0$; in tal caso esiste almeno una radice nell' intervallo (c_0, b_0).

Escludendo il primo caso, si può sempre dimezzare l'intervallo di ricerca nel quale esiste una radice. Lo schema dicotomico permette, data una tolleranza ϵ, di trovare un intervallo di lunghezza $< \epsilon$ che contiene una radice. L'algoritmo si può scrivere come segue:

[1] Isaac Newton 1643–1727, Joseph Raphson 1648–1715

Algorithm *Metodo dicotomico*
Input: a, b
1. $a_0 \leftarrow a$; $b_0 \leftarrow b$; $i \leftarrow 0$
2. **while** $b_i - a_i < \epsilon$
3. **do** $c \leftarrow \dfrac{a_i + b_i}{2}$
4. **if** $f(c) = 0$ **then return** c
5. **if** $f(c)f(a_i) < 0$
6. **then** ($*$ esiste almeno una radice nell'intervallo (a_i, c). $*$)
7. $a_{i+1} \leftarrow a_i$; $b_{i+1} \leftarrow c$
8. **else** ($*$ esiste almeno una radice nell'intervallo (c, b_i). $*$)
9. $a_{i+1} \leftarrow c$; $b_{i+1} \leftarrow b_i$
10. $i \leftarrow i + 1$
11. ($*$ c contiene l'approssimazione della radice cercata. $*$)
12. **return** c

Si può calcolare il numero minimo di passi necessari per approssimare una radice con l'accuratezza richiesta. Infatti,

$$\begin{aligned} b_i - a_i &= 2^{-1}(b_{i-1} - a_{i-1}), \\ &\vdots \\ &= 2^{-i}(b_0 - a_0), \\ &= 2^{-i}(b - a), \end{aligned}$$

quindi posto $b_i - a_i < \epsilon$ si ottenie

$$2^{-i}(b-a) < \epsilon, \qquad \Rightarrow \qquad \frac{b-a}{\epsilon} < 2^i,$$

e di conseguenza

$$i > \log_2\left(\frac{b-a}{\epsilon}\right) = \log_2(b-a) - \log_2(\epsilon),$$

per cui la parte intera del membro di destra fornisce il numero minimo di iterazioni necessarie per stimare una radice a meno di un errore minore o uguale ad ϵ.

4.3 Metodo delle false posizioni (regula falsi)

Sia data $f : [a, b] \mapsto \mathbb{R}$ supposta funzione continua sull'intervallo di definizione, e con $f(a)f(b) < 0$, per cui l' intervallo $[a, b]$ contiene almeno uno zero di $f(x)$.

Per stimare la posizione della radice costruiamo la retta interpolante i punti

$$\begin{bmatrix} a \\ f(a) \end{bmatrix}, \qquad \begin{bmatrix} b \\ f(b) \end{bmatrix},$$

cioè

$$y = f(a) + (x - a)\frac{f(b) - f(a)}{b - a}, \tag{4.2}$$

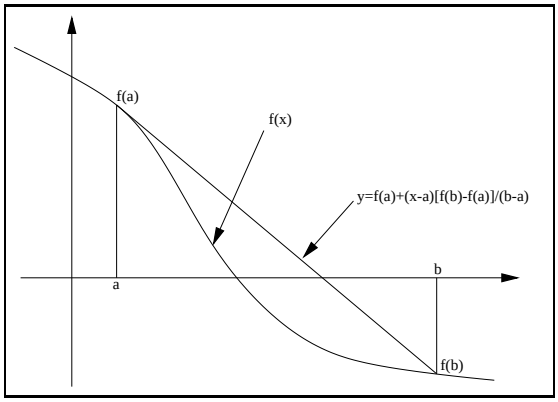


Figura 4.1: Costruzione della "regula falsi"

Lo zero della retta in (4.2) fornisce la seguente stima dello zero della funzione $f(x)$

$$0 = f(a) + (\hat{x} - a)\frac{f(b) - f(a)}{b - a}, \qquad \Rightarrow \qquad \hat{x} = \frac{af(b) - bf(a)}{f(b) - f(a)}.$$

A questo punto potremmo utilizzare un procedimento simile a quello per il metodo dicotomico e ripetere il procedimento per l'intervallo $[a, \hat{x}]$ o $[\hat{x}, b]$. Se usassimo lo stesso algoritmo dello schema dicotomico avremo che in generale la regula falsi non costruisce intervalli convergenti a zero come si vede dalla figura 4.2

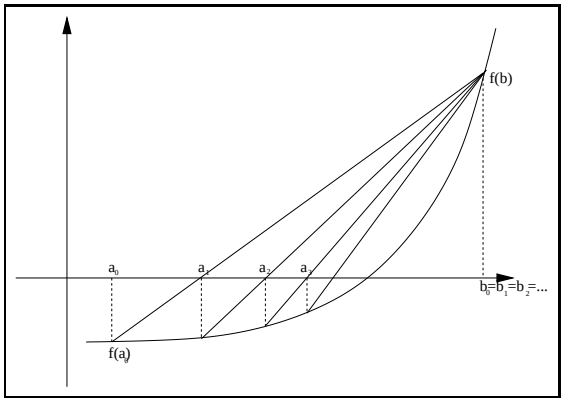


Figura 4.2: Andamento della “regula falsi”

perciò descriveremo l’andamento della regula falsi come la successione di due punti x_i e a_i dove x_i è una approssimazione della radice e a_i è tale che $f(x_i)f(a_i) < 0$.

Il procedimento è riassunto nel seguente algoritmo:

Algorithm *Regula falsi*
Input: a, b
1. $a_0 \leftarrow a$; $x_0 \leftarrow b$; $i \leftarrow 0$
2. **while** $|f(x_i)| \geq \epsilon$
3. **do** $\hat{x} \leftarrow \dfrac{a_i f(x_i) - x_i f(a_i)}{f(x_i) - f(a_i)}$
4. **if** $f(x_i)f(\hat{x}) < 0$
5. **then** $x_{i+1} \leftarrow \hat{x}$; $a_{i+1} \leftarrow a_i$
6. **else** $x_{i+1} \leftarrow x_i$; $a_{i+1} \leftarrow \hat{x}$
7. $i \leftarrow i + 1$
8. **return** x_i

4.3.1 Convergenza dalla “regula falsi”

Per discutere la convergenza della regula falsi ammettiamo che $f(x)$ abbia derivata seconda regolare in ogni punto dell’intervallo di definizione della funzione. Assumiamo, inoltre, che

$a)$ $x_0 < a_0$;

$b)$ $f(x_0) < 0$; $\qquad f(a_0) > 0$;

$c)$ $f''(x) \geq 0$ per ogni $x \in [x_0, a_0]$;

Con queste ipotesi, vale il seguente teorema.

Teorema 70. *La "regula falsi" con le ipotesi a–c converge e vale inoltre*

$$x_k \leq x_{k+1} < a_k = a_0, \qquad \text{per } k = 1, 2, \ldots.$$

Dimostrazione. Innanzitutto osserviamo che l'ipotesi $f''(x) \geq 0$ in $[x_0, a_0]$ implica che la funzione $f(x)$ sia convessa in $[x_0, a_0]$, cioè

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y), \qquad \forall x, y \in [x_0, a_0].$$

Inoltre poiché $f(x_0) < 0$ e $f(a_0) > 0$, almeno una radice si trova nell'intervallo $[x_0, a_0]$. Sia

$$\eta = \sup\{x \in [x_0, a_0] \,|\, f(x) < 0\}.$$

Ovviamente dalla continuità di $f(x)$ segue che $f(\eta) = 0$. Calcoliamo ora x_1

$$x_1 = \frac{a_0 f(x_0) - x_0 f(a_0)}{f(x_0) - f(a_0)} = x_0 + (a_0 - x_0)\frac{f(x_0)}{f(x_0) - f(a_0)}.$$

Poiché $f(x_0) < 0$ e $f(a_0) > 0$ e $x_0 < a_0$ segue che $x_1 > x_0$. Inoltre dalla convessità segue che

$$\begin{aligned} f(x_1) &= f\left(x_0 \frac{f(a_0)}{f(a_0) - f(x_0)} + a_0\left(1 - \frac{f(a_0)}{f(a_0) - f(x_0)}\right)\right), \\ &\leq \frac{f(a_0)}{f(a_0) - f(x_0)} f(x_0) + \left(1 - \frac{f(a_0)}{f(a_0) - f(x_0)}\right) f(a_0) = 0, \end{aligned}$$

e di conseguenza $x_1 \leq \eta$. In modo analogo si vede che

$$x_0 \leq x_1 \leq x_2 \ldots \leq x_k \leq \eta < a_0 = a_1 = \cdots = a_k.$$

Quindi x_k per $k = 0, 1, \ldots$ è una successione monotona crescente e quindi convergente. Indichiamo il suo limite col simbolo μ; allora avremo che

$$\mu = \lim_{k \mapsto \infty} x_k \leq \eta.$$

Inoltre varrà

$$\mu = \frac{a_0 f(\mu) - \mu f(a_0)}{f(\mu) - f(a_0)},$$

e di conseguenza

$$(\mu - a_0) f(\mu) = 0;$$

poiché $\mu - a_0 \leq \eta - a_0 < 0$ segue che $f(\mu) = 0$ cioè il metodo è convergente. ■

4.4 Metodo di Newton-Raphson

Sia data una funzione $f : [a, b] \mapsto \mathbb{R}$ continua e differenziabile. Se x_0 non è uno zero della funzione possiamo stimare la posizione dello zero approssimando la funzione $f(x)$ con la retta passante per $[\, x_0, f(x_0)\,]^T$ e tangente alla funzione,

$$y = f(x_0) + (x - x_0) f'(x_0). \tag{4.3}$$

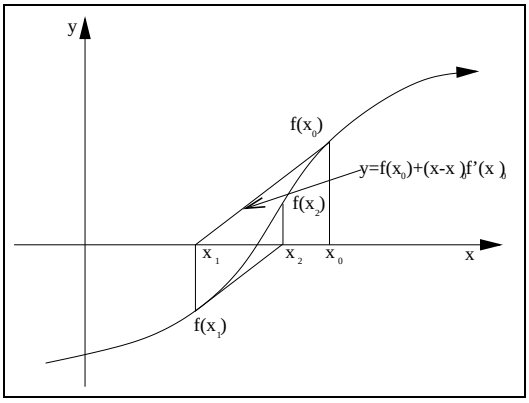


Figura 4.3: Costruzione del metodo di Newton-Raphson

Lo zero x_1 della retta (4.3) fornisce una approssimazione della radice di $f(x)$,

$$0 = f(x_0) + (x_1 - x_0) f'(x_0), \qquad \Rightarrow \qquad x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}.$$

Ripetendo il procedimento a partire da x_1 fornisce un punto x_2, che possiamo sperare essere una approssimazione ancora migliore della radice cercata. Questo suggerisce un procedimento iterativo della forma:

$$\boxed{x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \qquad k = 1, 2, \ldots} \tag{4.4}$$

Lo schema prende il nome di metodo di Newton-Raphson. Vale il seguente teorema:

Teorema 71. *Sia $f \in C^2([a,b])$ ed α una radice semplice di $f(x)$ cioè $f'(\alpha) \neq 0$. Allora se $x_0 \in (\alpha-\epsilon, \alpha+\epsilon)$ con ϵ sufficientemente piccolo avremo che la successione generate dallo schema* (4.4) *soddisfi:*

1. $x_k \in (\alpha-\epsilon, \alpha+\epsilon)$ *per* $k = 1, 2, 3, \ldots$
2. $|x_{k+1} - \alpha| \leq C\,|x_k - \alpha|^2$ *per* $k = 1, 2, 3, \ldots$

Dimostrazione. Sviluppando con la serie di Taylor e utilizzando (4.4):

$$\begin{aligned} 0 = f(\alpha) &= f(x_k) + f'(x_k)(\alpha - x_k) + f''(\eta_k)\frac{(x_k-\alpha)^2}{2}, \\ &= f'(x_k)\left(\frac{f(x_k)}{f'(x_k)} + \alpha - x_k\right) + \frac{f''(\eta_k)}{2}(\alpha - x_k)^2, \\ &= f'(x_k)\,(\alpha - x_{k+1}) + \frac{f''(\eta_k)}{2}(\alpha - x_k)^2, \end{aligned}$$

da cui

$$\alpha - x_{k+1} = \frac{f''(\eta_k)}{2f(x_k)}(\alpha - x_k)^2.$$

Prendendo in valore assoluto entrambi i membri e ponendo

$$M = \sup_{\eta \in [a,b]} |f''(\eta)|,$$

otteniamo

$$|x_{k+1} - \alpha| \leq |x_k - \alpha|^2 \frac{M}{2\,|f'(x_k)|}.$$

Poiché α è supposto zero semplice si ha che $|f'(\alpha)| > 0$. Inoltre, per la continuità di f', esisterà ϵ tale che $2\,|f'(x)| > |f'(\alpha)|$ per ogni $x \in (\alpha-\epsilon, \alpha+\epsilon)$. Se $x_k \in (\alpha-\epsilon, \alpha+\epsilon)$ allora avremo

$$|x_{k+1} - \alpha| \leq |x_k - \alpha|^2 \frac{M}{2\,|f'(x_k)|} \leq \epsilon^2 \frac{M}{|f'(\alpha)|}. \tag{4.5}$$

Eventualmente riducendo ϵ possiamo supporre $\epsilon \dfrac{M}{|f'(\alpha)|} < 1$, cosicché dalla (4.5) segue il punto 1,

$$|x_{k+1} - \alpha| \leq \epsilon, \qquad \Rightarrow \qquad x_{k+1} \in (\alpha-\epsilon, \alpha+\epsilon).$$

Inoltre ponendo $C = M/\left|f'(\alpha)\right|$ si trova verificato il punto 2. ■

4.5 Metodo delle secanti

Se invece di controllare ogni volta il segno del prodotto $f(x_i)f(a_i)$ utilizziamo i valori delle ultime due iterate per calcolare una approssimazione della radice, otteniamo il metodo delle secanti che prende la forma seguente:

$$\boxed{x_{i+1} = \frac{x_{i-1}f(x_i) - x_i f(x_{i-1})}{f(x_i) - f(x_{i-1})}, \qquad i = 1, 2, \ldots.} \tag{4.6}$$

Nel metodo delle secanti quando si arriva a convergenza si ha che $f(x_i) \approx f(x_{i-1})$, per cui ci possono essere problemi di cancellazione. Osseviamo che ciò non si verifica nella "regula falsi", dato che per quel procedimento vale $f(x_i)f(a_i) < 0$. Inoltre, se x_i non è sufficientemente vicino alla radice, non è detto che lo schema converga. Studiamo quindi la convergenza locale dello schema. Sia α una radice di $f(x)$, sottraendo α a entrambi i membri della (4.6) otteniamo

$$\begin{aligned}
x_{i+1} - \alpha &= \frac{x_{i-1}f(x_i) - x_i f(x_{i-1})}{f(x_i) - f(x_{i-1})} - \alpha, \\
&= \frac{x_{i-1}f(x_i) - x_i f(x_{i-1})}{f(x_i) - f(x_{i-1})} - \alpha\frac{f(x_i) - f(x_{i-1})}{f(x_i) - f(x_{i-1})}, \\
&= \frac{(x_{i-1} - \alpha)f(x_i) - (x_i - \alpha)f(x_{i-1})}{f(x_i) - f(x_{i-1})}, \\
&= (x_i - \alpha)(x_{i-1} - \alpha)\frac{\dfrac{f(x_i)}{x_i - \alpha} - \dfrac{f(x_{i-1})}{x_{i-1} - \alpha}}{f(x_i) - f(x_{i-1})}.
\end{aligned}$$

