Calcolo Numerico e Programmazione - A.A. 2003/04

Corso di Laurea triennale in Scienza dei Materiali
Università di Milano Bicocca

C. Tablino Possio
Dipartimento di Matematica e Applicazioni
Università di Milano Bicocca

Indice

Parte I
Algebra lineare teorica

1 Gli spazi vettoriali

1.1 Definizione ed esempi

Consideriamo come esempio di riferimento lo spazio $\mathbb{R}^n, n \geq 1$, ossia l'insieme delle $n-$ple di numeri reali con n fissato

$$\mathbb{R}^n = \{\underline{x} = (x_1, x_2, \ldots, x_n) \mid x_i \in \mathbb{R}^n \text{ per ogni } i = 1, 2, \ldots, n\}$$

Ogni $n-$pla di numeri reali viene detta vettore di $\mathbb{R}^n$.

Ora si vogliono introdurre due operazioni in $\mathbb{R}^n$.

Definizione 1.1 + ***Somma di due vettori***
Siano $\underline{x} = (x_1, x_2, \ldots, x_n), \underline{y} = (y_1, y_2, \ldots, y_n) \in \mathbb{R}^n$ *allora si definisce il vettore somma come*

$$\underline{x} + \underline{y} = (x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n).$$

Definizione 1.2 $\cdot$ ***Prodotto di un vettore per uno scalare***
Sia $\alpha \in \mathbb{R}$ *e* $\underline{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ *allora si definisce il vettore prodotto come*

$$\alpha \cdot \underline{x} = (\alpha x_1, \alpha x_2, \ldots, \alpha x_n).$$

Proprietà 1.1 *Chiamiamo per semplicità di notazione* V *lo spazio* $\mathbb{R}^n$ *e* K *lo spazio* $\mathbb{R}$. *L'operazione di somma di due vettori di Definizione 1.1 soddisfa le seguenti proprietà:*

1. *per ogni* $\underline{x}, \underline{y} \in V$ *si ha che* $\underline{x} + \underline{y} \in V$ *(Chiusura);*
2. *per ogni* $\underline{x}, \underline{y}, \underline{z} \in V$ *si ha che* $(\underline{x} + \underline{y}) + \underline{z} = \underline{x} + (\underline{y} + \underline{z})$ *(Associativa);*
3. *per ogni* $\underline{x}, \underline{y} \in V$ *si ha che* $\underline{x} + \underline{y} = \underline{y} + \underline{x}$ *(Commutativa);*
4. *esiste* $\underline{0} \in V$ *tale che* $\underline{x} + \underline{0} = \underline{0} + \underline{x} = \underline{x}$ *per ogni* $\underline{x} \in V$ *(Elemento neutro);*
5. *per ogni* $\underline{x} \in V$ *esiste unico* $-\underline{x} \in V$ *tale che* $\underline{x} + (-\underline{x}) = (-\underline{x}) + \underline{x} = \underline{0}$ *(Elemento opposto).*

L'operazione di prodotto di un vettore per uno scalare di Definizione 1.2 soddisfa le seguenti proprietà:

1. *per ogni* $\underline{x} \in V$ *e per ogni* $\alpha \in K$ *si ha che* $\alpha \cdot \underline{x} \in V$ *(Chiusura);*
2. *per ogni* $\underline{x}, \underline{y} \in V$ *e per ogni* $\alpha \in K$ *si ha che* $\alpha \cdot (\underline{x} + \underline{y}) = \alpha \cdot \underline{x} + \alpha \cdot \underline{y}$ *(Distributiva);*
3. *per ogni* $\underline{x} \in V$ *e per ogni* $\alpha, \beta \in K$ *si ha che* $(\alpha + \beta) \cdot \underline{x} = \alpha \cdot \underline{x} + \beta \cdot \underline{x}$ *(Distributiva);*
4. *per ogni* $\underline{x} \in V$ *e per ogni* $\alpha, \beta \in K$ *si ha che* $\alpha \cdot (\beta \cdot \underline{x}) = (\alpha\beta) \cdot \underline{x}$*;*
5. *esiste unico* $1 \in K$ *tale che* $1 \cdot \underline{x} = \underline{x}$ *per ogni* $\underline{x} \in V$ *(Elemento neutro).*

Dimostrazione: Si tratta di una semplice verifica facendo uso delle proprietà della somma e del prodotto di numeri reali. Più precisamente, essendo le operazioni definite delle operazioni componente per componente, è evidente che le proprietà della somma e del prodotto di numeri reali "salgono" direttamente alla struttura più complessa del vettore. Per semplicità si pensi al caso $n = 2$. Da notare che è evidentemente

$$\begin{aligned} \underline{0} &= (0, 0, \ldots, 0) \quad \text{Elemento neutro} \\ -\underline{x} &= (-x_1, -x_2, \ldots, -x_n) \quad \text{Elemento opposto} \end{aligned}$$

Ora, un ulteriore passo di astrazione permette di introdurre la nozione di spazio vettoriale. Le applicazioni di tale concetto si trovano nei più svariati campi delle scienze pure ed applicate.

Definizione 1.3 ***Spazio vettoriale***
Un insieme V in cui sia definita un operazione $+$ di somma tra elementi dell'insieme e un operazione di prodotto $\cdot$ di un elemento dell'insieme per uno scalare tale che valgano tutte le Proprietà 1.1 si dice spazio vettoriale (rispetto a $+$ e $\cdot$ fissati).

Esempio 1.1
Sia

$$\begin{aligned} V = \mathbb{P}_n\{p_n(x) &= a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n \\ &\quad \text{tale che } \ a_i \in \mathbb{R} \text{ per ogni } \ i = 0, \ldots, n\} \end{aligned}$$

insieme di tutti i polinomi di grado non superiore a n (con n fissato). Si considera l'operazione $+$ di somma di due polinomi

$$\begin{aligned} p_n(x) + q_n(x) &= (a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n) \\ &\quad +(b_0 + b_1 x + b_2 x^2 + \ldots + b_n x^n) \\ &= (a_0 + b_0) + (a_1 + b_1)x + (a_2 + b_2)x^2 + \ldots + (a_n + b_n)x^n \end{aligned}$$

e l'operazione $\cdot$ di prodotto per uno scalare $\alpha \in \mathbb{R}$

$$\begin{aligned} \alpha p_n(x) &= \alpha(a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n) \\ &= \alpha a_0 + \alpha a_1 x + \alpha a_2 x^2 + \ldots + \alpha a_n x^n. \end{aligned}$$

Si può verificare facilmente che V è spazio vettoriale (rispetto a $+$ e $\cdot$ fissati).

1.2 Basi di uno spazio vettoriale

Per cogliere una più evidente analogia tra lo spazio vettoriale $\mathbb{R}^n$ e lo spazio vettoriale $\mathbb{P}_n$ occorre introdurre la fondamentale nozione di base di uno spazio vettoriale. Ci occorrono alcune definizioni preliminari.

Definizione 1.4 ***Combinazione lineare di vettori***
Sia V uno spazio vettoriale. Si dice che $\underline{y} \in V$ è ***combinazione lineare*** *dei vettori $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_m \in V$ se esistono m scalari $\alpha_1, \alpha_2, \ldots, \alpha_m$ tali che*

$$\underline{y} = \alpha_1 \underline{x}_1 + \alpha_2 \underline{x}_2 + \ldots + \alpha_m \underline{x}_m.$$

Con stretto riferimento alla definizione precedente, si pone pure la seguente.

Definizione 1.5 ***Lineare dipendenza/indipendenza***
Sia V uno spazio vettoriale. Si dice che i vettori $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_m \in V$ sono linearmente dipendenti se esistono m scalari $\alpha_1, \alpha_2, \ldots, \alpha_m$ non tutti nulli tali che

$$\alpha_1 \underline{x}_1 + \alpha_2 \underline{x}_2 + \ldots + \alpha_m \underline{x}_m = \underline{0} \ (\textit{vettore nullo}).$$

In caso contrario i vettori $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_m \in V$ si dicono linearmente indipendenti, ossia il vettore nullo $\underline{0}$ è combinazione lineare dei vettori $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_m$ se e soltanto se

$$\alpha_1 = \alpha_2 = \ldots = \alpha_m = 0.$$

Esempio 1.2
Consideriamo $V = \mathbb{R}^2$ e i due vettori $\underline{x}_1 = [1, 0]$ e $\underline{x}_2 = [1, 1]$; valgono le seguenti affermazioni:

1. *i due vettori $\underline{x}_1$ e $\underline{x}_2$ sono linearmente indipendenti;*
2. *il vettore $\underline{x}_3 = [3, 2]$ è combinazione lineare dei due vettori $\underline{x}_1$ e $\underline{x}_2$;*
3. *I tre vettori $\underline{x}_1$, $\underline{x}_2$, $\underline{x}_3$ sono linearmente dipendenti.*

Infatti

1. *$\alpha \underline{x}_1 + \beta \underline{x}_2 = [\alpha + \beta, \beta]\underline{0}$ se e solo se vale contemporaneamente $\alpha + \beta = 0$ e $\beta = 0$, ossia $\alpha = \beta = 0$;*
2. *si verifica che $\underline{x}_3 = [3, 2] = \alpha \underline{x}_1 + \beta \underline{x}_2$ con $\alpha = 1, \beta = 2$;*
3. *si verifica che $\alpha \underline{x}_1 + \beta \underline{x}_2 + \gamma \underline{x}_3 = \underline{0}$ con ad esempio $\alpha = 1, \beta = 2, \gamma = -1$.*

Definizione 1.6 ***Base di uno spazio vettoriale***
Un insieme di vettori $\{\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_m \in V\}$ si dice ***base*** *dello spazio vettoriale V se*

a) i vettori $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_m$ sono linearmente indipendenti,

b) se ogni vettore di V si può ottenere come combinazione lineare dei vettori $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_m$, ossia, con notazione compatta,

$$V = span < \underline{x}_1, \underline{x}_2, \ldots, \underline{x}_m > .$$

Teorema 1.1 *La combinazione lineare con cui si rappresenta un fissato vettore di V a partire dai vettori della base è unica.*

Dimostrazione:
Sia $V = span < \underline{x}_1, \underline{x}_2, \ldots, \underline{x}_m >$ e $\underline{y} \in V$. Supponiamo che il vettore $\underline{y}$ si possa scrivere come combinazione lineare dei vettori della base $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_m$ in due diversi modi, ossia

$$\begin{aligned} \underline{y} &= \alpha_1 \underline{x}_1 + \alpha_2 \underline{x}_2 + \ldots + \alpha_m \underline{x}_m, \\ \underline{y} &= \beta_1 \underline{x}_1 + \beta_2 \underline{x}_2 + \ldots + \beta_m \underline{x}_m. \end{aligned}$$

Sottraendo membro a membro si ha

$$\underline{0} = \underline{y} - \underline{y} = (\alpha_1 - \beta_1)\underline{x}_1 + (\alpha_2 - \beta_2)\underline{x}_2 + \ldots + (\alpha_m - \beta_m)\underline{x}_m,$$

ma poiché i vettori $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_m$ sono linearmente indipendenti fra di loro si ha che deve essere

$$\begin{array}{rcl} \alpha_1 - \beta_1 = 0 & & \alpha_1 = \beta_1 \\ \alpha_2 - \beta_2 = 0 & \text{ossia} & \alpha_2 = \beta_2 \\ \vdots & & \vdots \\ \alpha_m - \beta_m = 0 & & \alpha_m = \beta_m, \end{array}$$

ossia la combinazione lineare è la stessa.

Pertanto, si può osservare come l'importanza della nozione di base di uno spazio vettoriale consista nel fatto che costituisce una sorta di informazione "sintetica", e non ambigua, tramite cui si è certi di poter rappresentare uno qualsivoglia degli infiniti elementi dello spazio vettoriale in esame.

Esempio 1.3 *Sia* $V = \mathbb{R}^n$.