Inoltre poiché $f(\alpha) = 0$

$$\begin{aligned}
\frac{\dfrac{f(x_i)}{x_i - \alpha} - \dfrac{f(x_{i-1})}{x_{i-1} - \alpha}}{f(x_i) - f(x_{i-1})} &= \frac{\dfrac{f(x_i) - f(\alpha)}{x_i - \alpha} - \dfrac{f(x_{i-1}) - f(\alpha)}{x_{i-1} - \alpha}}{f(x_i) - f(x_{i-1})}, \\
&= \frac{\dfrac{f(x_i) - f(\alpha)}{x_i - \alpha} - \dfrac{f(x_{i-1}) - f(\alpha)}{x_{i-1} - \alpha}}{x_i - x_{i-1}}\left(\frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}}\right)^{-1},
\end{aligned} \tag{4.7}$$

osserviamo che vale il teorema di Lagrange [2], per cui si ha

$$\frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}} = f'(\eta_i), \qquad \eta \in I[x_{i-1}, x_i], \tag{4.8}$$

dove con $I[a, b, c, \ldots]$ si intende il più piccolo intervallo chiuso contenente $a, b, c, \ldots$

Lemma 72. *Se $f \in C^2$ vale la seguente uguaglianza*

$$\frac{\dfrac{f(\alpha+h) - f(\alpha)}{h} - \dfrac{f(\alpha-k) - f(\alpha)}{k}}{h+k} = \frac{1}{2} f''(\zeta), \qquad \zeta \in [\alpha - k, \alpha + h]. \tag{4.9}$$

Dimostrazione. Data la funzione

$$G(t) := \frac{\dfrac{f(\alpha+th) - f(\alpha)}{h} - \dfrac{f(\alpha-tk) - f(\alpha)}{k}}{h+k},$$

è immediato verificare che $G(1)$ è proprio (4.9). La funzione

$$H(t) := G(t) - G(1)t^2,$$

si annulla per $t = 0, 1$, quindi per il teorema di Rolle (Michel Rolle 1652–1719) esiste un punto $\eta \in (0, 1)$ per cui $H'(\eta) = 0$. Osserviamo che

$$H'(t) = G'(t) - 2G(1)t, \qquad G'(t) = \frac{f'(\alpha+th) - f'(\alpha-tk)}{h+k}, \tag{4.10}$$

e calcolando la (4.10) in η otteniamo

$$0 = G'(\eta) - 2G(1)\eta, \qquad G(1) = \frac{1}{2\eta} G'(\eta) = \frac{f'(\alpha+\eta h) - f'(\alpha-\eta k)}{2\eta(h+k)} = \frac{1}{2} f''(\zeta).$$

Ponendo $h = x_i - \alpha$ e $k = \alpha - x_{i-1}$ nel lemma 72 otteniamo

$$\frac{\dfrac{f(x_i) - f(\alpha)}{x_i - \alpha} - \dfrac{f(x_{i-1}) - f(\alpha)}{x_{i-1} - \alpha}}{x_i - x_{i-1}} = \frac{1}{2} f''(\zeta), \tag{4.11}$$

sostituendo (4.7) ed (4.8) in (4.11) otteniamo

$$x_{i+1} - \alpha = (x_i - \alpha)(x_{i-1} - \alpha) \frac{f''(\zeta_i)}{2f'(\eta_i)}, \qquad \zeta_i, \eta_i \in I[x_{i-1}, x_i, \alpha] \tag{4.12}$$

questa equazione è utile per dimostrare il seguente teorema:

[2] Joseph-Louis Lagrange 1736—1813

Teorema 73. *Se α è uno zero semplice di $f \in C^2([a,b])$ (cioè $f'(\alpha) \neq 0$) allora esiste un intervallo $J = (x-\epsilon, x+\epsilon)$ per il quale presi comunque $x_0, x_1 \in J$ il metodo delle secanti è convergente e inoltre*

$$|x_i - \alpha| \leq Cq^{p^i}, \qquad 0 < q < 1, \quad p = \frac{1+\sqrt{5}}{2} = 1.618034\ldots$$

Dimostrazione. Poiché α è uno zero semplice e $f'(x)$ è continua esisterà un $\epsilon > 0$ per cui vale

$$|f'(x)| > \frac{1}{2}|f'(\alpha)|, \qquad \forall x \in (\alpha-\epsilon, \alpha+\epsilon),$$

inoltre poiché $f''(x)$ è continua esisterà una costante K per cui vale

$$|f''(x)| \leq K, \qquad \forall x \in (\alpha-\epsilon, \alpha+\epsilon),$$

e quindi posto $J = (x-\epsilon, x+\epsilon)$ e $M = K/|f'(\alpha)|$ avremo

$$\left|\frac{1}{2}\frac{f''(\zeta)}{f'(\eta)}\right| \leq \left|\frac{f''(\zeta)}{f'(\alpha)}\right| \leq \left|\frac{K}{f'(\alpha)}\right| = M, \qquad \forall \eta \in J,\ \forall \zeta \in J$$

possiamo supporre che $\epsilon M < 1$ (altrimenti basta ridurre ϵ), allora avremo che $x_i \in J$ per $i = 1, 2, \ldots$ infatti dalla (4.12)

$$|x_{i+1} - \alpha| \leq |x_i - \alpha|\,|x_{i-1} - \alpha|\, M \leq \epsilon^2 M < \epsilon, \qquad \Rightarrow \qquad x_{i+1} \in J \tag{4.13}$$

ponendo $e_i = M|x_i - \alpha|$ dalla (4.13) otteniamo

$$\frac{e_{i+1}}{M} \leq \frac{e_i}{M}\frac{e_{i-1}}{M}M, \qquad \Rightarrow \qquad e_{i+1} \leq e_i e_{i-1},$$

osserviamo che posto $E_0 \geq e_0$ ed $E_1 \geq e_1$ ed $E_{i+1} = E_i E_{i-1}$ abbiamo $e_i \leq E_i$ per $i = 0, 1, 2, \ldots$ Infatti se vale fino a k, allora

$$e_{k+1} \leq e_k e_{k-1} \leq E_k E_{k-1} = E_{k+1}.$$

Ponendo $E_i = \exp(cz^i)$ otteniamo

$$\exp(cz^{i+1}) = \exp(cz^i)\exp(cz^{i-1}) = \exp(c(z^i + z^{i-1})),$$

e prendendo il logaritmo da entrambe le parti e dividendo per cz^{i-1} otteniamo

$$z^2 = z + 1, \qquad \Rightarrow \qquad z_{1,2} = \frac{1 \pm \sqrt{5}}{2} = \begin{cases} 1.618034\ldots \\ -0.618034\ldots \end{cases}$$

scegliendo $p = \dfrac{1+\sqrt{5}}{2}$ una soluzione particolare della relazione di ricorrenza $E_{k+1} = E_k E_{k-1}$ è la seguente:

$$E_0 = \exp(c), \qquad E_1 = \exp(cp), \qquad E_k = \exp(cp^k),$$

e scegliendo la costante c tale che

$$\begin{aligned} e_0 \leq E_0 = \exp(c), \qquad & c \geq \log(e_0), \\ e_1 \leq E_1 = \exp(cp), \qquad & c \geq \frac{\log(e_1)}{p}, \end{aligned}$$

e ponendo $q = \exp(c)$ avremo che

$$e_i \leq \exp(cp^i), \qquad \Rightarrow \qquad M\,|x_i - \alpha| \leq \exp(c)^{p^i} = q^{p^i}.$$

Osserviamo infine che $e_i = |x_i - \alpha|\, M < \epsilon M < 1$ di conseguenza possiamo scegliere $c < 0$ e $q = \exp(c) < 1$. ■

4.6 Iterazioni di punto fisso

Vogliamo ora dare una formulazione più generale alle iterazioni del tipo Newton-Raphson.

Definizione 57. Data una funzione $g : \mathbb{R} \mapsto \mathbb{R}$ diremo che α è un punto fisso se vale:

$$\alpha = g(\alpha).$$

Cercare gli zeri di una funzione $f(x)$ è equivalente a cercare i punti fissi della funzione

$$g(x) = x - f(x).$$

Un modo molto naturale per cercare i punti fissi è quello di usare delle iterazioni. Ad esempio dato un punto iniziale x_0 costruendo la successione

$$\boxed{x_{k+1} = g(x_k), \qquad k = 1, 2, \ldots} \tag{4.14}$$

possiamo chiederci quando la successione $\{x_i\}_{i=0}^{\infty}$ è convergente e se converge a α. Ad esempio lo schema di Newton (4.4) può essere messo nella forma (4.14) ponendo

$$g(x) = x - \frac{f(x)}{f'(x)},$$

fondamentale per lo studio di schemi iterativi della forma (4.14) è il seguente teorema:

Teorema 74. *Sia $g(x)$ funzione continua e α un suo punto fisso, cioè $\alpha = g(\alpha)$. Sia $g(x)$ soddisfacente la seguente condizione*

$$|g(x) - g(x')| \leq L\,|x - x'|\,, \tag{4.15}$$

per ogni x, x' in un intervallo chiuso $I \equiv [\alpha - \rho, \alpha + \rho]$, dove la costante L soddisfa:

$$0 \leq L < 1,$$

allora

(i) *Posto $x_0 \in I$ allora tutte le iterate x_k definite da* (4.14) *stanno in I, cioè*

$$\alpha - \rho \leq x_k \leq \alpha + \rho, \qquad k = 1, 2, \ldots$$

(ii) *(esistenza) le iterate convergono ad α,*

$$\lim_{k \mapsto \infty} x_k = \alpha.$$

(iii) *(unicità) α è l'unico punto fisso di $g(x)$ in $[x_0 - \rho, x_0 + \rho]$.*

Dimostrazione. (i) basta osservare che

$$\begin{aligned} |\alpha - x_k| &= |g(\alpha) - g(x_{k-1})|\,, \\ &\leq L\,|\alpha - x_{k-1}| = L\,|g(\alpha) - g(x_{k-2})|\,, \\ &\leq L^2\,|\alpha - x_{k-2}| = L^2\,|g(\alpha) - g(x_{k-3})|\,, \\ &\vdots \\ &\leq L^k\,|\alpha - x_0| < \rho, \end{aligned} \tag{4.16}$$

e quindi x_k è nell'intervallo $[\alpha - \rho, \alpha + \rho]$.

(ii) sempre dalla (4.16) e dal fatto che $0 \leq L < 1$ segue che $\lim_{k \mapsto \infty} |\alpha - x_k| = 0$.

(iii) siano α e β due punti fissi allora dalla (4.15) abbiamo

$$|\alpha - \beta| = |g(\alpha) - g(\beta)| \leq L\,|\alpha - \beta| < |\alpha - \beta|\,.$$

Questa contraddizione implica che $\alpha = \beta$. ■

Osservazione 49. Se la funzione $g(x)$ è differenziabile possiamo scrivere

$$g(x) - g(x') = g'(\eta)(x - x'), \qquad \eta \in (x, x')$$
$$|g(x) - g(x')| \leq |g'(\eta)||x - x'|,$$

e quindi se abbiamo

$$|g'(\zeta)| \leq L < 1, \qquad \zeta \in [\alpha - \rho, \alpha + \rho],$$

dove α è un punto fisso di $g(x)$ allora per il teorema precedente le iterazioni (4.14) convergono ad α.

Esempio 34. Considerando il metodo di Newton-Raphson come uno schema a punto fisso abbiamo

$$g(x) = x - \frac{f(x)}{f'(x)}, \qquad g'(x) = \frac{f(x)f''(x)}{(f'(x))^2},$$

Se α è una radice semplice di $f(x)$ cioè $f(\alpha) \neq 0$ allora

$$g'(\alpha) = \frac{f(\alpha)f''(\alpha)}{(f'(\alpha))^2} = 0,$$

se $f(x)$ è una funzione C^2 allora $g'(x)$ è continua nell'intorno di α e scegliendo un intorno sufficientemente piccolo di α avremo

$$|g'(x)| \leq L < 1, \qquad x \in [\alpha - \rho, \alpha + \rho]$$

e quindi per il teorema precedente lo schema di Newton-Raphson è convergente perlomeno purché si parti da una approssimazione sufficientemente vicina alla radice e la radice sia semplice.

Osservazione 50. Se α è un punto fisso di $g(x)$ e inoltre $g'(\alpha) = 0$ è chiaro che nella iterazione (4.14) più ci avviciniamo alla radice α più piccola è la costante L e quindi più rapidamente ci avviciniamo al punto fisso. Supponiamo ora che $g''(x)$ esista e sia continua, allora sviluppando con Taylor attorno al punto fisso α la funzione $g(x)$ otteniamo:

$$g(x) = g(\alpha) + 0 + \frac{(x - \alpha)^2}{2} g''(\eta),$$

e da questa

$$|x_{k+1} - \alpha| = |g(x_k) - g(\alpha)| = \left| \frac{(x_k - \alpha)^2}{2} g''(\eta_k) \right| \leq \left| \frac{g''(\eta_k)}{2} \right| |x_k - \alpha|^2 . \qquad (4.17)$$

Da questa equazione segue che l'errore alla iterata successiva è proporzionale al quadrato dell'errore precedente. In tal caso diremo che il metodo è del *secondo ordine*. Quindi in particolare il metodo di Newton-Raphson è del secondo ordine nel caso di radici semplici.

Sia ora $I \equiv [\alpha - \rho, \alpha + \rho]$ un intervallo in cui lo schema è convergente e

$$\left| \frac{g''(x)}{2} \right| \leq M, \qquad x \in I$$

allora dalla (4.17)

$$\begin{aligned}
|x_k - \alpha| &\leq M |x_{k-1} - \alpha|^2 , \\
&\leq M \left(M |x_{k-2} - \alpha|^2 \right)^2 = M \cdot M^2 |x_{k-2} - \alpha|^4 , \\
&\leq M \cdot M^2 \left(M |x_{k-3} - \alpha|^2 \right)^4 = M \cdot M^2 \cdot M^4 |x_{k-3} - \alpha|^8 , \\
&\vdots \\
&\leq M^{1+2+4+8+\cdots+2^{k-1}} |x_0 - \alpha|^{2^k} ,
\end{aligned}$$

osserviamo che $1 + 2 + 4 + \cdots + 2^{k-1} = 2^k - 1$ e quindi

$$|x_k - \alpha| \leq (M |x_0 - \alpha|)^{2^{k-1}} |x_0 - \alpha| .$$

Scegliamo x_0 sufficientemente vicino ad α in modo che

$$M |x_0 - \alpha| < 1,$$

allora possiamo calcolare il numero di iterazioni necessarie per ridurre l'errore iniziale $|x_0 - \alpha|$ di ad esempio 10^{-m} ponendo

$$(M |x_0 - \alpha|)^{2^{k-1}} \approx 10^{-m}, \qquad (4.18)$$

e prendendo il logaritmo da entrambe le parti

$$k \approx \frac{1}{\log_{10} 2} \log \left(\frac{m}{-\log_{10}(M |x_0 - \alpha|)} \right) ,$$

osserviamo che sempre dalla (4.18) otteniamo

$$m \approx 2^k \log_{10}\left(1/(M\,|x_0-\alpha|)\right), \tag{4.19}$$

m è proporzionale al numero di cifre esatte nella approssimazione di α. quindi la (4.19) significa che ad ogni iterazione un metodo al secondo ordine circa raddoppia le cifre esatte.