1. *Si dice* ***base canonica*** *di* $\mathbb{R}^n$ *la base costituita dagli* n *vettori*

$$\begin{array}{rcl} \underline{e}_1 & = & (1, 0, \ldots, 0) \text{ (prima componente pari a 1, le rimanenti a 0)} \\ \underline{e}_2 & = & (0, 1, \ldots, 0) \text{ (seconda componente pari a 1, le rimanenti a 0)} \\ \vdots & & \vdots \\ \underline{e}_n & = & (0, 0, \ldots, 1) \text{ (ultima componente pari a 1, le rimanenti a 0).} \end{array}$$

Infatti, ogni vettore $\underline{x} = (x_1, x_2, \ldots, x_n)$ *si può scrivere come combinazione lineare dei vettori* $\underline{e}_1, \underline{e}_2, \ldots, \underline{e}_n$ *nel modo seguente*

$$\underline{x} = x_1\underline{e}_1 + x_2\underline{e}_2 + \ldots + x_n\underline{e}_n$$

e i vettori $\underline{e}_1, \underline{e}_2, \ldots, \underline{e}_n$ *sono linearmente indipendenti in quanto* $\underline{x} = x_1\underline{e}_1 + x_2\underline{e}_2 + \ldots + x_n\underline{e}_n = \underline{0}$ *se e solo se*

$$x_1 = x_2 = \ldots = x_n = 0.$$

2. *Si dice* ***base a bandiera*** *di* $\mathbb{R}^n$ *la base costituita dagli* n *vettori*

$$\begin{array}{rcl} \underline{v}_1 & = & (1, 0, 0, \ldots, 0) \text{ (prima componente pari a 1, le rimanenti a 0)} \\ \underline{v}_2 & = & (1, 1, 0, \ldots, 0) \text{ (prime due componenti pari a 1, le rimanenti a 0)} \\ \underline{v}_3 & = & (1, 1, 1, \ldots, 0) \text{ (prime tre componenti pari a 1, le rimanenti a 0)} \\ \vdots & & \vdots \\ \underline{v}_n & = & (1, 1, 1, \ldots, 1) \text{ (tutte le componenti pari a 1).} \end{array}$$

Infatti, fissato ad esempio $n = 4$, *ogni vettore* $\underline{x} = (x_1, x_2, x_3, x_4)$ *si può scrivere come combinazione lineare dei vettori* $\underline{v}_1, \underline{v}_2, \underline{v}_3, \underline{v}_4$ *nel modo seguente* $\underline{x} = (x_1 - x_2)\underline{v}_1 + (x_2 - x_3)\underline{v}_2 + (x_3 - x_4)\underline{v}_3 + x_4\underline{v}_4$ *e i vettori* $\underline{v}_1, \underline{v}_2, \underline{v}_3, \underline{v}_4$, *sono linearmente indipendenti in quanto* $\underline{x} = (x_1 - x_2)\underline{v}_1 + (x_2 - x_3)\underline{v}_2 + (x_3 - x_4)\underline{v}_3 + x_4\underline{v}_4 = \underline{0}$ *se e solo se*

$$\begin{array}{rcl} x_1 - x_2 = 0 & & x_1 = x_2 \\ x_2 - x_3 = 0 & \textit{ossia} & x_2 = x_3 \\ x_3 - x_4 = 0 & & x_3 = x_4 \\ x_4 = 0 & & \end{array}$$

e quindi $x_1 = x_2 = x_3 = x_4 = 0$.
Per n *generico si ha che ogni vettore* $\underline{x} = (x_1, x_2, \ldots, x_n)$ *si può scrivere come combinazione lineare dei vettori* $\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_n$ *nel modo seguente* $\underline{x} = (x_1 - x_2)\underline{v}_1 + (x_2 - x_3)\underline{v}_2 + \ldots + (x_{n-1} - x_n)\underline{v}_{n-1} + x_n\underline{v}_n$ *e i vettori* $\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_n$ *sono linearmente indipendenti in quanto* $\underline{x} = (x_1 - x_2)\underline{v}_1 + (x_2 - x_3)\underline{v}_2 + \ldots + (x_{n-1} - x_n)\underline{v}_{n-1} + x_n\underline{v}_n = \underline{0}$ *se e solo se*

$$\begin{array}{rcl} x_1 - x_2 = 0 & & x_1 = x_2 \\ x_2 - x_3 = 0 & \textit{ossia} & x_2 = x_3 \\ \vdots & & \vdots \\ x_{n-1} - x_n = 0 & & x_{n-1} = x_n \\ x_n = 0 & & \end{array}$$

e quindi $x_1 = x_2 = \ldots = x_{n-1} = x_n = 0$.

Ora, il fatto che nell'Esempio 1.3 siano riportati due esempi distinti di base di $\mathbb{R}^n$ è di estrema importanza in quanto mette in luce che la base di uno spazio vettoriale non è unica.
Tuttavia, si osserva che entrambe le basi sono costituite dallo stesso numero n di vettori. Tale fatto non è una pura coincidenza; vale infatti il seguente teorema.

Teorema 1.2 ***della base***
Tutte le basi di uno spazio vettoriale V *sono costituite dallo stesso numero di vettori, ossia hanno la medesima cardinalità.*

Il precedente Teorema 1.2 giustifica la seguente definizione.

Definizione 1.7 ***Dimensione dello spazio vettoriale***
Si dice ***dimensione*** *dello spazio vettoriale* V *il numero intero che indica la cardinalità di una qualsiasi base di* V.

La possibilità di considerare basi diverse per un medesimo vettore di un assegnato spazio vettoriale non è un puro gioco matematico. A seconda delle applicazioni può essere più conveniente considerare una base piuttosto che un'altra. Ritorneremo in seguito su questo argomento; più precisamente nella sezione 4.

Esempio 1.4 ***Spazi vettoriali a dimensione finita***

1. *Sia* $V = \mathbb{R}^n$.
   Si ha $\dim V = n$. *Esempi di base sono riportati in Esempio 1.3.*

2. *Sia* $V = \mathbb{P}_n = \{p_n(x) = a_0 + a_1x + a_2x^2 + \ldots + a_nx^n$ *tale che* $a_i \in \mathbb{R}$ *per ogni* $i = 0, \ldots, n\}$, *insieme polinomi di grado non superiore* n.
Si ha $\dim V = n + 1$. *Base canonica: gli* $n + 1$ *monomi* $1, x, x^2, \ldots, x^n$.

Esempio 1.5 ***Spazi vettoriali a dimensione infinita***
Sia $V = \mathbb{P}$ *insieme polinomi di qualsiasi grado* $n \in \mathbb{N}$.
Si ha $\dim V = \infty$. *Base canonica: i monomi* $1, x, x^2, \ldots, x^n, \ldots$.

Tramite il concetto di base, e più precisamente quello di rappresentazione di ogni elemento dello spazio vettoriale come combinazione lineare di vettori della base, diventa chiaro come l'insieme $\mathbb{P}_n$ costituito da elementi apparentemente così diversi da quelli di $\mathbb{R}^m$, non sia poi così differente: il polinomio

$$p_n(x) = a_0 + a_1x + a_2x^2 + \ldots + a_nx^n$$

è identificato in modo unico rispetto alla base $1, x, x^2, \ldots, x^n$ dagli $n + 1$ coefficienti $a_0, a_1, \ldots, a_n \in \mathbb{R}$, ossia da una $m-$pla di $\mathbb{R}^m$ con $m = n + 1$

$$(a_0, a_1, a_2, \ldots, a_n)$$

detta **vettore delle coordinate rispetto alla base considerata**.

Semplice conseguenza del Teorema della base 1.2 è il seguente corollario.

Corollario 1.1

1. *Sia* V *uno spazio vettoriale di dimensione* N, *allora ogni insieme di* n *vettori linearmente indipendenti è una base per* V;
2. *Sia* V *uno spazio vettoriale di dimensione* N, *allora ogni insieme di* n *vettori che genera tutto* V *è una base per* V;

1.3 Sottospazi vettoriali

È di notevole interesse la seguente definizione.

Definizione 1.8 ***Sottospazio vettoriale*** *Un sottoinsieme* S *di uno spazio vettoriale* V *si dice sottospazio vettoriale di* V *se*

1. *per ogni* $\underline{x}, \underline{y} \in S$ *vale che* $\underline{x} + \underline{y} \in S$;
2. *per ogni* $\underline{x} \in S$ *e per ogni* $\alpha \in K$ *vale che* $\alpha\underline{x} \in S$;

ossia se il sottoinsieme S *è chiuso rispetto alle operazioni di somma e di moltiplicazione per uno scalare.*

Esempio 1.6 *Sia* $V = \mathbb{R}^3$. *Il sottoinsieme*

$$S = \{(x_1, x_2, 0) \text{ tali che } x_1, x_2 \in \mathbb{R}\}$$

è un sottospazio vettoriale di V.
Di più vale $S = span < \underline{y}_1, \underline{y}_2 >$, *con ad esempio* $\underline{y}_1 = [1, 0], \underline{y}_2 = [0, 1]$, *oppure* $\underline{y}_1 = [1, 0], \underline{y}_2 = [1, 1]$.

Per altri esempi di sottospazi si veda la sezione 2.9.

1.4 Esercizi

1. Verificare che $V = \mathbb{R}^3$ ($K = \mathbb{R}$) è spazio vettoriale rispetto alle operazioni di somma e prodotto per uno scalare definite in Definizione 1.1 e in 1.2.

2. Verificare che $V = \mathbb{P}_2$ ($K = \mathbb{R}$) è spazio vettoriale rispetto alle operazioni di somma e prodotto per uno scalare definite in Esempio 1.1.

3. Verificare che $\underline{x} = (1, 2, 3)$ e $\underline{y} = (5, 10, 15)$ sono linearmente dipendenti. Verificare che $\underline{x} = (1, 2, 3)$ e $\underline{z} = (5, -10, 15)$ sono linearmente indipendenti.

4. Determinare i valori del parametro α per cui i vettori $\underline{x} = (1, 2, -1)$ e $\underline{y} = (3, \alpha, -3)$ e i vettori $\underline{x} = (1, 2, -1)$ e $\underline{y} = (\alpha, 2\alpha, -\alpha)$sono linearmente dipendenti.

2 Generalità sulle matrici

2.1 Definizione e casi particolari

Definizione 2.1 ***Matrice*** $n \times m$
Una matrice $n \times m$ è una tabella rettangolare di n righe e m colonne i cui elementi sono numeri reali (o complessi) indicizzati con la seguente notazione: a_{ij} indica l'elemento di posizione i rispetto alle righe e di posizione j rispetto alle colonne.

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1j} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2j} & \dots & a_{2m} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{i1} & a_{i2} & \dots & a_{ij} & \dots & a_{im} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nj} & \dots & a_{nm} \end{bmatrix} = [a_{ij}]_{i=1,\dots,n;j=1,\dots,m} \in \mathbb{R}^{n\times m}$$

Se $n = m$, la matrice si dice ***quadrata****.*

Si tenga presente che i vettori sono casi particolari di matrici aventi $n = 1$ (si dicono **vettori riga**) o $m = 1$ (si dicono **vettori colonna**).

2.2 Operazioni con le matrici

Sia $V = \{A \text{ tali che } A \in \mathbb{R}^{n\times m}\}$ l'insieme delle matrici di n righe e m colonne a elementi reali. Come precedentemente introdotto, si usa la notazione a_{ij} per indicare l'elemento di posizione i rispetto alle righe e di posizione j rispetto alle colonne.
Si definiscono le seguenti operazioni.