Osservazione 51. Supponiamo che sia α un punto fisso di $g(x)$ e inoltre $g \in C^p$ con

$$g'(\alpha) = g''(\alpha) = \cdots = g^{(p-1)}(\alpha) = 0,$$

per il teorema di Taylor

$$g(x) = g(\alpha) + \frac{(x-\alpha)^p}{p!} g^{(p)}(\eta),$$

anche qui avremo

$$|x_{k+1}-\alpha| = |g(x_k)-g(\alpha)| \le \frac{\left|g^{(p)}(\eta)\right|}{p!}\,|x-\alpha|^p\,.$$

Da questa equazionesegue che l'errore alla iterata successiva è proporzionale alla $p-$esima potenza dell'errore precedente. In tal caso diremo che il metodo è *di ordine* p.

Esempio 35. Nel metodo di Newton-Raphson se α è una radice multipla di $f(x)$ cioè

$$f(x) = (x-\alpha)^n h(x), \qquad n > 1$$

dove $h(\alpha) \neq 0$ segue che

$$\begin{aligned} f'(x) &= n(x-\alpha)^{n-1}h(x) + (x-\alpha)^n h'(x) \\ f''(x) &= n(n-1)(x-\alpha)^{n-2}h(x) + 2n(x-\alpha)^{n-1}h'(x) + (x-\alpha)^n h''(x), \end{aligned}$$

e di conseguenza

$$\begin{aligned} g'(x) &= \frac{f(x)f''(x)}{(f'(x))^2} = \frac{h(x)\left(n(n-1)h(x) + 2n(x-\alpha)h'(x) + (x-\alpha)^2 h''(x)\right)}{\left(nh(x) + (x-\alpha)h'(x)\right)^2}, \\ g'(\alpha) &= \frac{n(n-1)h(\alpha)^2}{n^2 h(\alpha)^2} = 1 - \frac{1}{n}, \end{aligned}$$

quindi

$$|g'(\alpha)| = 1 - \frac{1}{n} < 1$$

e lo schema iterativo converge localmente anche se ora la convergenza è solo al primo ordine.

4.7 Zeri di polinomi

4.7.1 Eliminazione delle radici multiple

Il metodo di Newton-Raphson è un metodo generalmente del secondo ordine, ma se vogliamo approssimare una radice multipla di un polinomio la convergenza degrada al primo ordine. Un modo per evitare questo degrado delle prestazioni è quello di sostituire al polinomio che contiene radici multiple un polinomio con le stesse radici ma semplici. Questo problema apparentemente complicato ha una soluzione molto semplice: Sia $p(x)$ un polinomio monico fattorizzato come segue[3]

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

allora il polinomio derivata prima si scriverà come

$$p'(x) = q(x) \prod_{i=1}^{k} (x - x_i)^{m_i - 1},$$

$$q(x) = \sum_{i=1}^{k} m_i \prod_{j=1}^{k\ (i)} (x - x_i),$$

osserviamo che il polinomio $q(x)$ non ha radici in comune con il polinomio $p(x)$ infatti

$$q(x_\ell) = \sum_{i=1}^{k} m_i \prod_{j=1}^{k\ (i)} (x_\ell - x_i) = m_\ell \prod_{j=1}^{k\ (\ell)} (x_\ell - x_i) \neq 0, \qquad \ell = 1, 2, \ldots, k$$

Quindi il polinomio $p(x)$ e $p'(x)$ hanno in comune solo le radici di $p(x)$ con molteplicità maggiore di 1. Se x_i è una radice comune con molteplicità m_i allora la sua molteplicità in $p'(x)$ sarà $m_i - 1$.

[3] nella formula $\prod_{j=1}^{k\ (i)}$ significa che vengono fatti tutti i prodotti con indice j che va da 1 a k escluso i

Ricordiamo che un polinomio $M(x)$ è il massimo comun divisore tra due polinomi $P(x)$ e $Q(x)$ se vale

1. $M(x)$ divide $P(x)$ e $Q(x)$, cioè

$$P(x) = M(x)S(x), \qquad Q(x) = M(x)T(x), \tag{4.20}$$

   dove $S(x)$ e $T(x)$ sono a loro volta polinomi (anche di grado 0)

2. Se $N(x)$ è un altro polinomio che divide $P(x)$ e $Q(x)$ allora $N(x)$ divide $M(x)$ cioè

$$M(x) = W(x)N(x),$$

   per $W(x)$ opportuno polinomio.

Teorema 75. *Se $M(x)$ è il massimo comun divisore tra due polinomi $P(x)$ e $Q(x)$ e se α è una radice comune a $P(x)$ e $Q(x)$ cioè:*

$$P(\alpha) = 0, \qquad Q(\alpha) = 0,$$

allora necessariamente α è anche una radice di $M(x)$.

Dimostrazione. Se così non fosse dalla (4.20) avremo che il polinomio $\mathbf{x} - \alpha$ divide sia $S(x)$ che $T(x)$ cioè

$$P(x) = M(x)(\mathbf{x} - \alpha)S^1(x), \qquad Q(x) = M(x)(\mathbf{x} - \alpha)T^1(x),$$

e $M(x)(\mathbf{x} - \alpha)$ sarebbe un divisore comune a $S(x)$ e $T(x)$ che non divide $M(x)$. ■

Teorema 76. *Se α è una radice di $M(x)$ massimo comun divisore tra i polinomi $P(x)$ e $Q(x)$ allora α è una radice comune a $P(x)$ e $Q(x)$ cioè:*

$$P(\alpha) = 0, \qquad Q(\alpha) = 0,$$

Dimostrazione. dalla (4.20) segue immediatamente

$$P(\alpha) = M(\alpha)S(\alpha) = 0,$$
$$Q(\alpha) = M(\alpha)T(\alpha) = 0,$$

Conseguenza di questi due teoremi è che il massimo comun divisore tra due polinomi è costituito dal prodotto delle radici in comune con molteplicità il minimo tra le due.

Esempio 36. Siano

$$P(x) = 3(x-1)^3(x+1)^2(x-3),$$
$$Q(x) = 2(x-1)^2(x+1)(x-3)^3,$$

il loro massimo comun divisore è

$$M(x) = (x-1)^2(x+1)(x-3).$$

Sia $m(x)$ il massimo comun divisore tra i polinomi $p(x)$ e $p'(x)$ allora per i discorsi precedenti vale

$$m(x) = \prod_{i=1}^{k}(x-x_i)^{m_i-1}.$$

Allora il polinomio

$$\frac{p(x)}{m(x)} = \prod_{i=1}^{k}(x-x_i),$$

conterrà tutte le radici di $p(x)$ e solo radici semplici. Il problema è come trovare questo massimo comun divisore. Per fare questo ci viene in aiuto un algoritmo classico, l'algoritmo di Euclide (Euclide di Alessandria circa 365–300 a.C.) per il calcolo del massimo comun divisore.

Algorithm *Algoritmo di Euclide per il M.C.D. di* $p(x)$ *e* $q(x)$
Input: $p(x)$, $q(x)$
1. **if** $\partial p > \partial q$
2. **then** $p_0 \leftarrow p$; $p_1 \leftarrow q$
3. **else** $p_0 \leftarrow q$; $p_1 \leftarrow p$
4. $i \leftarrow 1$
5. **repeat**
6. ($*$ p_{i+1} è il resto della divisione di p_{i-1} con p_i. $*$)
7. ($*$ $p_{i-1}(x) = s_i(x)p_i(x) + p_{i+1}(x)$ $*$)
8. $i \leftarrow i+1$
9. **until** $p_i \equiv 0$
10. ($*$ p_{i-1} è il massimo comun divisore. $*$)

Esempio 37. Sia dato

$$\begin{aligned} p(x) &= (x-1)^3(x-2)(x+1)^2, \\ &= x^6 - 3x^5 + 6x^3 - 3x^2 - 3x + 2, \\ p'(x) &= 6x^5 - 15x^4 + 18x^2 - 6x - 3, \end{aligned}$$

calcoliamo il massimo comun divisore $m(x)$ con l'algoritmo di Euclide:

1. Inizializzazione:

$$\begin{aligned} p_0(x) &= x^6 - 3x^5 + 6x^3 - 3x^2 - 3x + 2, \\ p_1(x) &= 6x^5 - 15x^4 + 18x^2 - 6x - 3, \end{aligned}$$

2. Calcolo p_2:

$$\begin{aligned} p_0(x) &= p_1(x)s_1(x) + p_2(x), \\ x^6 - 3x^5 + 6x^3 - 3x^2 - 3x + 2 &= \left(6x^5 - 15x^4 + 18x^2 - 6x - 3\right)\left(\frac{x}{6} - \frac{1}{12}\right) \\ &\quad + \left(\frac{7}{4} - 3x - \frac{x^2}{2} + 3x^3 - \frac{5}{4}x^4\right). \end{aligned}$$

3. Calcolo p_3:

$$\begin{aligned} p_1(x) &= p_2(x)s_1(x) + p_3(x), \\ 6x^5 - 15x^4 + 18x^2 - 6x - 3 &= \left(-\frac{24x}{5} + \frac{12}{25}\right)\left(\frac{7}{4} - 3x - \frac{x^2}{2} + 3x^3 - \frac{5}{4}x^4\right) \\ &\quad + \frac{96}{25}\left(-x^3 + x^2 + x - 1\right). \end{aligned}$$

4. Calcolo p_4:

$$\begin{aligned} p_2(x) &= p_3(x)s_1(x) + p_4(x), \\ 6x^5 - 15x^4 + 18x^2 - 6x - 3 &= \left(\frac{125x}{384} - \frac{175}{384}\right)\frac{96}{25}\left(-x^3 + x^2 + x - 1\right) + 0. \end{aligned}$$

5. Poiché $p_4 \equiv 0$ il massimo comun divisore è p_3

$$\frac{96}{25}\left(-x^3 + x^2 + x - 1\right),$$

e poiché è unico a meno di una costante moltiplicativa scegliamo

$$m(x) = -x^3 + x^2 + x - 1.$$

dividendo $p(x)$ per $m(x)$ otteniamo

$$\frac{p(x)}{m(x)} = -\frac{x^3}{2} + x^2 + \frac{x}{2} - 1.$$

Per questioni computazionali (ad esempio nella costruzione di successioni di Sturm) a volte è più conveniente usare una versione modificata dell'algoritmo tenendo cento del fatto che se $m(x)$ è il massimo comun divisore tra due polinomi anche $cm(x)$ lo è dove c è un qualunque scalare non nullo. La versione è la seguente

Algorithm *Algoritmo di Euclide per il M.C.D. di $p(x)$ e $q(x)$ (versione modificata)*
Input: $p(x)$, $q(x)$
1. **if** $\partial p > \partial q$
2. **then** $p_0 \leftarrow p$; $p_1 \leftarrow q$
3. **else** $p_0 \leftarrow q$; $p_1 \leftarrow p$
4. $i \leftarrow 1$
5. **repeat**
6. ($*$ p_{i+1} è il resto della divisione di p_{i-1} con p_i. $*$)
7. ($*$ $p_{i-1}(x) = s_i(x)p_i(x) - c_i p_{i+1}(x)$ $*$)
8. $i \leftarrow i + 1$
9. **until** $p_i \equiv 0$
10. ($*$ p_{i-1} è il massimo comun divisore $*$)

Osserviamo che l'algoritmo *Algoritmo di Euclide per il M.C.D. di $p(x)$ e $q(x)$* produce effettivamente il massimo comun divisore. Vale infatti il seguente teorema:

Teorema 77. *L'algoritmo Algoritmo di Euclide per il M.C.D. di $p(x)$ e $q(x)$ termina in un numero finito m di passi e l'ultimo resto non nullo è il massimo comun divisore dei polinomi.*

Dimostrazione. Innanzitutto osserviamo che l'algoritmo termina in un numero finito di passi. Infatti il resto della divisione ha grado strettamente minore del divisore e quindi ad ogni divisione si riduce il grado dei polinomi coinvolti. Quando il grado è 0 il polinomio è uno scalare. La divisione di un polinomio per uno scalare ha resto nullo e quindi anche

in questo caso estremo l' algoritmo termina. L'algoritmo può essere quindi scritto come segue

$$p_0(x) = s_1(x)p_1(x) - c_1p_2(x), \tag{4.20.0}$$

$$p_1(x) = s_2(x)p_2(x) - c_2p_3(x), \tag{4.20.1}$$

$$\vdots$$

$$p_{k-1}(x) = s_k(x)p_k(x) - c_kp_{k+1}(x), \tag{4.20.k}$$

$$\vdots$$

$$p_{m-2}(x) = s_{m-1}(x)p_{m-1}(x) - c_{m-1}p_m(x), \tag{4.20.m-1}$$

$$p_{m-1}(x) = s_m(x)p_m(x), \tag{4.20.m}$$

e segue subito che $p_m(x)$ l'ultimo resto non nullo divide $p_{m-1}(x)$, $p_{m-2}(x)$,..., $p_0(x)$ e quindi è un divisore comune. Viceversa sia $q(x)$ un divisore di $p_0(x)$ e $p_1(x)$, allora dalla (4.20.0) segue se $q(x)$ divide $p_2(x)$ e dalla (4.20.1) segue che $q(x)$ divide $p_3(x)$ e così via fino ad arrivare alla (4.20.m) da cui segue che $q(x)$ divide $p_m(x)$ e quindi $p_m(x)$ è il massimo comun divisore. ■

4.7.2 Separazione delle radici

E' desiderabile avere il modo per conoscere il numero di *radici reali* in un dato intervallo. Se questo è possibile è possibile tramite un metodo di bisezione separare tutte le radici reali e approssimarle fino alla approssimazione voluta. Questo problema può essere risolto grazie alle successioni che prendono il nome dal suo scopritore Jacques Charles Francois Sturm 1803–1855:

Definizione 58. La successione di funzioni continue definite su $[a, b]$:

$$\mathcal{F}(x) = \{f_0(x), f_1(x), \ldots, f_m(x)\},$$

forma una *successione di Sturm su* $[a, b]$ se vale:

1. $f_0(x)$ ha al più zero semplici in $[a, b]$;
2. $f_m(x)$ non ha zeri su $[a, b]$;
3. per ogni $\alpha \in [a, b]$ tale che $f_k(\alpha) = 0$, allora $f_{k-1}(\alpha)f_{k+1}(\alpha) < 0$;

4. per ogni $\alpha \in [a, b]$ tale che $f_0(\alpha) = 0$, allora $f_0'(\alpha) f_1(\alpha) < 0$;

Definizione 59. Data una successione di Sturm $\mathcal{F}(x) = \{f_0(x), f_1(x), \ldots, f_m(x)\}$ per ogni punti x associamo un vettore di $m+1$ numeri reali. A questo vettore possiamo associare un numero intero detto variazione di segno. Questo numero rappresenta il numero di volte che scorrendo la successione di numeri c'è un cambio di segno. Ad esempio la successione

$$\{\underbrace{1, 0, -2}_{*}, \underbrace{-3, 4}_{*}, 3, 0, 1\},$$

ha 2 variazioni di segno (marcate con $*$). Osserviamo che $3, 0, 1$ ha variazione di segno nulla. Infatti lo 0 va considerato come elemento neutro e va rimosso dal computo.

Per ogni successione di Sturm vale il seguente teorema:

Teorema 78 (Sturm). *Data una successione di Sturm $\mathcal{F}(x) = \{f_0(x), f_1(x), \ldots, f_m(x)\}$ su $[a, b]$ il numero di zeri di $f_0(x)$ in (a, b) è dato dalla differenza tra il numero di variazioni di segno in $\mathcal{F}(b)$ e $\mathcal{F}(a)$.*

Dimostrazione. Il numero di variazioni di segno cambia man mano che x passa da a a b solo a causa del cambio di segno di qualche funzione della successione di Sturm. Assumiamo che per un qualche $\hat{x} \in (a, b)$ valga $f_k(\hat{x}) = 0$ per $0 < k < m$. In un intorno di $\hat{x}$ dalla condizione 3 sarà

x	$f_{k-1}(x)$	$f_k(x)$	$f_{k+1}(x)$
$\hat{x} + \epsilon$	$+$	$\pm$	$-$
$\hat{x}$	$+$	0	$-$
$\hat{x} - \epsilon$	$+$	$\pm$	$-$

oppure

x	$f_{k-1}(x)$	$f_k(x)$	$f_{k+1}(x)$
$\hat{x} + \epsilon$	$-$	$\pm$	$+$
$\hat{x}$	$-$	0	$+$
$\hat{x} - \epsilon$	$-$	$\pm$	$+$

in ogni caso il passaggio da $\hat{x}$ non cambia il numero di variazioni di segno. Sia ora $\hat{x}$ uno zero di $f_0(x)$ e vediamo la variazione di segno nell'intorno di $\hat{x}$, osserviamo che dalla condizione 4 della definizione 58 avremo

x	$f_0(x)$	$f_1(x)$
$\hat{x} + \epsilon$	$+$	$+$
$\hat{x}$	0	$+$
$\hat{x} - \epsilon$	$-$	$+$

oppure

x	$f_0(x)$	$f_1(x)$
$\hat{x} + \epsilon$	$-$	$-$
$\hat{x}$	0	$-$
$\hat{x} - \epsilon$	$+$	$-$

in ogni caso in numero delle variazioni di segno cresce al passare di x per uno zero di $f_0(x)$. Combinando queste osservazioni otteniamo il teorema. ■

Costruzione della successione di Sturm per un polinomio

E‘ relativamente facile costruire una successione di Sturm per un polinomio. Sia $f_0 \equiv p_n$ un polinomio di grado n, definiamo $f_1 \equiv -p_n'$. Dividiamo $f_0(x)$ per $f_1(x)$ e chiamiamo il resto $-f_2(x)$. Poi dividiamo $f_1(x)$ per $f_2(x)$ e chiamiamo il resto $-f_3(x)$. Continuiamo così finché il procedimento termina. Abbiamo così la successione:

$$\begin{aligned}
f_0(x) &= q_1(x)f_1(x) - f_2(x),\\
f_1(x) &= q_2(x)f_2(x) - f_3(x),\\
&\vdots\\
f_{k-1}(x) &= q_k(x)f_k(x) - f_{k+1}(x),\\
&\vdots\\
f_{m-2}(x) &= q_{m-1}(x)f_{m-1}(x) - f_m(x),\\
f_{m-1}(x) &= q_m(x)f_m(x).
\end{aligned} \tag{4.21}$$

Questa successione a parte il segno di f_i è esattamente la successione per il calcolo del massimo comun divisore di un polinomio, e f_m è il massimo comun divisore tra f_0 e f_1. La successione di polinomi

$$p_i(x) = \frac{f_i(x)}{f_m(x)}, \qquad i = 0, 1, \ldots, m$$

è una successione di Sturm, infatti

1. $p_0(x)$ ha al più zero semplici in $[a, b]$; essendo il rapporto tra il polinomio originario ed il massimo comun divisore tra il polinomio originario e la sua derivata prima.

2. $p_m(x)$ non ha zeri su $[a, b]$; infatti è una costante.

3. per ogni $\alpha \in [a, b]$ tale che $p_k(\alpha) = 0$, allora $p_{k-1}(\alpha)p_{k+1}(\alpha) < 0$; infatti dalla (4.21) abbiamo

   $$p_{k-1}(\alpha) = -p_{k+1}(\alpha).$$

   Osserviamo che se $p_{k-1}(\alpha) = p_k(\alpha) = p_{k+1}(\alpha) = 0$ allora dalla (4.21) seguirebbe

4. per ogni $\alpha \in [a, b]$ tale che $p_0(\alpha) = 0$, allora $p_0'(\alpha)p_1(\alpha) = -p_0'(\alpha)^2 < 0$;

4.7.3 Limitazione delle radici

Per poter usare il metodo di bisezione e separare le radici con Sturm è necessario conoscere una prima stima anche se grossolana dell' intervallo in cui stanno tutte le radici di un polinomio. Per fare questo occorre osservare che la seguente matrice in forma di Frobenius (Ferdinand Georg Frobenius 1849–1917)

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & & \vdots \\ \vdots & & & \ddots & \ddots & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ \dfrac{a_0}{a_n} & -\dfrac{a_1}{a_n} & \dfrac{a_2}{a_n} & \cdots & (-1)^{n-1}\dfrac{a_{n-2}}{a_n} & (-1)^{n}\dfrac{a_{n-1}}{a_n} \end{bmatrix},$$

ha come polinomio caratteristico

$$p(\lambda) = |\,\mathbf{A} - \lambda\mathbf{I}\,| = \frac{(-1)^n}{a_n}\left(a_0 + a_1\lambda + a_2\lambda^2 + \cdots + a_{n-2}\lambda^{n-2} + a_{n-1}\lambda^{n-1} + a_n\lambda^n\right), \tag{4.22}$$

(basta sviluppare lungo l' ultima riga). Applicando il teorema di Gershgorin alla matrice otteniamo che le radici del polinomio (4.22) soddisfano

$$\left|\lambda - \frac{a_{n-1}}{a_n}\right| \le \left|\frac{a_0}{a_n}\right| + |a_1| + \cdots + \left|\frac{a_{n-2}}{a_n}\right|, \qquad |\lambda| \le 1.$$

Applicando il teorema di Gershgorin alla matrice trasposta otteniamo che le radici del polinomio (4.22) soddisfano

$$|\lambda| \le \left|\frac{a_0}{a_n}\right|, \qquad |\lambda| \le \left|\frac{a_i}{a_n}\right| + 1, \qquad \left|\lambda - \frac{a_{n-1}}{a_n}\right| \le 1.$$

Possiamo applicare le disegualianze appena viste (indebolendole un poco) ad un polinomio qualunque,

$$p(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_{n-1}x^{n-1} + a_n x^n,$$

ottenendo le seguenti limitazioni

$$|x| \le \max\left\{1, \sum_{i=0}^{n-1}\left|\frac{a_i}{a_n}\right|\right\},$$

oppure

$$|x| \leq \max\left\{\left|\frac{a_0}{a_n}\right|, 1 + \max_{i=1}^{n-1}\left|\frac{a_i}{a_n}\right|\right\}.$$

CAPITOLO

CINQUE

INTERPOLAZIONE POLINOMIALE

5.1 Interpolazione polinomiale

Teorema 79. *Siano $x_0, x_1, \ldots, x_n \in \mathbb{R}$, $n+1$ punti distinti ed $f : \mathbb{R} \mapsto \mathbb{R}$ una funzione, allora* esiste un unico *polinomio di grado al più n, $p(x)$ per cui vale*

$$p(x_i) = f(x_i), \qquad i = 0, 1, \ldots, n$$

Dimostrazione. Consideriamo un generico polinomio di grado al più n,

$$p(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n,$$

ed imponiamo che soddisfi le $n+1$ condizioni della forma $(x_i, f(x_i))$ per $i = 0, 1, \ldots, n$. Si ottiene il sistema lineare

$$\begin{cases} a_0 + a_1 x_0 + a_2 x_0^2 + \cdots + a_n x_0^n &= f(x_0) \\ a_0 + a_1 x_1 + a_2 x_1^2 + \cdots + a_n x_1^n &= f(x_1) \\ \vdots \qquad\qquad\qquad\qquad\qquad\quad \vdots & \quad \vdots \\ a_0 + a_1 x_n + a_2 x_n^2 + \cdots + a_n x_n^n &= f(x_n) \end{cases}$$

Il sistema si scrive in forma compatta matriciale $\mathbf{Ma} = \mathbf{b}$ con

$$\mathbf{M} = \begin{bmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ \vdots & & & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{bmatrix}, \quad \mathbf{a} = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} f(x_0) \\ f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix},$$

quindi il problema ha una unica soluzione se e solo se il determinante della matrice $\mathbf{M}$ è non nullo.

La matrice $\mathbf{M}$ è nota col nome di matrice di Vandermonde[1] ed il suo determinante ha la forma seguente:

$$\boxed{\prod_{i>j}^{n} (x_i - x_j)} \tag{5.1}$$

[1] Alexandre Théophile Vandermonde 1735–1796, l'attribuzione è dovuta a Henri Léon Lebesgue 1875–1941, tuttavia questa matrice non compare nei lavori di Vandermonde.

Dall'ipotesi che $x_i \neq x_j$ per $i \neq j$ (punti distinti) segue che il determinante di M è non nullo ed il problema ha una unica soluzione.

Per completare la dimostrazione calcoliamo il determinante (5.1). Sia

$$V(x_0, x_1, \ldots, x_{n-1}, x) = \begin{vmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ \vdots & & & & \vdots \\ 1 & x_{n-1} & x_{n-1}^2 & \cdots & x_{n-1}^n \\ 1 & x & x^2 & \cdots & x^n \end{vmatrix}. \tag{5.2}$$

Sviluppando il determinante in (5.2) per cofattori sull'ultima riga si ottiene un polinomio nella variabile x di grado al più n. Osserviamo che sostituendo al posto di x successivamente i valori x_0, x_1,... fino a x_{n-1} il determinante si annulla, cioè

$$V(x_0, x_1, \ldots, x_{n-1}, x_k) = 0, \qquad k = 0, 1, \ldots, n-1.$$

Ciò accade perché si sta calcolando il determinante di una matrice con due righe uguali, l'ultima e di volta in volta la $k-esima$. Le ascisse x_0, x_1, ..., x_{n-1} sono radici del polinomio $V(x_0, x_1, \ldots, x_{n-1}, x)$, che può essere quindi fattorizzato nella forma

$$K(x - x_0)(x - x_1) \cdots (x - x_{n-1}), \tag{5.3}$$

dove lo scalare K è il coefficiente del termine di grado massimo.

Infatti, moltiplicando i binomi in (5.3)

$$V(x_0, x_1, \ldots, x_{n-1}, x) = Kx^n + Lx^{n-1} \cdots$$

si vede subito che K è il coefficiente che moltiplica x^n.

Il determinante della matrice di Vandermonde coincide con il valore del polinomio in (5.3) calcolato nell'ascissa x_n, per cui si rende necessario determinare esplicitamente il coefficiente K. Sviluppando $V(x_0, x_1, \ldots, x_{n-1}, x)$ per minori rispetto alla ultima riga

otteniamo

$$V(x_0, x_1, \ldots, x_{n-1}, x) = \begin{vmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ \vdots & & & & \vdots \\ 1 & x_{n-1} & x_{n-1}^2 & \cdots & x_{n-1}^n \\ 1 & x & x^2 & \cdots & x^n \end{vmatrix}$$

$$= x^n \begin{vmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^{n-1} \\ 1 & x_1 & x_1^2 & \cdots & x_1^{n-1} \\ \vdots & & & & \vdots \\ 1 & x_{n-1} & x_{n-1}^2 & \cdots & x_{n-1}^{n-1} \end{vmatrix} + x^{n-1} \cdots$$

da cui segue immediatamente che

$$K = \begin{vmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^{n-1} \\ 1 & x_1 & x_1^2 & \cdots & x_1^{n-1} \\ \vdots & & & & \vdots \\ 1 & x_{n-1} & x_{n-1}^2 & \cdots & x_{n-1}^{n-1} \end{vmatrix} = V(x_0, x_1, \ldots, x_{n-1}).$$

Sostituendo questa espressione per K in (5.3) abbiamo

$$V(x_0, x_1, \ldots, x_{n-1}, x_n) = V(x_0, x_1, \ldots, x_{n-1})(x_n - x_0)(x_n - x_1) \cdots (x_n - x_{n-1}),$$

in cui è facile riconoscere una relazione di ricorrenza

$$\begin{aligned} V(x_0, x_1, \ldots, x_{n-1}, x_n) &= V(x_0, x_1, \ldots, x_{n-1}) \prod_{i=0}^{n-1} (x_n - x_i), \\ &= V(x_0, x_1, \ldots, x_{n-2}) \left(\prod_{i=0}^{n-2} (x_{n-1} - x_i) \right) \left(\prod_{i=0}^{n-1} (x_n - x_i) \right), \\ &\vdots \\ &= V(x_0, x_1, \ldots, x_k) \prod_{j=k}^{n} \prod_{i=0}^{j-1} (x_j - x_i), \qquad k \geq 1. \end{aligned} \tag{5.4}$$

La ricorsione all'indietro si arresta sull'indice $k = 1$

$$V(x_0, x_1) = \begin{vmatrix} 1 & x_0 \\ 1 & x_1 \end{vmatrix} = x_1 - x_0.$$

che in (5.4) produce proprio la formula (5.1). ■

Definizione 60 (Nodi di interpolazione). Le coppie ordinate della forma $(x_i,\ y_i)$ con gli x_i assunti distinti, cioè $x_i \neq x_j$ per $i \neq j$, si indicano normalmente col termine di *nodi di interpolazione*.

Se necessario, si può anche supporre che $y_i = f(x_i)$, cioè che i valori assunti nei nodi di interpolazione siano tabulati a partire da una funzione nota $y = f(x)$.

Definizione 61 (Gradi di libertà). Gli $n+1$ coefficienti a_i per $i = 0, 1, \ldots n$ che determinano il polinomio di grado n si indicano normalmente col termine di *gradi di libertà*. Il numero a_n si chiama *coefficiente del termine di grado massimo*, ed è ovviamente assunto non nullo quando il polinomio è di grado n.

Osservazione 52. Il teorema 79 afferma che dati $n+1$ nodi di interpolazione (distinti) il polinomio interpolante di grado al piú n esiste ed è unico. Nel teorema è stato scritto come combinazione lineare dei polinomi $1,\ x,\ x^2,\ldots,\ x^n$. Tuttavia può essere scritto anche come combinazione lineare di funzioni polinomiali generali, che indicheremo con i simboli $L_0(x),\ L_1(x),\ldots,L_n(x)$.

5.1.1 Generalizzazione

Esaminiamo cosa succede se consideriamo una funzione della forma:

$$L(x) = a_0 L_0(x) + a_1 L_1(x) + \cdots + a_n L_n(x),$$

per cui valga

$$L(x_k) = f(x_k), \qquad k = 0, 1, \ldots, n$$

Imponendo esplicitamente il passaggio per i nodi di interpolazione si ha un sistema lineare

$$\begin{cases} a_0 L_0(x_0) + a_1 L_1(x_0) + \cdots + a_n L_n(x_0) &= f(x_0) \\ a_0 L_0(x_1) + a_1 L_1(x_1) + \cdots + a_n L_n(x_1) &= f(x_1) \\ \vdots & \vdots \quad \vdots \\ a_0 L_0(x_n) + a_1 L_1(x_n) + \cdots + a_n L_n(x_n) &= f(x_n) \end{cases}$$

che scriviamo in forma matriciale compatta $\mathbf{Ma} = \mathbf{b}$ con

$$\mathbf{M} = \begin{bmatrix} L_0(x_0) & L_1(x_0) & L_1(x_0) & \cdots & L_n(x_0) \\ L_0(x_1) & L_1(x_1) & L_2(x_1) & \cdots & L_n x_1) \\ \vdots & & & & \vdots \\ L_0(x_n) & L_1(x_n) & L_2(x_n) & \cdots & L_n(x_n) \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} f(x_0) \\ f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix},$$

nelle incognite $\mathbf{a} = (a_0,\ a_1,\ \ldots,\ a_n)^T$. Quindi la soluzione esiste ed è unica se e solo se la matrice $\mathbf{M}$ ha determinante non nullo.