Definizione 2.2 $+$ ***Somma di matrici***
Siano $A, B \in \mathbb{R}^{n\times m}$ allora si definisce la matrice somma come

$$C = A + B \text{ ove } c_{ij} = a_{ij} + b_{ij}, \text{ per ogni } i = 1, \dots, n; j = 1, \dots, m$$

Definizione 2.3 $\cdot$ ***Moltiplicazione per uno scalare***
Siano $A \in \mathbb{R}^{n\times m}$ e $\alpha \in \mathbb{R}$ allora si definisce la matrice prodotto per uno scalare come

$$D = \alpha \cdot A \text{ ove } d_{ij} = \alpha a_{ij}, \text{ per ogni } i = 1, \dots, n; j = 1, \dots, m$$

Esempio 2.1 *Sia $n = m = 2$ e*

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \text{ e } B = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix},$$

allora

$$C = A + B = \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} \\ a_{21} + b_{21} & a_{22} + b_{22} \end{bmatrix}$$

e

$$D = \alpha \cdot A = \begin{bmatrix} \alpha a_{11} & \alpha a_{12} \\ \alpha a_{21} & \alpha a_{22} \end{bmatrix}.$$

Proposizione 2.1 *L'insieme* $V = \{A \text{ tali che } A \in \mathbb{R}^{n\times m}\}$ *è uno spazio vettoriale rispetto alle due operazioni di somma di matrici e moltiplicazione di una matrice per uno scalare di Definizione 2.2 e 2.3.*
La sua dimensione è nm. *Una base (base canonica) è data da*

$$\{E^{st} \in \mathbb{R}^{n\times m}\}_{s=1,\dots,n;t=1,\dots,m}$$

con

$$E^{st} \text{ tale che } (E^{st})_{ij} = \begin{cases} 1 & se \quad i = j \\ 0 & se \quad i \neq j. \end{cases}$$

Esempio 2.2 *Sia* $V = \{A \in \mathbb{R}^{2\times 3}\}$. *La sua dimensione è* 6.
La base canonica è data da $\{E^{st} \in \mathbb{R}^{2\times 3}\}_{s=1,\dots,2;t=1,\dots,3}$ *con*

$$\begin{aligned} E^{11} &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, E^{12} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, E^{13} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \\ E^{21} &= \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}, E^{22} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, E^{23} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \end{aligned}$$

Si noti che come conseguenza importante della precedente proposizione si ha che la somma di matrici gode delle stesse proprietà della somma di numeri reali.

Ora, si introduce un'altra operazione di prodotto, ossia l'operazione di prodotto fra matrici: tale operazione è possibile solo nel caso in cui il numero di colonne della matrice di sinistra sia uguale al numero di righe della matrice di destra; in caso contrario l'operazione di prodotto non è consistente e quindi non può essere effettuata.

Definizione 2.4 ***Prodotto di matrici***
Sia $A \in \mathbb{R}^{n\times m}$ *e* $B \in \mathbb{R}^{m\times l}$ *allora* $C = A \cdot B \in \mathbb{R}^{n\times l}$ *ove*

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

Graficamente

$$\begin{bmatrix} a_{11} & a_{12} & \dots & \dots & \dots & a_{1m} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \boxed{a_{i1} \quad a_{i2} \quad \dots \quad a_{ik} \quad \dots \quad a_{im}} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \dots & \dots & \dots & a_{nm} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} & \dots & \boxed{b_{1j}} & \dots & b_{1l} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \dots & b_{kj} & \dots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ b_{m1} & b_{m2} & \dots & b_{mj} & \dots & b_{ml} \end{bmatrix}$$

$$= \begin{bmatrix} c_{11} & \dots & c_{1j} & \dots & c_{1l} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ c_{i1} & \dots & \boxed{c_{ij}} & \dots & c_{il} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ c_{n1} & \dots & c_{nj} & \dots & c_{nl} \end{bmatrix}$$

Si noti che il numero di righe della matrice di sinistra definisce il numero di righe della matrice risultato; mentre il numero di colonne della matrice di destra definisce il numero di colonne della matrice risultato.

Casi particolari

Definizione 2.5 ***Prodotto matrice per vettore colonna***
Sia $A \in \mathbb{R}^{n\times m}$ e $b \in \mathbb{R}^m$ (vettore colonna a m componenti). Poiché b può essere pensato come una matrice $m \times 1$ allora è definito $c = A \cdot b \in \mathbb{R}^{n\times 1} = \mathbb{R}^n$ (vettore colonna a n componenti) ove

$$c_i = \sum_{k=1}^{m} a_{ik} b_k, \ \ i = 1, \dots, n.$$

Graficamente

$$\begin{bmatrix} a_{11} & a_{12} & \dots & \dots & \dots & a_{1m} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \boxed{a_{i1} & a_{i2} & \dots & a_{ik} & \dots & a_{im}} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \dots & \dots & \dots & a_{nm} \end{bmatrix} \boxed{\begin{bmatrix} b_1 \\ \vdots \\ b_k \\ \vdots \\ b_m \end{bmatrix}} = \begin{bmatrix} c_1 \\ \vdots \\ \boxed{c_i} \\ \vdots \\ c_n \end{bmatrix}$$

Definizione 2.6 ***Prodotto vettore riga per matrice***
Sia $c \in \mathbb{R}^{1\times m}$ (vettore riga a m componenti) e $B \in \mathbb{R}^{m\times l}$. Poiché c può essere pensato come una matrice $1 \times m$ allora è definito $d = c \cdot B \in \mathbb{R}^{1\times l}$ (vettore riga a l componenti) ove

$$d_j = \sum_{k=1}^{m} c_k b_{kj}, \ \ j = 1, \dots, l.$$

Graficamente

$$\boxed{\begin{bmatrix} c_1 & \dots & c_k & \dots & c_m \end{bmatrix}} \begin{bmatrix} b_{11} & \dots & \boxed{b_{1j}} & \dots & b_{1l} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \dots & b_{kj} & \dots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ b_{m1} & \dots & b_{mj} & \dots & b_{ml} \end{bmatrix} = \begin{bmatrix} d_1 & \dots & \boxed{d_j} & \dots & d_l \end{bmatrix}$$

2.3 Un caso notevole: le matrici quadrate ($n = m$)

Si vuole analizzare in maggior dettaglio il caso delle matrici quadrate, ovvero il caso in cui il numero di colonne è pari al numero di righe. È evidente che matrici quadrate della medesima dimensione possono essere moltiplicate liberamente fra loro ottenendo come risultato una matrice della medesima dimensione.
La successiva domanda che è naturale porsi è:
il prodotto di matrici quadrate della medesima dimensione gode delle stesse proprietà del prodotto di numeri reali?
La risposta è negativa, ma occorre precisare meglio.

Vediamo la questione in dettaglio: come già nel caso del prodotto di numeri reali valgono le seguenti proprietà.