5.1.2 Condizione di Haar

Definizione 62. Date $n+1$ funzioni polinomiali generali $L_0(x)$, $L_1(x)$,...,$L_n(x)$ ed $n+1$ nodi di interpolazione con ascisse (x_i) supposte distinte, sia $\mathbf{M}$ la matrice di componenti $M_{ij} = L_j(x_i)$. La condizione $\mid M \mid \neq 0$ si chiama *condizione di Haar*[2] .

Osservazione 53. La *condizione di Haar* esprime l'indipendenza lineare delle funzioni polinomiali $L_i(x)$ per $i = 0, 1, \ldots n$ sui nodi di interpolazione ed è *necessaria e sufficiente* per la risolubilità del problema di interpolazione.

5.1.3 Interpolazione di Lagrange

Scegliamo le funzioni generali di interpolazione $L_k(x)$ come

$$L_k(x) = \frac{\omega_{n+1}(x)}{(x - x_k)\omega'_{n+1}(x_k)},$$

dove si è introdotto il polinomio $\omega_{n+1}(x)$ di grado $n+1$

$$\omega_{n+1}(x) = (x - x_0)(x - x_1)\cdots(x - x_n),$$

la cui derivata prima è [3]

$$\omega'_{n+1}(x) = \sum_{k=0}^{n} \prod_{i=0}^{n}{}^{(k)} (x - x_i).$$

[2] Alfréd Haar 1885–1933

[3] L'apice k nell'espressione $\prod_{i=0}^{n}{}^{(k)}$ significa che si esclude dalla produttoria il termine con indice $i = k$.

Si noti che $\omega'_{n+1}(x_k)$, cioè la derivata prima di $\omega_{n+1}(x)$ calcolata nell'ascissa del k-esimo nodo di interpolazione assume una forma particolarmente semplice,

$$\omega'_{n+1}(x_k) = \prod_{i=0}^{n}{}^{(k)} (x_k - x_i).$$

Definizione 63 (Polinomi elementari di Lagrange). [4] Ogni $L_k(x)$ è un polinomio di grado n della forma

$$L_k(x) = \frac{\omega_{n+1}(x)}{(x - x_k)\omega'_{n+1}(x_k)} = \prod_{i=0}^{n}{}^{(k)} \left(\frac{x - x_i}{x_k - x_i}\right).$$

Questi polinomi sono noti come *polinomi elementari di Lagrange*.

Osservazione 54. I polinomi L_k sono completamente caratterizza dalla proprietà che

$$L_k(x_i) = \delta_{ik}.$$

Si ha quindi che $\mathbf{M} = \mathbf{I}$ ed il problema della interpolazione si riduce a $a_k = f(x_k), \quad k = 0, 1 \ldots n$. Il polinomio interpolante di grado n si scrive immediatamente come

$$\boxed{p(x) = \sum_{k=0}^{n} f(x_k) L_k(x) = \sum_{k=0}^{n} f(x_k) \frac{\omega_{n+1}(x)}{(x - x_k)\omega'_{n+1}(x_k)}}$$

5.1.4 Interpolazione di Newton

Nel caso dell'algoritmo di Newton[5] scegliamo le funzioni generali di interpolazione $L_k(x)$ per $k = 1, 2, \ldots n$ come segue:

$$\begin{aligned} L_k(x) &= \omega_k(x), \\ &= (x - x_0)(x - x_1) \cdots (x - x_{k-1}), \end{aligned}$$

e poniamo per coerenza $\omega_0 \equiv 1$ per l'indice $k = 0$. Il polinomio interpolatore si scrive quindi come

$$p(x) = a_0\omega_0(x) + a_1\omega_1(x) + \cdots + a_n\omega_n(x).$$

[4] Joseph-Louis Lagrange 1736–1813

[5] Sir Isaac Newton 1643–1727

Osservando che $\omega_k(x_i) = 0$ per $i \leq k$ il problema di interpolazione si può esprimere come sistema lineare

$$\begin{cases} f(x_0) &= a_0 \\ f(x_1) &= a_0 + a_1\omega_1(x_1) \\ f(x_2) &= a_0 + a_1\omega_1(x_2) + a_2\omega_2(x_2) \\ & \vdots \\ f(x_n) &= a_0 + a_1\omega_1(x_n) + \cdots + a_n\omega_n(x_n) \end{cases}$$

o equivalentemente in forma matriciale compatta $\mathbf{Ma} = \mathbf{b}$ con

$$\mathbf{M} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & \omega_1(x_1) & 0 & \cdots & 0 \\ \vdots & & & & \vdots \\ 1 & \omega_1(x_n) & \omega_2(x_n) & \cdots & \omega_n(x_n) \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} f(x_0) \\ f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix},$$

nelle incognite $\mathbf{a} = (a_0, a_1, \ldots, a_n)^T$.

La matrice $\mathbf{M}$ in questo caso è una matrice triangolare inferiore, il cui determinante

$$|\,\mathbf{M}\,| = \omega_0(x_0)\omega_1(x_1)\cdots\omega_n(x_n),$$

è uguale al prodotto degli elementi diagonali $\omega_k(x_k)$ dove

$$\omega_k(x_k) = \prod_{i=0}^{k-1}(x_k - x_i).$$

Se assumiamo come sempre che i nodi di interpolazione hanno ascisse distinte, cioè $x_i \neq x_j$ per $i \neq j$, allora il determinante è non nullo ed il problema di interpolazione ammette una unica soluzione.

Relazioni di ricorrenza Sia la $n+1$-upla di coefficienti $(a_0, a_1, \ldots, a_n)$ una soluzione del problema di interpolazione. Allora gli $n+1$ polinomi

$$p_k(x) = a_0\omega_0(x) + a_1\omega_1(x) + \cdots + a_k\omega_k(x),$$

per $k = 0, 1, 2, \ldots, n$ interpolano i nodi $(x_0, f(x_0))$, $(x_1, f(x_1))$,…,$(x_k, f(x_k))$.

Ovviamente, per $k = n$ ritroviamo il polinomio di interpolazione, cioè $p_n(x) = p(x)$.

Per $i \leq k$ si ha

$$\omega_{k+1}(x_i) = \omega_{k+1}(x_i) = \ldots = \omega_n(x_i) = 0.$$

Infatti, se per esempio esaminiamo il primo di questi polinomi, quello con indice $k+1$, notiamo che

$$\omega_{k+1}(x_i) = (x_i - x_0)(x_i - x_1)\ldots(x_i - x_k),$$

per cui, essendo $i \leq k$, uno dei binomi è sicuramente nullo. Lo stesso ragionamento si può ripetere anche per gli altri polinomi con indici $k+2$, $k+3$, $\ldots n$.

Il polinomio interpolatore calcolato nel nodo x_i si scrive formalmente come

$$\begin{aligned} p_k(x_i) &= \underbrace{a_0\omega_0(x_i) + a_1\omega_1(x_i) + \cdots + a_k\omega_k(x_i)}_{=p(x_i)} + \\ &+ \underbrace{a_{k+1}\omega_{k+1}(x_i) + \cdots + a_n\omega_n(x_i)}_{=0}, \\ &= p(x_i), \qquad i = 1, 2, \ldots, k. \end{aligned}$$

Possiamo esprimere questa proprietà per mezzo delle seguenti *relazioni di ricorrenza* per $k \geq 0$

$$\begin{aligned} \omega_{k+1}(x) &= (x - x_k)\omega_k(x), \\ p_{k+1}(x) &= p_k(x) + a_{k+1}\omega_{k+1}(x). \end{aligned}$$

La condizione di interpolazione $p_{k+1}(x_{k+1}) = f(x_{k+1})$ permette di ricavare immediatamente un'espressione per a_{k+1}, cioè per il *coefficiente del termine di grado massimo* del polinomio $p_{k+1}(x)$

$$a_{k+1} = \frac{f(x_{k+1}) - p_k(x_{k+1})}{\omega_{k+1}(x_{k+1})},$$

che sostituita nelle relazioni di ricorrenza permette di scrivere un algoritmo di calcolo per il polinomio di interpolazione.

5.1.5 Algoritmo di Newton

1. Si inizializzi $\omega_0(x) = 1$, $p_0(x) = f(x_0)$.
2. e per $k = 0, 1, \ldots, n-1$ si calcoli

$$\begin{aligned} \omega_{k+1}(x) &= (x - x_k)\omega_k(x), \\ p_{k+1}(x) &= p_k(x) + \frac{f(x_{k+1}) - p_k(x_{k+1})}{\omega_{k+1}(x_{k+1})}\omega_{k+1}(x). \end{aligned}$$

5.1.6 Differenze divise

Il sistema lineare che definisce il problema di interpolazione secondo il metodo di Newton

$$\begin{cases} f(x_0) = a_0 \\ f(x_1) = a_0 + a_1\omega_1(x_1) \\ f(x_2) = a_0 + a_1\omega_1(x_2) + a_2\omega_2(x_2) \\ \qquad \vdots \\ f(x_n) = a_0 + a_1\omega_1(x_n) + \cdots + a_n\omega_n(x_n) \end{cases}$$

è triangolare inferiore e può essere risolto esplicitamente con la tecnica delle *sostituzioni in avanti*.

La soluzione del sistema lineare può essere scritta come

$$\begin{aligned} a_0 &= f(x_0), \\ a_1 &= \frac{f(x_1) - f(x_0)}{\omega_1(x_1)}, \\ a_2 &= \frac{f(x_2) - f(x_0) - \omega_1(x_1)\dfrac{f(x_1) - f(x_0)}{\omega_1(x_1)}}{\omega_2(x_2)}, \\ &\vdots \end{aligned}$$

Si noti che ogni a_i dipende dai nodi di interpolazione $(x_j, f(x_j))$ per $j = 0, 1, \ldots i$ e non dai successivi.

Osservazione 55. Questo tipo di dipendenza non deve stupire, perché è ancora un modo per descrivere le relazioni di ricorrenza di cui si è già parlato.

Indichiamo questa dipendenza con il simbolo speciale (notare le parentesi quadre)

$$a_j = f[x_0, x_1, \ldots, x_j],$$

che si chiama *differenza divisa* di ordine j.

Tramite le differenze divise possiamo scrivere il polinomio interpolante come segue

$$\boxed{p(x) = \sum_{k=0}^{n} f[x_0, x_1, \ldots, x_k]\omega_k(x)}$$

Le differenze divise godono di una serie notevole di proprietà interessanti.

- *Proprietà 1*
  La differenza divisa $f[x_0, x_1, \ldots, x_k]$ è il coefficiente del termine di grado massimo k del polinomio interpolatore che interpola i $k+1$ nodi di ascisse $x_0, x_1, \ldots, x_k$.

- *Proprietà 2*
  La differenza divisa $f[x_0, x_1, \ldots, x_k]$ non dipende dall'ordine con cui si prendono i nodi di interpolazione, cioè

  $$f[x_0, x_1, \ldots, x_k] = f[x_{i_0}, x_{i_1}, \ldots, x_{i_k}],$$

  dove $i_0, i_1, \ldots, i_k$ è una qualunque permutazione dei numeri $0, 1, \ldots, k$.

Come conseguenza della proprietà 2 si ha che permutando la sequenza dei nodi di interpolazione il polinomio non cambia. Il problema di interpolazione sui nodi di ascissa $x_0, x_1, \ldots, x_{k-1}, x_k$ ha come soluzione il polinomio che si esprime con le differenze divise come

$$\begin{aligned} p(x) = {} & f[x_0] + \\ & f[x_0, x_1](x - x_0) + \\ & + \cdots + \\ & f[x_0, \ldots, x_{k-2}, x_{k-1}](x - x_0) \cdots (x - x_{k-2}) + \\ & f[x_0, \ldots, x_{k-2}, x_{k-1}, x_k](x - x_0) \cdots (x - x_{k-2})(x - x_{k-1}), \end{aligned}$$

In maniera analoga possiamo costruire *lo stesso* polinomio interpolatore per i nodi di ascissa x_0, x_1,..., x_k, con un ordine di interpolazione diverso, ad esempio x_0, x_1,..., x_k, x_{k-1}.

In questo caso la formula di interpolazione di Newton è

$$\begin{aligned} p(x) = {} & f[x_0] + \\ & f[x_0, x_1](x - x_0) + \\ & + \cdots + \\ & f[x_0, \ldots, x_{k-2}, x_k](x - x_0) \cdots (x - x_{k-2}) + \\ & f[x_0, \ldots, x_{k-2}, x_k, x_{k-1}](x - x_0) \cdots (x - x_{k-2})(x - x_k), \end{aligned}$$

Confrontando le due espressioni (si sottragga membro a membro e si divida per $(x - x_0)(x - x_1) \cdots (x - x_{k-2})$) si ottiene la relazione

$$\begin{aligned} 0 = {} & f[x_0, \ldots, x_{k-2}, x_{k-1}] + f[x_0, \ldots, x_{k-2}, x_{k-1}, x_k](x - x_{k-1}) - \\ & f[x_0, \ldots, x_{k-2}, x_k] - f[x_0, \ldots, x_{k-2}, x_k, x_{k-1}](x - x_k), \end{aligned}$$

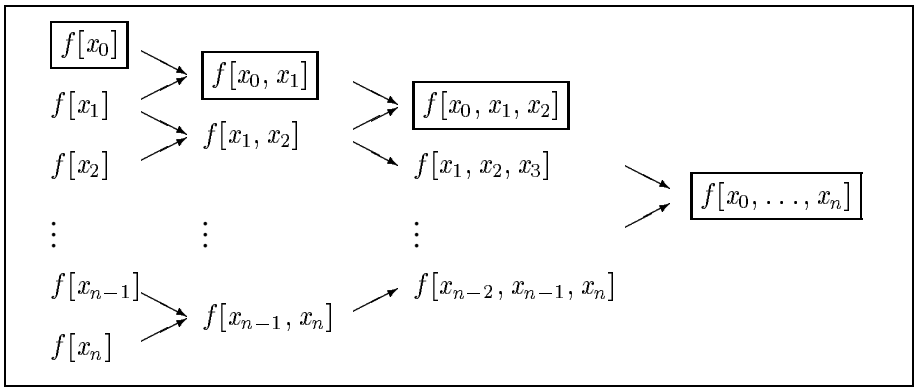
Usando il fatto che $f[x_0, \ldots, x_k, x_{k-1}] = f[x_0, \ldots, x_{k-1}, x_k]$ si ottiene

$$f[x_0, \ldots, x_k, x_{k-1}] = \frac{f[x_0, \ldots, x_{k-2}, x_{k-1}] - f[x_0, \ldots, x_{k-2}, x_k]}{x_k - x_{k-1}},$$

- *Proprietà 3*
  Ripetendo il ragionamento con un ordinamento qualsiasi dei nodi di interpolazione si ottiene una definizione *ricorsiva* delle differenze divise di ogni ordine

$$\begin{aligned} f[x_i] &= f(x_i), \\ f[\ldots, x_i, x_j] &= \frac{f[\ldots, x_i] - f[\ldots, x_j]}{x_i - x_j}. \end{aligned}$$

Tabella 5.1: Schema grafico del procedimento ricorsivo per il calcolo delle differenze divise del polinomio interpolatore di ordine n



5.1.7 L'algoritmo di Aitken-Neville

Per presentare l'algoritmo di Aitken-Neville [6] introduciamo l'insieme formato da $n+1$ punti di $\mathbb{R} \times \mathbb{R}$

$$I = \{(x_i, f(x_i)) \mid i = 0, 1, 2, \ldots, n,\ x_i \neq x_j\},$$

e tutti i suoi sottoinsiemi

$$I_{i_0,i_1,i_2,\ldots,i_k} = \{(x_{i_s}, f(x_{i_s})) \mid s = 0, 1, 2, \ldots, k\},$$

con $k = 0, 1, \ldots, n$.