Proposizione 2.2 *Per ogni $A, B, C \in \mathbb{R}^{n\times n}$ vale che*

1. *A(BC)=(AB)C (associativa)*
2. *A(B+C)=AB+AC (distributiva a destra)*
3. *(B+C)A = BA+CA (distributiva a sinistra)*
4. *Esiste $I \in \mathbb{R}^{n\times n}$ (matrice identica) tale che per ogni $A \in \mathbb{R}^{n\times n}$*

$$A \cdot I = I \cdot A = A$$

con

$$I \text{ tale che} \qquad I_{ij} = \begin{cases} 1 & se \quad i = j \\ 0 & se \quad i \neq j, \end{cases}$$

ossia

$$I = \begin{bmatrix} 1 & 0 & \dots & \dots & \dots & 0 \\ 0 & 1 & 0 & \dots & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & 0 & 1 & 0 \\ 0 & \dots & \dots & \dots & 0 & 1 \end{bmatrix}.$$

Ora, è già evidente che nel caso di matrici rettangolari il prodotto di matrici non è commutativo, visto che, in genere, l'operazione non sarà più consistente. Tuttavia è del tutto lecito chiedersi se lo diventi nel caso di matrici quadrate della stessa dimensione.
Consideriamo il seguente esempio.

Esempio 2.3 *Si considerino*

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

Si ha che

$$A \cdot B = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \neq B \cdot A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}.$$

Quindi, a differenza del prodotto di numeri reali, non vale la proprietà commutativa neanche nel caso di prodotto di matrici quadrate della medesima dimensione.
Infatti, nell'Esempio 2.3 si è visto che esistono $A, B \in \mathbb{R}^{n\times n}$ tali che

$$A \cdot B \neq B \cdot A \quad (\textbf{ NO commutativa})$$

Infine, l'ultima questione aperta è quella delle condizioni sotto le quali si può garantire l'esistenza di un elemento inverso per $A \in \mathbb{R}^{n\times n}$ in accordo alla seguente definizione.

Definizione 2.7 ***Matrice Inversa***
Sia $A \in \mathbb{R}^{n\times n}$. Si definisce matrice inversa di A e si indica con $A^{-1} \in \mathbb{R}^{n\times n}$ la matrice (se esiste - e in tal caso è unica) tale che

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

con I matrice identica, ossia $I_{ij} = 1$ se $i = j$; 0 altrimenti.

Chiaramente occorre escludere la matrice $A \equiv 0$ (matrice nulla, ossia con tutti gli elementi uguali a 0) in quanto è evidente che non esiste alcuna matrice che moltiplicata per tale matrice possa dare la matrice identica. Tuttavia, questo non è sufficiente come evidenzia l'esempio sotto riportato.

Esempio 2.4 *Si consideri*

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

e una generica matrice

$$B = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \in \mathbb{R}^{2\times 2}.$$

Si ha che

$$AB = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} b_{11} & b_{12} \\ 0 & 0 \end{bmatrix} \neq I$$

qualsiasi siano b_{11} *e* b_{12}.

È quindi evidente che tale questione richiede particolare attenzione, non essendo banale la risposta. Più precisamente le condizioni sotto le quali esiste la matrice inversa di un'assegnata matrice $A \in \mathbb{R}^{n\times n}$ sono riportate nel seguente Teorema.

Teorema 2.1 *Sia* $A \in \mathbb{R}^{n\times n}$. *La matrice* A *è invertibile (o non singolare), ossia esiste ed è unica* $A^{-1} \in \mathbb{R}^{n\times n}$ *tale che*

$$AA^{-1} = A^{-1}A = I \quad \textit{matrice identica},$$

se e solo se le colonne della matrice A*, pensate come vettori (colonna) di* $\mathbb{R}^n$*, sono fra loro linearmente indipendenti (ossia formano una base di* $\mathbb{R}^n$ *- si veda il Corollario 1.1), ovvero se e solo se il determinante della matrice* A *è diverso da zero.*

La funzione determinante di una matrice quadrata, che verrà definita nella sezione 2.5, è quindi uno strumento (molto comodo) per verificare la lineare indipendenza dei vettori.

2.4 Esercizi

1. Verificare che $V = A \in \mathbb{R}^{2\times 3}$ è spazio vettoriale rispetto alle operazioni di somma e prodotto per uno scalare di Definizione 2.2 e 2.3.

2. Sia

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \quad B = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}, \quad \underline{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad \underline{y} = [y_1,\ y_2].$$

   Calcolare le due espressioni equivalenti $(AB)\underline{x}$ e $A(B\underline{x})$. Contare il numero di operazioni moltiplicative effettuate e stabilire quale espressione è più conveniente.
   Calcolare le due espressioni equivalenti $\underline{y}(AB)$ e $(\underline{y}A)B$. Contare il numero di operazioni moltiplicative effettuate e stabilire quale espressione è più conveniente.

3. Sia
$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \quad B = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}.$$
Calcolare la matrice $C = AB$ e la matrice $D = BA$, verificando che, in generale $C \neq D$.

4. Sia
$$A = \begin{bmatrix} 1 & 1 & 2 \\ 2 & 0 & 2 \\ -1 & 3 & 2 \end{bmatrix}, \quad \underline{b} = \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}.$$
   - Verificare che $A\underline{b} = \underline{0}$;
   - dire se le colonne di A sono linearmente indipendenti fra loro;
   - dire se la matrice A è invertibile.

5. Sia $A \in \mathbb{R}^{n\times n}$ una matrice invertibile. Verificare che $A^{-1}A^2 = A$.
Sia
$$A = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}.$$
Sapendo che $A^2 = 2I$ e utilizzando la relazione $A^{-1}A^2 = A$, dare una formula esplicita per A^{-1}.

2.5 La funzione determinante di matrici quadrate

In questa sezione non considereremo la definizione formale di determinante di una matrice quadrata, ma ci limiteremo alla sua definizione nel caso di matrici 2×2 e alla **Formula di Laplace** (per righe o per colonne) per matrici di dimensione superiore, ovvero ci limiteremo alle formule che vengono usate per il calcolo effettivo.

Definizione 2.8 ***Determinante esempio di riferimento*** $n = 2$
Sia
$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \in \mathbb{R}^{2\times 2},$$
allora si definisce determinante la funzione
$$\begin{aligned} \det &: \mathbb{R}^{2\times 2} \to \mathbb{R} \\ \det(A) &= a_{11}a_{22} - a_{12}a_{21} \end{aligned}$$

Ora l'applicazione, eventualmente ripetuta, della **Formula di Laplace**, che stiamo per introdurre, permette di ricondurre il calcolo del determinante di una matrice $A \in \mathbb{R}^{n\times n}$, $n \geq 3$ al calcolo del determinate di opportune matrici 2×2. Ci occorre preliminarmente la seguente definizione.

Definizione 2.9 ***Minore di indici*** i, k
Sia $A \in \mathbb{R}^{n\times n}$ si dice Minore di indici i, k M_{ik} il determinante della sottomatrice di dimensione $n-1$ che si ottiene dalla matrice A cancellando l'$i-$sima riga e la $k-$sima colonna.

Esempio 2.5 *Sia*

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

allora il minore M_{23} *vale*

$$M_{23} = \det \begin{bmatrix} a_{11} & a_{12} \\ a_{31} & a_{32} \end{bmatrix} = a_{11}a_{32} - a_{12}a_{31},$$

ove la sottomatrice

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{31} & a_{32} \end{bmatrix}$$

è ottenuta dalla matrice A *cancellando la seconda riga e la terza colonna.*

Teorema 2.2 ***Formula di Laplace (per righe)***
Vale la seguente formula di sviluppo del determinante rispetto alla generica riga $i-$*sima di una matrice* $A \in \mathbb{R}^{n\times n}$.

$$\det(A) = \sum_{k=1}^{n} (-)^{i+k} a_{ik} M_{ik} \qquad \textit{con } i \textit{ fissato}$$

Esempio 2.6 ***Determinante esempio di riferimento*** $n = 3$
Sia

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \in \mathbb{R}^{3\times 3}.$$

- *Sviluppando rispetto alla prima riga* ($i = 1$) *si ha:*

$$\begin{aligned} \det(A) &= (-)^{1+1} a_{11} \det \begin{bmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{bmatrix} + (-)^{1+2} a_{12} \det \begin{bmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{bmatrix} \\ &\quad + (-)^{1+3} a_{13} \det \begin{bmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \\ &= a_{11}(a_{22}a_{33} - a_{23}a_{32}) - a_{12}(a_{21}a_{33} - a_{23}a_{31}) + a_{13}(a_{21}a_{32} - a_{22}a_{31}) \end{aligned}$$

- *Sviluppando rispetto alla seconda riga* ($i = 2$) *si ha:*

$$\begin{aligned} \det(A) &= (-)^{2+1} a_{21} \det \begin{bmatrix} a_{12} & a_{13} \\ a_{32} & a_{33} \end{bmatrix} + (-)^{2+2} a_{22} \det \begin{bmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{bmatrix} \\ &\quad + (-)^{2+3} a_{23} \det \begin{bmatrix} a_{11} & a_{12} \\ a_{31} & a_{32} \end{bmatrix} \\ &= -a_{21}(a_{12}a_{33} - a_{13}a_{32}) + a_{22}(a_{11}a_{33} - a_{13}a_{31}) - a_{23}(a_{11}a_{32} - a_{12}a_{31}) \end{aligned}$$

- *Sviluppando rispetto alla terza riga* ($i = 3$) *si ha:*

$$\begin{aligned} \det(A) &= (-)^{3+1} a_{31} \det \begin{bmatrix} a_{12} & a_{13} \\ a_{22} & a_{23} \end{bmatrix} + (-)^{3+2} a_{32} \det \begin{bmatrix} a_{11} & a_{13} \\ a_{21} & a_{23} \end{bmatrix} \\ &\quad + (-)^{33} a_{33} \det \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \\ &= a_{31}(a_{12}a_{23} - a_{13}a_{22}) - a_{32}(a_{11}a_{23} - a_{13}a_{21}) + a_{33}(a_{11}a_{22} - a_{12}a_{21}) \end{aligned}$$

Esempio 2.7 *In taluni casi può essere estremamente vantaggioso per ridurre i calcoli sviluppare il determinante rispetto ad una certa riga, piuttosto che rispetto ad altre, come evidenziato in questo esempio.*
Sia

$$A = \begin{bmatrix} 1 & 2 & 5 \\ 0 & 1 & 0 \\ 3 & 4 & 2 \end{bmatrix} \in \mathbb{R}^{3\times 3}.$$

Sviluppando rispetto alla seconda riga si ha da calcolare un solo determinante di sottomatrice 2×2 anziché tre. Infatti

$$\det(A) = (-)^{2+2} a_{22} \det \begin{bmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{bmatrix} = (-)^{2+2} 1 \det \begin{bmatrix} 1 & 5 \\ 3 & 2 \end{bmatrix} = 2 - 15 = -13$$

Vale un analogo del precedente Teorema 2.2 con sviluppo per colonne.

Teorema 2.3 ***Formula di Laplace (per colonne)***
Vale la seguente formula di sviluppo del determinante rispetto alla generica colonna $k-$sima di una matrice $A \in \mathbb{R}^{n\times n}$.

$$\det(A) = \sum_{i=1}^{n} (-)^{i+k} a_{ik} M_{ik} \qquad \text{con } k \text{ fissato}$$

Esempio 2.8 *Esempio di riferimento $n = 3$*
Sia

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \in \mathbb{R}^{3\times 3}.$$

- *Sviluppando rispetto alla prima colonna ($k = 1$) si ha:*

$$\begin{aligned} \det(A) &= (-)^{1+1} a_{11} \det \begin{bmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{bmatrix} + (-)^{1+2} a_{21} \det \begin{bmatrix} a_{12} & a_{13} \\ a_{32} & a_{33} \end{bmatrix} \\ &\quad + (-)^{1+3} a_{31} \det \begin{bmatrix} a_{12} & a_{13} \\ a_{22} & a_{23} \end{bmatrix} \\ &= a_{11}(a_{22}a_{33} - a_{23}a_{32}) - a_{21}(a_{12}a_{33} - a_{13}a_{32}) + a_{31}(a_{12}a_{23} - a_{22}a_{13}) \end{aligned}$$