Ad ogni sottoinsieme $I_{i_0 i_1 i_2 \ldots i_k}$ associamo il polinomio $p_{i_0 i_1 i_2 \ldots i_k}$ di grado k che interpola i punti dell'insieme

$$p_{i_0 i_1 i_2 \ldots i_k}(x_{i_j}) = f(x_{i_j}) \qquad j = 0, 1, \ldots, k.$$

Osservazione 56. Conoscendo i due polinomi interpolatori $p_{i_0 i_1 i_2 \ldots i_{k-1}}(x)$ e $p_{i_1 i_2 \ldots i_{k-1} i_k}(x)$ possiamo facilmente costruire il polinomio interpolatore $p_{i_0 i_1 i_2 \ldots i_{k-1} i_k}(x)$, che è dato dalla formula

$$p_{i_0 i_1 \ldots i_k}(x) = \frac{(x_{i_k} - x)\, p_{i_0 i_1 \ldots i_{k-1}}(x) + (x - x_{i_0})\, p_{i_1 i_2 \ldots i_k}(x)}{x_{i_k} - x_{i_0}}$$

[6] Alexander Craig Aitken 1895–1967

$$\boxed{p_{i_0 i_1 \ldots i_k}(x) = \frac{(x_{i_k} - x)\, p_{i_0 i_1 \ldots i_{k-1}}(x) + (x - x_{i_0})\, p_{i_1 i_2 \ldots i_k}(x)}{x_{i_k} - x_{i_0}}}$$

Per $s = 1, 2, \ldots, k-1$ vale ovviamente

$$\begin{aligned} p_{i_0 i_1 \ldots i_k}(x_{i_s}) &= \frac{(x_{i_k} - x_{i_s})\, p_{i_0 i_1 \ldots i_{k-1}}(x_{i_s}) + (x_{i_s} - x_{i_0})\, p_{i_1 i_2 \ldots i_k}(x_{i_s})}{x_{i_k} - x_{i_0}}, \\ &= \frac{(x_{i_k} - x_{i_s}) f(x_{i_s}) + (x_{i_s} - x_{i_0}) f(x_{i_s})}{x_{i_k} - x_{i_0}}, \\ &= f(x_{i_s}), \end{aligned}$$

$$\boxed{p_{i_0 i_1 \ldots i_k}(x) = \frac{(x_{i_k} - x)\, p_{i_0 i_1 \ldots i_{k-1}}(x) + (x - x_{i_0})\, p_{i_1 i_2 \ldots i_k}(x)}{x_{i_k} - x_{i_0}}}$$

Nel caso $s = 0$ abbiamo

$$\begin{aligned} p_{i_0 i_1 \ldots i_k}(x_{i_0}) &= \frac{(x_{i_k} - x_{i_0})\, p_{i_0 i_1 \ldots i_{k-1}}(x_{i_0}) + (x_{i_0} - x_{i_0})\, p_{i_1, i_2, \ldots, i_k}(x_{i_0})}{x_{i_k} - x_{i_0}}, \\ &= \frac{(x_{i_k} - x_{i_0}) f(x_{i_0})}{x_{i_k} - x_{i_0}} = f(x_{i_0}). \end{aligned}$$

$$\boxed{p_{i_0 i_1 \ldots i_k}(x) = \frac{(x_{i_k} - x)\, p_{i_0 i_1 \ldots i_{k-1}}(x) + (x - x_{i_0})\, p_{i_1 i_2 \ldots i_k}(x)}{x_{i_k} - x_{i_0}}}$$

Nel caso $s = k$ abbiamo

$$\begin{aligned} p_{i_0 i_1 \ldots i_k}(x_{i_k}) &= \frac{(x_{i_k} - x_{i_k})\, p_{i_0, i_1, \ldots, i_{k-1}}(x_{i_k}) + (x_{i_k} - x_{i_0})\, p_{i_1, i_2, \ldots, i_k}(x_{i_k})}{x_{i_k} - x_{i_0}}, \\ &= \frac{(x_{i_k} - x_{i_0}) f(x_{i_k})}{x_{i_k} - x_{i_0}} = f(x_{i_k}). \end{aligned}$$

5.1.8 Osservazioni finali sull'algoritmo di Aitken-Neville

I polinomi $p_i(x)$ con $i = 0, 1, \ldots, n$ sono semplicemente le costanti $f(x_0), f(x_1), \ldots f(x_n)$; cioè:

$$p_i(x) = f(x_i), \qquad i = 0, 1, \ldots, n$$

Analogamente

$$p_{ii+1}(x) = \frac{(x_{i+1} - x)p_i(x) + (x - x_i)p_{i+1}(x)}{x_{i+1} - x_i},$$

per $i = 0, 1, \ldots, n-1$.

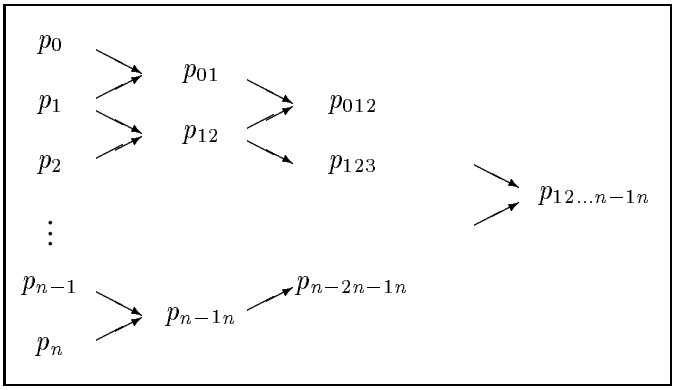
La formula di interpolazione dell'algoritmo di Aitken-Neville può essere utilizzata in modo efficiente per calcolare il valore del polinomio interpolatore in un punto generico $\bar{x}$ come segue:

1. Per $i = 0, 1, \ldots, n$ si pone $p_i(\bar{x}) = f(x_i)$.
2. Per $k = 1, 2, \ldots, n$ si calcola
   (a) Per $i = 0, 1, \ldots, n-k$

$$p_{i\ldots i+k}(\bar{x}) = \frac{(x_{i+k} - \bar{x})p_{i\ldots i+k-1}(\bar{x}) + (\bar{x} - x_i)p_{i+1\ldots i+k}(\bar{x})}{x_{i+k} - x_i}.$$

3. $p_{01\ldots n}(\bar{x})$ restituisce il valore del polinomio interpolatore calcolato in $\bar{x}$.

Tabella 5.2: Schema grafico del procedimento di Aitken-Neville



5.1.9 Errore di interpolazione

Definizione 64. Data la funzione $f(x)$ ed il polinomio $p(x)$ interpolante gli $n+1$ nodi (x_i, f_i) per $i = 0, 1, \ldots n$, definiamo come *errore di interpolazione* nel punto x la differenza (in valore assoluto) tra il valore calcolato con la funzione ed il valore calcolato con il polinomio

$$E(x) = |\, f(x) - p(x) \,|.$$

Osservazione 57. L'accuratezza di ogni procedimento di interpolazione segue da una *stima* dell'errore di interpolazione.

Teorema 80 (dell'errore di interpolazione). *Sia $f \in C^{n+1}[a,b]$ e $p(x)$ il suo polinomio interpolatore negli $n+1$ punti distinti $x_0, x_1, \ldots, x_n$. Allora per ogni x esiste un $\eta \in I(x, x_0, \ldots, x_n)$ tale che:*

$$E(x) = f(x) - p(x) = \frac{f^{(n+1)}(\eta)}{(n+1)!}\omega_{n+1}(x), \tag{5.5}$$

dove $I(x; x_0, \ldots, x_n)$ è il minimo intervallo contenente i punti x, x_0,...,x_n

$$I(x, x_0, \ldots, x_n) = (\min\{x, x_0, \ldots, x_n\}, \max\{x, x_0, \ldots, x_n\}), \qquad e$$

$$\omega_{n+1}(x) = (x - x_0)(x - x_1)\cdots(x - x_n).$$

Dimostrazione. Osserviamo che

$$E(x_i) = f(x_i) - p(x_i) = 0, \qquad i = 0, 1, \ldots, n$$

Quindi potremmo scrivere

$$E(x) = e(x)\omega_{n+1}(x).$$

Consideriamo la seguente funzione

$$G(z; x) := f(z) - p(z) - e(x)\omega_{n+1}(z), \tag{5.6}$$

dove con $G(z; x)$ intendiamo una funzione in z che dipende da un parametro x. Ovviamente per questa funzione abbiamo

$$G(x_i; x) = \underbrace{f(x_i) - p(x_i)}_{=0} - e(x)\underbrace{\omega_{n+1}(x_i)}_{=0} = 0, \qquad i = 0, 1, \ldots, n$$

$$G(x; x) = f(x) - p(x) - e(x)\omega_{n+1}(x) = 0, \qquad \text{per costruzione}$$

quindi $G(z; x)$ si annulla in $n + 2$ punti. Grazie al teorema di Rolle[7] la sua derivata prima (si calcola in z perché x è solo un parametro) si annulla in $n + 1$ punti. Ripetendo il ragionamento per le derivate successive otteniamo che esiste almeno un punto $\eta \in I(x; x_0, \ldots, x_n)$ per cui vale

$$G^{(n+1)}(\eta; x) = 0.$$

[7] Michel Rolle 1652–1719

Ricordiamo che $p^{(n+1)} \equiv 0$, essendo un polinomio di grado n, e che vale $\omega_{n+1}^{(n+1)}(x) = (n+1)!$, per cui si ottiene

$$0 = G^{(n+1)}(\eta; x) = f^{(n+1)}(\eta) - e(x)(n+1)!,$$

e quindi

$$e(x) = \frac{f^{(n+1)}(\eta)}{(n+1)!},$$

che sostituito nella (5.6) produce l'espressione (5.5). ■

Con il teorema precedente si può dare immediatamente una stima dell'errore se si conosce una stima della derivata $(n+1)-$esima come segue

$$|f(x) - p(x)| \leq \sup_{\eta \in I(x; x_0, \ldots, x_n)} \left| \frac{f^{(n+1)}(\eta)}{(n+1)!} \right| |\omega_{n+1}(x)| .$$

5.2 Equazioni Normali e Minimi Quadrati

5.2.1 Equazioni Normali

Sia data la matrice $\mathbf{A} \in \mathbb{K}^{m\times n}$ con $m > n$ e supponiamo che $\mathcal{R}\{\mathbf{A}\} = n$, cioè che $\mathbf{A}$ sia di rango massimo. Dalle considerazioni sul teorema di Rouché-Capelli, segue che il problema $\mathbf{Ax} = \mathbf{b}$ è *sovra-determinato*, quindi *non* ammette soluzione. Se moltiplichiamo a sinistra per $\mathbf{A}^T$, si ottiene

$$\mathbf{A}^T\mathbf{Ax} = \mathbf{A}^T\mathbf{b}.$$

Che cosa sappiamo dire su questo nuovo problema? Osserviamo che la matrice prodotto $\mathbf{A}^T\mathbf{A}$ è *simmetrica* (ovvio) e *definita positiva*:

$$\mathbf{y}^T\mathbf{A}^T\mathbf{Ay} = \left(\mathbf{Ay}\right)^T\mathbf{Ay} = \|\mathbf{Ay}\|_2^{\,2} \geq 0$$

$$\|\mathbf{Ay}\|_2 = 0 \Rightarrow \mathbf{Ay} = \mathbf{0} \Rightarrow \mathbf{y} = \mathbf{0} \qquad [\mathbf{A} \text{ ha rango massimo}]$$

Poiché $\mathbf{A}^T\mathbf{A}$ è SPD, è *non-singolare*, quindi $\mathbf{A}^T\mathbf{Ax} = \mathbf{b}$ ammette *una ed una sola* soluzione $\mathbf{x}$.

Questa "buona" proprietà ci porta ad introdurre la seguente definizione.

Definizione 65. Data una matrice rettangolare $\mathbf{A} \in \mathbb{K}^{m\times n}$ di rango massimo con $m \geq n$, indicheremo il sistema lineare di equazioni della forma $\mathbf{A}^T\mathbf{Ax} = \mathbf{A}^T\mathbf{b}$ col termine di **equazioni normali**.

Osservazione 58. Le equazioni normali si presentano in modo naturale nel problema di approssimazione ai *Minimi Quadrati*.

5.2.2 Minimi Quadrati

Consideriamo il seguente problema di *approssimazione*:

Problema

Trovare la retta che *passa al meglio* per gli n punti del piano che supponiamo *distinti*[8]

$$\mathbf{p}_1 = \begin{bmatrix} x_1 \\ y_1 \end{bmatrix}, \quad \cdots \quad \mathbf{p}_n = \begin{bmatrix} x_n \\ y_n \end{bmatrix}.$$

[8] Si noti che in questa sezione consideriamo n punti numerati partendo da 1 e non, come nella sezione dedicata all'interpolazione polinomiale, $(n+1)$ punti numerati partendo da 0.

Osservazione 59. Dalla geometria analitica nel piano sappiamo che

- per $n = 1$, si ha un solo punto assegnato sul piano cartesiano, che è il centro di un fascio formato da *infinite* rette;
- per $n = 2$ due punti assegnati, esiste *una ed una sola* retta che passa *esattamente* per i due punti dati;
- per $n \geq 2$ piú di due punti, *non esiste* alcuna retta che passa *esattamente* per gli n punti dati.

Dato che non esiste una retta che passa esattamente per tutti i punti assegnati, cercheremo una retta che, come enunciato nel problema, "passi al meglio", qualcosa tipo ciò che è mostrato in figura 5.2.2

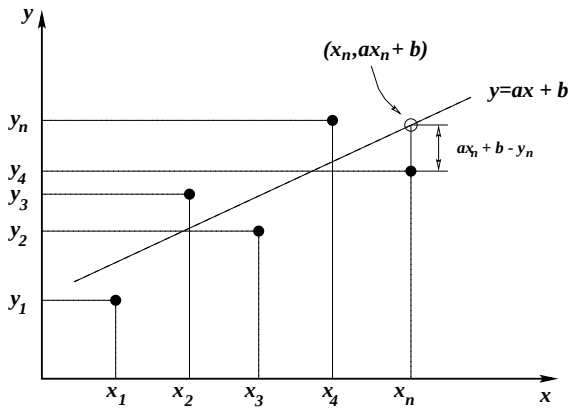


Figura 5.1: Retta ai minimi quadrati.

Che cosa significa *matematicamente* trovare la retta che *passa al meglio*?

Prendiamo una retta generica della forma $y(x) = ax + b$, con a e b coefficienti parametrici da determinare. Se consideriamo per il generico punto i assegnato (x_i, y_i) il valore assunto dalla retta $y(x_i) = ax_i + b \neq y_i$ come approssimazione del valore nodale y_i commettiamo l'errore

$$e_i(a, b) = y(x_i) - y_i = ax_i + b - y_i.$$

Formiamo il vettore degli errori

$$\mathbf{e}^T(a, b) = (e_1, e_2, \ldots, e_n)^T,$$

e scegliamo (a, b) in modo che la norma $\|\cdot\|_2$ del vettore degli errori sia minima

$$\|\mathbf{e}(a,b)\|_2 \leq \|\mathbf{e}(\alpha,\beta)\|_2, \qquad \forall(\alpha,\beta)$$

Rimane ora solo da specificare come si calcolano i parametri (a, b) che minimizzano gli errori. A tale scopo, introduciamo la funzione $E(a, b)$

$$E(a,b) = \|\mathbf{e}(a,b)\|_2{}^2 = \sum_{i=1}^{n} |e_i|^2 = \sum_{i=1}^{n} |ax_i + b - y_i|^2 ,$$

a e b sono un punto di minimo della funzione $E(a, b)$ e sappiamo che in un punto di minimo il gradiente si annulla, cioè imponiamo la condizione di annullamento delle derivate parziali

$$\begin{cases} \dfrac{\partial E(a,b)}{\partial a} = 0 \\ \dfrac{\partial E(a,b)}{\partial b} = 0 \end{cases}$$

Ricordando la definizione di $E(a, b)$

$$E(a,b) = \sum_{i=1}^{n} |ax_i + b - y_i|^2 ,$$

calcoliamo esplicitamente le derivate parziali ed imponiamo l'annullamento

$$\begin{aligned} \frac{\partial E(a,b)}{\partial a} &= 2\sum_{i=1}^{n}(ax_i + b - y_i)\,x_i = 2a\sum_{i=1}^{n} x_i^2 + 2b\sum_{i=1}^{n} x_i - 2\sum_{i=1}^{n} x_i y_i = 0, \\ \frac{\partial E(a,b)}{\partial b} &= 2\sum_{i=1}^{n}(ax_i + b - y_i) \quad = 2a\sum_{i=1}^{n} x_i + 2nb - 2\sum_{i=1}^{n} y_i \qquad = 0. \end{aligned}$$

In forma matriciale compatta si ottiene il sistema

$$\left[\begin{array}{c:c} \displaystyle\sum_{i=1}^{n} x_i^2 & \displaystyle\sum_{i=1}^{n} x_i \\ \hdashline \displaystyle\sum_{i=1}^{n} x_i & n \end{array}\right] \left[\begin{array}{c} a \\ \hdashline b \end{array}\right] = \left[\begin{array}{c} \displaystyle\sum_{i=1}^{n} x_i y_i \\ \hdashline \displaystyle\sum_{i=1}^{n} y_i \end{array}\right]$$

che ha per soluzione i coefficienti a e b richiesti. Si dice che la retta $y(x) = ax + b$ passa per gli n punti assegnati *nel senso dei minimi quadrati*.

Osservazione 60. Il procedimento dei minimi quadrati di fatto risolve un sistema lineare di equazioni normali. Perché?

Imponiamo che *tutti* i punti assegnati soddisfino la generica retta $y(x) = ax + b$

$$\begin{cases} y(x_1) = ax_1 + b = y_1 \\ y(x_2) = ax_2 + b = y_2 \\ \dots \\ y(x_n) = ax_n + b = y_n \end{cases}.$$

In forma matriciale compatta

$$\begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \\ x_n & 1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix},$$

oppure introducendo la matrice $\mathbf{A} = (\mathbf{x}, \mathbf{1})$ partizionata per colonne

$$\begin{bmatrix} \mathbf{x} & \mathbf{1} \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \mathbf{y} \end{bmatrix},$$

dove $\mathbf{1}$ è la colonna di tutti 1, e le altre colonne sono ovvie.

Il sistema è sovra-determinato, quindi scriviamo le *equazioni normali*

$$\mathbf{A}^T\mathbf{A} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \mathbf{x}^T \\ \mathbf{1}^T \end{bmatrix} \begin{bmatrix} \mathbf{x} & \mathbf{1} \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \mathbf{A}^T \begin{bmatrix} \mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{x}^T \\ \mathbf{1}^T \end{bmatrix} \begin{bmatrix} \mathbf{y} \end{bmatrix},$$

da cui si ricava

$$\left[\begin{array}{c|c} \mathbf{x}^T\mathbf{x} & \mathbf{x}^T\mathbf{1} \\ \hline \mathbf{1}^T\mathbf{x} & \mathbf{1}^T\mathbf{1} \end{array}\right] \left[\begin{array}{c} a \\ \hline b \end{array}\right] = \left[\begin{array}{c} \mathbf{x}^T\mathbf{y} \\ \hline \mathbf{1}^T\mathbf{y} \end{array}\right].$$

Calcolando esplicitamente i termini $\mathbf{x}^T\mathbf{x}$, $\mathbf{x}^T\mathbf{1}$, $\mathbf{1}^T\mathbf{x}$, $\mathbf{1}^T\mathbf{1}$, $\mathbf{x}^T\mathbf{y}$ e $\mathbf{1}^T\mathbf{y}$, si riottiene il sistema dei minimi quadrati:

$$\left[\begin{array}{c|c} \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i \\ \hline \sum_{i=1}^{n} x_i & n \end{array}\right] \left[\begin{array}{c} a \\ \hline b \end{array}\right] = \left[\begin{array}{c} \sum_{i=1}^{n} x_i y_i \\ \hline \sum_{i=1}^{n} y_i \end{array}\right].$$

Conclusione: il procedimento di minimizzazione dell'errore quadratico (minimi quadrati) per le rette ed il procedimento di risoluzione delle equazioni normali sul sistema ottenuto imponendo il passaggio di una retta generica per tutti i punti assegnati coincidono.

Esempio 38. Trovare la retta che approssima nel senso dei minimi quadrati i seguenti punti: $(-4,-9)$, $(-3,-2)$, $(-1,-7)$, $(1,2)$.

$$\begin{cases} -4\alpha + \beta = -9 \\ -3\alpha + \beta = -2 \\ -\alpha + \beta = -7 \\ \alpha + \beta = +2 \end{cases}, \quad \Rightarrow \quad \begin{bmatrix} -4 & 1 \\ -3 & 1 \\ -1 & 1 \\ +1 & 1 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} -9 \\ -2 \\ -7 \\ +2 \end{bmatrix},$$

da cui segue sistema delle equazioni normali con soluzioni α e β

$$\begin{bmatrix} 27 & -7 \\ -7 & +4 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} +51 \\ -16 \end{bmatrix}, \quad \Rightarrow \quad \begin{cases} \alpha = +\frac{92}{59} \\ \\ \beta = -\frac{75}{59} \end{cases}$$

5.2.3 Generalizzazione al caso polinomiale

Il procedimento algebrico che porta alla risoluzione di un problema di approssimazione ai minimi quadrati attraverso le equazioni normali si generalizza facilmente al caso in cui le funzioni non siano rette e dipendano da un numero di parametri eventualmente superiore a due. Praticamente si procede come nel caso delle rette, imponendo il passaggio per i punti assegnati, ed ottenendo un sistema sovra-determinato nei parametri che si risolve con la tecnica delle equazioni normali. Nella presente sezione approfondiremo questo argomento discutendo il caso polinomiale.

Supponiamo di voler approssimare n punti assegnati nel piano con un polinomio di grado m

$$g(x) = a_0 + a_1 x + \cdots + a_m x^m,$$

invece che con una retta come nel caso precedente. Dato che un polinomio di grado m ha $m+1$ gradi di libertà, deve ovviamente essere $m+1 < n$, affinché il problema sia sovradeterminato [9]. Ripetendo le argomentazioni della sezione precedente, dobbiamo determinare i coefficienti a_i con

[9] Altrimenti?

$i = 0, 1, \ldots, m$ in modo tale che gli errori siano minimizzati. Introduciamo la funzione "errore quadratico", che indicheremo col simbolo $E(a_0, a_1, \ldots, a_m)$, attraverso la seguente relazione

$$E(a_0, a_1, \ldots, a_m) = \sum_{i=1}^{n} |a_0 + a_1 x_i + \cdots + a_m x^m - y_i|^2 .$$

Diremo che $g(x) = a_0 + a_1 x + \cdots + a_m x^m$ approssima "al meglio" i dati assegnati, o che il polinomio $g(x)$ passa per i punti dati nel piano *nel senso dei minimi quadrati* se i coefficienti a_i per $i = 0, 1, \ldots, m$ minimizzano la funzione errore, cioè si ha

$$E(a_0, a_1, \ldots, a_m) \leq E(\alpha_0, \alpha_1, \ldots, \alpha_m), \qquad \forall\, (\alpha_0, \alpha_1, \ldots, \alpha_m) .$$

Quindi a_0, a_1,..., a_m sono un punto di minimo della funzione errore quadratico $E(\ldots)$. Poiché in un punto di minimo il gradiente si annulla, imponiamo la condizione di annullamento delle derivate parziali

$$\begin{cases} \dfrac{\partial E(a_0, a_1, \ldots, a_m)}{\partial a_0} = 0 \\ \dfrac{\partial E(a_0, a_1, \ldots, a_m)}{\partial a_1} = 0 \\ \qquad\qquad \vdots \\ \dfrac{\partial E(a_0, a_1, \ldots, a_m)}{\partial a_m} = 0 \end{cases} . \tag{5.7}$$

Dato che

$$\begin{aligned} \frac{\partial E(a_0, a_1, \ldots, a_m)}{\partial a_0} &= 2n a_0 + 2a_1 \sum_{i=1}^{n} x_i + \cdots + 2a_m \sum_{i=1}^{n} x_i^m - 2 \sum_{i=1}^{n} y_i \\ \frac{\partial E(a_0, a_1, \ldots, a_m)}{\partial a_1} &= 2a_0 \sum_{i=1}^{n} x_i + 2a_1 \sum_{i=1}^{n} x_i^2 + \cdots + 2a_m \sum_{i=1}^{n} x_i^{m+1} - 2 \sum_{i=1}^{n} x_i y_i \\ &\vdots \\ \frac{\partial E(a_0, a_1, \ldots, a_m)}{\partial a_m} &= 2a_0 \sum_{i=1}^{n} x_i^m + 2a_1 \sum_{i=1}^{n} x_i^{m+1} + \cdots + 2a_m \sum_{i=1}^{n} x_i^{2m} - 2 \sum_{i=1}^{n} x_i^m y_i \end{aligned}$$

si ottiene per sostituzione in (5.7) il seguente sistema lineare

$$\left[\begin{array}{c:c:c:c} n & \sum_{i=1}^{n} x_i & \cdots & \sum_{i=1}^{n} x_i^m \\ \hdashline \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 & \cdots & \sum_{i=1}^{n} x_i^{m+1} \\ \hdashline \vdots & \ddots & \ddots & \vdots \\ \hdashline \sum_{i=1}^{n} x_i^m & \sum_{i=1}^{n} x_i^{m+1} & \cdots & \sum_{i=1}^{n} x_i^{2m} \end{array}\right] \left[\begin{array}{c} a_0 \\ \hdashline a_1 \\ \hdashline \vdots \\ \hdashline a_m \end{array}\right] = \left[\begin{array}{c} \sum_{i=1}^{n} y_i \\ \hdashline \sum_{i=1}^{n} x_i\, y_i \\ \hdashline \vdots \\ \hdashline \sum_{i=1}^{n} x_i^m\, y_i \end{array}\right]. \quad (5.8)$$

Osserviamo che definendo la matrice $\mathbf{M}$ e il vettore $\mathbf{b}$ come segue

$$M_{i+1\,j+1} = \sum_{k=1}^{n} x_k^{i+j}, \qquad i, j = 0, 1, \ldots, m$$

$$b_{i+1} = \sum_{k=1}^{n} x_k^i\, y_k, \qquad i = 0, 1, \ldots, m$$

il sistema (5.8) si può scrivere come $\mathbf{Ma} = \mathbf{b}$. Se introduciamo la matrice $\mathbf{V} \in \mathbb{R}^{(m+1)\times n}$ definita come

$$\mathbf{V} = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ x_0 & x_1 & x_2 & \cdots & x_n \\ x_0^2 & x_1^2 & x_2^2 & \cdots & x_n^2 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ x_0^m & x_1^m & x_2^m & \cdots & x_n^m \end{bmatrix}$$

la matrice $\mathbf{M}$ del sistema (5.8) si scrive semplicemente come $\mathbf{V}^T\mathbf{V}$. Per la proprietà dei ranghi delle matrici

$$\mathcal{R}\left\{\mathbf{V}^T\mathbf{V}\right\} = \mathcal{R}\left\{\mathbf{V}\right\} \qquad \Rightarrow \qquad \mathcal{R}\left\{\mathbf{M}\right\} = \mathcal{R}\left\{\mathbf{V}\right\}$$

La matrice $\mathbf{M}$ è invertibile se e solo se ha rango massimo cioè $\mathcal{R}\{\mathbf{M}\} = m+1$, dato che abbiamo supposto $m+1 < n$. La matrice $\mathbf{V}$ ha rango $m+1$ se e solo se esistono almeno $m+1$ colonne linearmente indipendenti. Se assumiamo che almeno $m+1$ tra i punti assegnati abbiano ascisse distinte, le rispettive colonne formano una matrice un po' "speciale", per l'appunto la *matrice di Vandermonde*. Nella sezione dedicata all'interpolazione polinomiale è stato mostrato che la matrice di Vandermonde è non singolare – e quindi ha rango massimo – se è costruita partendo da nodi di interpolazione le cui ascisse sono distinte. Possiamo concludere che l'approssimazione ai minimi quadrati nel caso polinomiale ammette una ed una sola soluzione.

Esempio 39. Con i dati riportati nella tabella 5.3

Tabella 5.3:

x	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0	10.0
y	1.5	1.8	3.0	4.1	5.6	6.1	6.0	8.1	10.0	5.0

l'approssimazione ai minimi quadrati produce il seguente polinomio di grado 2

$$g(x) = -1.033 + 1.798x - 0.097x^2,$$

vedi figura 5.2.

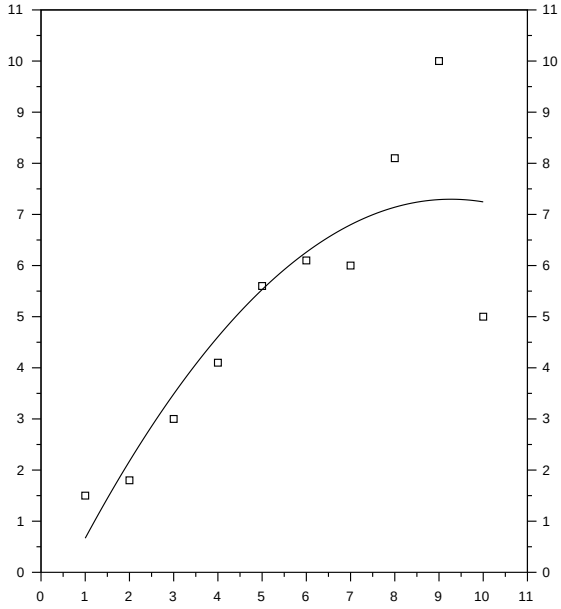


Figura 5.2: Approssimazione ai minimi quadrati dei dati nella tabella 5.3 con un polinomio di secondo grado.

CAPITOLO

SEI

INTEGRAZIONE NUMERICA

6.1 Problema dell'integrazione numerica

Data una funzione $f : \mathbb{R} \mapsto \mathbb{R}$ e l'intervallo chiuso e limitato $[a, b]$ si vuole stimare *numericamente* l'integrale

$$\mathcal{I}(f; a, b) = \int_a^b f(x)\, dx$$

Si parla di *stima numerica* perché in realtà non si sta chiedendo di calcolare il valore esatto dell'integrale che è definito come differenza dei valori assunti negli estremi di integrazione dalla primitiva della funzione integranda f. Siamo infatti interessati a determinare un valore, ovviamente approssimato, dell'integrale che coinvolga soltanto la conoscenza della funzione integranda f e non della sua primitiva.

Il problema è di particolare interessa perché, pur esistendo finito l'integrale, la funzione integranda potrebbe non ammettere una primitiva esprimibile in maniera analitica, oppure la primitiva potrebbe essere molto complessa da determinare e molto costosa da calcolare. Addirittura, la stessa funzione integranda f potrebbe non essere nota in forma analitica, ma solo tabulata in certi punti dell'intervallo $[a, b]$, rendendo quindi di fatto impossibile il cercarne una primitiva.

Riprendendo una "vecchia idea" dell'integrale come di una sorta di "somma generalizzata", cercheremo di stimare il valore dell'integrale di f per mezzo di espressioni del tipo:

$$\int_a^b f(x)\, dx \approx \sum_{k=0}^{n} w_k\, f(x_k)$$

in cui la funzione da integrare è valutata in un certo insieme di punti $x_i \in [a, b]$ e l'integrale con una somma "pesata" con coefficienti w_k opportuni.

Definizione 66. L'espressione della forma

$$\mathcal{S}(\{x_i, w_i\}_{i=0,1,\ldots,n}) = \sum_{k=0}^{n} w_k\, f(x_k)$$

definita da $\{x_i, w_i\}_{i=0,1,\ldots,n}$ si chiama *formula di quadratura*. I punti $\{x_i\}_{i=0,1,\ldots,n}$ sono i *nodi* della formula di quadratura ed i coefficienti $\{w_i\}_{i=0,1,\ldots,n}$ sono i *pesi*.

Ripetiamo ancora che la formula di quadratura fornisce giusto una stima e non il valore dell'integrale, il che implica un errore di approssimazione, che potrà in certe situazioni essere significativo. Ci porremo dunque la questione di quanto una data formula di quadratura sia precisa nella stima

dell'integrale di una funzione generica e da quali fattori sia influenzata l'accuratezza del risultato numerico.

Per definire una formula di quadratura è necessario

- scegliere i nodi $\{x_i\}_{i=0,1,\ldots,n}$ della formula di quadratura;
- scegliere i pesi $\{w_i\}_{i=0,1,\ldots,n}$ della formula di quadratura.

Quindi è ovvio che il criterio con cui si scelgono nodi e pesi è fondamentale nel caratterizzare l'accuratezza di una formula di quadratura.

Inoltre, è ragionevole aspettarsi che la precisione del valore stimato dipenda anche dalla funzione che stiamo cercando di integrare. In altre parole, la stessa formula di quadratura potrebbe produrre un risultato molto accurato – o addirittura il valore esatto – dell'integrale per alcune famiglie di funzioni e fornire un cattivo risultato in altri casi.

Questi argomenti saranno oggetto di discussione nel presente capitolo.

6.2 Strategia "interpolatoria"

(i) approssimiamo la funzione f con un polinomio interpolatore di grado n scegliendo con un *criterio da stabilire* $n+1$ nodi di interpolazione con ascisse distinte $x_i \in [a, b]$ per $i = 0, 1, \ldots, n$;

(ii) stimiamo l'integrale di f calcolando l'integrale del polinomio interpolante:

$$\int_a^b f(x)\,dx \approx \int_a^b p_n(x)\,dx.$$

Definizione 67. Le formule di quadratura costruite con la strategia interpolatoria prendono il nome di *formule di quadratura interpolatorie*

6.2.1 Classificazione (largamente incompleta)

- *Formule di Newton-Cotes* : i nodi di integrazione sono scelti *equidistanti* nell'intervallo $[a, b]$; distinguiamo tra

  - *formule aperte* : gli estremi a e b non sono compresi;
  - *formule chiuse* : $x_0 = a$ e $x_n = b$;

- *Formule Gaussiane* : i nodi di integrazione sono eventualmente gli estremi dellintervallo di integrazione piú *gli zeri* di polinomi appartenenti a famiglie speciali. Citiamo, per esempio, i *polinomi ortogonali* che producono le formule di *Gauss-Legendre*, di *Gauss-Lobatto*, di *Gauss-Radau*, ...

Nella nostra esposizione tratteremo solo il caso delle formule di Newton-Cotes.

6.2.2 Formule di Newton-Cotes

I pesi si determinano immediatamente calcolando l’integrale del polinomio interpolatore. Esaminiamo il caso delle formule “chiuse”:

$$\begin{aligned} h &= \frac{b-a}{n}, && && \text{passo di integrazione,} \\ x_i &= a + i\,h, && i = 0, 1, \ldots, n, && \text{nodo } i, \\ x &= a + t\,h, && t \in [0,\, n], && \text{generico punto } x \in [a,\, b]. \end{aligned}$$

Sia $p_n(x)$ il polinomio interpolatore di Lagrange

$$p_n(x) = f(x_0)L_0(x) + f(x_1)L_1(x) + \ldots + f(x_n)L_n(x),$$

dove gli $L_k(x)$ sono i polinomi elementari di Lagrange, definiti dalla relazione $L_i(x_j) = \delta_{ij}$. Ricordiamo la definizione data nel capitolo sull’interpolazione,

$$L_i(x) = \frac{(x-x_0)\ldots(x-x_{i-1})(x-x_{i+1})\ldots(x-x_n)}{(x_i-x_0)\ldots(x_i-x_{i-1})(x-x_{i+1})\ldots(x_i-x_n)} = \prod_{j=0}^{n}{}^{(i)} \left(\frac{x-x_j}{x_i-x_j}\right).$$

Dato che possiamo scrivere per ogni indice i e j

$$\begin{aligned} x - x_j &= (a + t\,h) - (a + j\,h) = (t-j)\,h \\ x_i - x_j &= (a + i\,h) - (a + j\,h) = (i-j)\,h, \end{aligned}$$

si ha dalla definizione dei polinomi di Lagrange un'espressione particolare per nodi di interpolazione equidistanti

$$\begin{aligned}
\int_a^b f(x)\,dx &\approx \int_a^b p_n(x)\,dx,\\
&= \int_a^b \sum_{j=0}^{n} f(x_j) L_j(x)\,dx,\\
&= \sum_{j=0}^{n} f(x_j) \int_a^b L_j(x)\,dx,\\
&= \sum_{j=0}^{n} f(x_j) \int_a^b \prod_{j=0}^{n}{}^{(i)} \left(\frac{x-x_j}{x_i-x_j}\right) dx,\\
&= \sum_{j=0}^{n} f(x_j)\, h \int_0^n \prod_{j=0}^{n}{}^{(i)} \left(\frac{t-j}{i-j}\right) dt.
\end{aligned}$$

Si ottiene quindi la formula di quadratura

$$\begin{aligned}
\mathcal{S}(\{x_i, w_i\}_{i=0,1,\ldots,n}) &= \sum_{k=0}^{n} w_k\, f(x_k)\\
&= \sum_{j=0}^{n} f(x_j)\, h \int_0^n \prod_{j=0}^{n}{}^{(i)} \left(\frac{t-j}{i-j}\right) dt.
\end{aligned}$$

definita dai nodi e dai pesi

$$\begin{aligned}
x_i &= a + i\,h, \qquad i = 0, 1, \ldots n\\
w_i &= h \int_0^n \prod_{j=0}^{n}{}^{(i)} \left(\frac{t-j}{i-j}\right) dt.
\end{aligned}$$

6.2.3 Accuratezza

Definizione 68. Si chiama *errore di integrazione* la differenza tra il valore esatto dell'integrale ed il suo valore stimato con la formula di quadratura

$$\begin{aligned}
\mathcal{E}(f; a, b) &= \left| \int_a^b f(x)\,dx - \sum_{k=0}^{n} w_k\, f(x_k) \right|\\
&= \left| \mathcal{I}(f; a, b) - \mathcal{S}(\{x_i, w_i\}_{i=0,1,\ldots,n}) \right|.
\end{aligned}$$

Ovviamente $\mathcal{E}(f; a, b)$ può dipendere dalla funzione $f(x)$ e dall'intervallo di integrazione $[a,\, b]$.

Definizione 69. *L'ordine di accuratezza* o di *precisione* della formula di quadratura è dato dal massimo grado dei polinomi che sono integrati esattamente dal metodo. Piú correttamente diremo che una formula di quadratura

- ha ordine **almeno** k se integra *esattamente* **tutti** i polinomi di grado fino a k;
- ha ordine k se ha ordine almeno k ed **esiste almeno un** polinomio di grado $k+1$ che non è integrato *esattamente*.

6.2.4 Metodo dei Coefficienti Indeterminati

Il metodo dei Coefficienti Indeterminati permette di ricondurre il calcolo dei pesi delle formule di quadratura, una volta stabiliti i nodi, alla risoluzione di un problema algebrico. Osserviamo che l'integrale di un polinomio $p_n(x)$ della forma

$$p_n(x) = a_n x^n + a_{n-1} x^{n-1} + \ldots a_1 x + a_0$$

si scrive come

$$\begin{aligned}\int_a^b p_n(x)\,dx &= \int_a^b \left(a_n x^n + a_{n-1} x^{n-1} + \ldots + a_1 x + a_0\right)\,dx \\ &= a_n \int_a^b x^n\,dx + a_{n-1}\int_a^b x^{n-1}\,dx \ldots \qquad a_1 \int_a^b x\,dx + a_0 \int_a^b 1\,dx\end{aligned}$$

Quindi l'integrale è esatto *indipendentemente* dai coefficienti $a_0, a_1, \ldots, a_n$ se sono esatti *tutti* gli integrali della forma

$$\int_a^b x^i\,dx, \qquad \text{per } i = 0, 1, \ldots n.$$

L'osservazione precedente suggerisce il seguente procedimento per la determinazione dei *pesi* di una formula di quadratura, una volta scelti i nodi (non necessariamente equidistanti).

- *Si applica la formula di quadratura a tutti i monomi x^i, per $i = 0, 1, \ldots, n$ e si impone che il risultato sia uguale al risultato dell'integrazione esatta.*

In questo modo si ricava un sistema nelle $n+1$ equazioni per $i = 0, 1, \ldots, n$ nelle $n+1$ incognite $w_0, w_1, \ldots w_n$.

Praticamente, si deve imporre la condizione

$$\sum_{j=0}^{n} w_j x_j^i = \int_a^b x^i \, dx = \frac{b^{i+1} - a^{i+1}}{i+1}, \qquad \text{per } i = 0, 1, \ldots, n,$$

per cui si ottiene un sistema lineare nelle w_i

$$\begin{aligned}
&i = 0 && w_0 + w_1 + \ldots + w_n && = b - a \\
&i = 1 && w_0 x_0 + w_1 x_1 + \ldots + w_n x_n && = \frac{b^2 - a^2}{2} \\
&i = 2 && w_0 x_0^2 + w_1 x_1^2 + \ldots + w_n x_n^2 && = \frac{b^3 - a^3}{3} \\
&\ldots \\
&i = n && w_0 x_0^n + w_1 x_1^n + \ldots + w_n x_n^n && = \frac{b^{n+1} - a^{n+1}}{n+1}.
\end{aligned}$$

Il sistema lineare si può riscrivere in forma matriciale compatta $\mathbf{Vw} = \mathbf{q}$ dove

$$\mathbf{V} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_0 & x_1 & \cdots & x_n \\ x_0^2 & x_1^2 & \cdots & x_n^2 \\ \vdots & & & \vdots \\ x_0^n & x_1^n & \cdots & x_n^n \end{bmatrix}, \qquad \mathbf{q} = \begin{bmatrix} b - a \\ \frac{b^2 - a^2}{2} \\ \frac{b^3 - a^3}{3} \\ \vdots \\ \frac{b^{n+1} - a^{n+1}}{n+1} \end{bmatrix},$$

nelle incognite $w_0, w_1, \ldots, w_n$.

La matrice $\mathbf{V}$ è una matrice di *Vandermonde*, il cui determinante

$$V(x_0, x_1, \ldots, x_n) = \prod_{i>j}^{n} (x_i - x_j)$$

è sicuramente diverso da zero se i nodi sono scelti a due a due distinti, cioè vale $x_i \neq x_j$ per $i \neq j$. La non singolarità della matrice $\mathbf{V}$ garantisce l'esistenza di una ed una sola soluzione formata da una $n+1$-upla di pesi $w_0, w_1, \ldots, w_n$.

Osservazione 61. La formula di quadratura così costruita ha ordine *almeno* n. Qual'è il massimo ordine che può avere?

Se lasciamo libertà di scegliere sia gli $n+1$ nodi (purché distinti) che gli $n+1$ pesi, abbiamo in totale $2(n+1)$ *gradi di libertà*, per cui possiamo chiedere che il metodo dei coefficienti indeterminati soddisfi un ugual numero di condizioni, il che ci da la possibilità di integrare esattamente tutti i monomi fino all'ordine $2n+1$ [1].

6.2.5 Stima dell'errore di integrazione

In questo paragrafo riportiamo per completezza l'enunciato di un teorema assai importante, che permette di stimare l'ordine di accuratezza per le formule di Newton[2] -Cotes[3].

Il teorema si riassume usualmente con l'affermazione che *le formule di Newton-Cotes di grado pari guadagnano un ordine di accuratezza*. Per esempio, la formula di Simpson, pur essendo costruibile come un formula interpolatoria che integra esattamente tutti i polinomi di grado *fino a due*, ha in realtà ordine *quattro*, e non *tre* come si potrebbe erroneamente pensare.

Questa proprietà dipende essenzialmente dal fatto che l'integrale di $\omega_{n+1}(x)$, che compare nell'espressione dell'errore di fronte al termine $n+1$ dello sviluppo di Taylor, si annulla sull'intervallo $[a,b]$ per ragioni di simmetria. L'errore è dominato dal termine successivo dello sviluppo di Taylor che coinvolge la derivata $n+2$-esima.

L'enunciato del teorema mostra la forma generale dell'errore di integrazione per formule interpolatorie alla Newton-Cotes che interpolano esattamente polinomi di ordine fino ad n sia pari che dispari. Ricordiamo tuttavia che nella pratica non si usano formule di Newton-Cotes con n grandi, maggiori di 7-8, per ragioni di stabilità numerica.

La dimostrazione si può trovare per esempio in [1].

Teorema 81. *Sia* $\mathcal{E}(f;a,b) = |\mathcal{I}(f;a,b) - \mathcal{S}(\{x_i, w_i\}_{i=0,1,\ldots,n})|$ *l'errore di integrazione che si ha applicando una formula di quadratura di Newton-Cotes con* $n+1$ *nodi*

$$\mathcal{S}(\{x_i, w_i\}_{i=0,1,\ldots,n}) = \sum_{k=0}^{n} w_k\, f(x_k)$$

per il calcolo dell'integrale della funzione $f(x)$ *su* $[a\,b]$

$$\mathcal{I}(f;a,b) = \int_a^b f(x)\,dx$$

[1] Attenzione: la numerazione degli esponenti parte da zero, essendo $x^0 = 1$ il primo monomio di ogni polinomio di qualsiasi ordine.

[2] Sir Isaac Newton 1643–1727

[3] Roger Cotes 1682–1716

Allora valgono i seguenti risultati:

(i) *sia* $f \in \mathcal{C}^{n+2}(a,b)$*, con* n ***pari****; allora esiste un* $\bar{x} \in (a,b)$ *tale che*

$$\mathcal{E}(f;a,b) = \frac{1}{(n+2)!} f^{n+2}(\bar{x}) \int_a^b x\,\omega_{n+1}(x)\,dx$$

(ii) *sia* $f \in \mathcal{C}^{n+1}(a,b)$*, con* n ***dispari****; allora esiste un* $\bar{x} \in (a,b)$ *tale che*

$$\mathcal{E}(f;a,b) = \frac{1}{(n+1)!} f^{n+1}(\bar{x}) \int_a^b \omega_{n+1}(x)\,dx$$

dove $\omega_{n+1}(x) = (x-x_0)(x-x_1)\cdots(x-x_n)$.

BIBLIOGRAFIA

[1] Roberto Bevilacqua, Dario Bini, Milvio Capovani, and Ornella Menchi. *Metodi Numerici*. Zanichelli, 1992.

INDICE ANALITICO