- *Sviluppando rispetto alla seconda colonna ($k = 2$) si ha:*

$$\begin{aligned} \det(A) &= (-)^{2+1} a_{12} \det \begin{bmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{bmatrix} + (-)^{2+2} a_{22} \det \begin{bmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{bmatrix} \\ &\quad + (-)^{2+3} a_{32} \det \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{23} \end{bmatrix} \\ &= -a_{12}(a_{21}a_{22} - a_{31}a_{23}) + a_{22}(a_{11}a_{33} - a_{13}a_{31}) - a_{32}(a_{11}a_{23} - a_{13}a_{21}) \end{aligned}$$

- *Sviluppando rispetto alla terza colonna ($k = 3$) si ha:*

$$\begin{aligned} \det(A) &= (-)^{3+1} a_{13} \det \begin{bmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} + (-)^{3+2} a_{23} \det \begin{bmatrix} a_{11} & a_{12} \\ a_{31} & a_{32} \end{bmatrix} \\ &\quad + (-)^{33} a_{33} \det \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \\ &= a_{13}(a_{21}a_{32} - a_{31}a_{22}) - a_{23}(a_{11}a_{32} - a_{31}a_{12}) + a_{33}(a_{11}a_{22} - a_{12}a_{21}) \end{aligned}$$

Esempio 2.9 *In taluni casi può essere estremamente vantaggioso per ridurre i calcoli sviluppare il determinante rispetto ad una certa colonna, piuttosto che rispetto ad altre colonne o rispetto alle righe, come evidenziato in questo esempio.*
Sia

$$A = \begin{bmatrix} 1 & 0 & 3 \\ 2 & 1 & 4 \\ 5 & 0 & 2 \end{bmatrix} \in \mathbb{R}^{3\times 3}.$$

Sviluppando rispetto alla seconda colonna si ha da calcolare un solo determinante di sottomatrice 2×2 *anziché tre (come nel caso delle altre due colonne) o anziché due o tre (come nel caso dello sviluppo per righe). Infatti*

$$\det(A) = (-)^{2+2} a_{22} \det \begin{bmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{bmatrix} = (-)^{2+2} 1 \det \begin{bmatrix} 1 & 3 \\ 5 & 2 \end{bmatrix} = 2 - 15 = -13$$

2.6 Proprietà caratteristiche del determinante

La funzione determinante gode di alcune proprietà interessanti.

1. **Omogeneità**
   Se una costante $c \in \mathbb{R}$ moltiplica una fissata riga (colonna), il determinante risulta moltiplicato per c.
   Ad esempio, si ha

   $$\det \begin{bmatrix} ca_{11} & a_{12} & a_{13} \\ ca_{21} & a_{22} & a_{23} \\ ca_{31} & a_{32} & a_{33} \end{bmatrix} = c \det \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

   Infatti, basta considerare la Formula di Laplace applicata alla prima colonna: è evidente che si può mettere in evidenza la costante c e poi ricomporre la matrice 3×3.

   Come conseguenza si ha che se una matrice ha una riga (o una colonna) tutta nulla, il suo determinante è nullo (si può pensare quella riga (o colonna) come moltiplicata per $c = 0$).

2. **Invarianza per scorrimento**
   Sommando ad una riga (colonna) un'altra qualsiasi riga (colonna) moltiplicata per una costante $c \in \mathbb{R}$, il determinante risulta invariato.
   Ad esempio, si ha

   $$\det \begin{bmatrix} a_{11} + ca_{13} & a_{12} & a_{13} \\ a_{21} + ca_{23} & a_{22} & a_{23} \\ a_{31} + ca_{33} & a_{32} & a_{33} \end{bmatrix} = \det \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

   Come conseguenza si ha che se una matrice ha due righe (o una colonne) uguali, il suo determinante è nullo (si sottraggono le due righe (o colonne) uguali, ottenendo una riga (o colonna) tutta nulla).

3. **Linearità**
   La funzione determinate è lineare rispetto a ciascuna delle sue righe (colonne).

Ad esempio, si ha

$$\det \begin{bmatrix} a_{11}+\tilde{a}_{11} & a_{12} & a_{13} \\ a_{21}+\tilde{a}_{21} & a_{22} & a_{23} \\ a_{31}+\tilde{a}_{31} & a_{32} & a_{33} \end{bmatrix} =$$

$$= \det \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} + \det \begin{bmatrix} \tilde{a}_{11} & a_{12} & a_{13} \\ \tilde{a}_{21} & a_{22} & a_{23} \\ \tilde{a}_{31} & a_{32} & a_{33} \end{bmatrix}$$

Infatti, basta considerare la Formula di Laplace applicata alla prima colonna: è evidente che si possono separare gli addendi in evidenza $a_{ii}+\tilde{a}_{ii}$ e poi ricomporre le due matrici 3×3.

4. **Alternanza**
   Scambiando fra loro due righe (o colonne), il determinate della matrice cambia semplicemente di segno.
   Ad esempio, si ha

$$\det \begin{bmatrix} a_{12} & a_{11} & a_{13} \\ a_{22} & a_{21} & a_{23} \\ a_{32} & a_{31} & a_{33} \end{bmatrix} = -\det \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

2.7 Esercizi

1. Verificare le proprietà del determinante di sezione 2.6 sulle seguenti matrici

$$A = \begin{bmatrix} ca_{11} & a_{12} \\ ca_{21} & a_{22} \end{bmatrix}, B = \begin{bmatrix} b_{11}+cb_{12} & b_{12} \\ b_{21}+cb_{22} & b_{22} \end{bmatrix},$$

$$C = \begin{bmatrix} c_{11}+\tilde{c}_{11} & c_{12} \\ c_{21}+\tilde{c}_{21} & c_{22} \end{bmatrix}, D = \begin{bmatrix} d_{12} & d_{11} \\ d_{22} & d_{21} \end{bmatrix}$$

2. Calcolare il determinante di

$$A = \begin{bmatrix} 0 & \alpha_1 & 0 & \dots & \dots & 0 \\ 0 & 0 & \alpha_2 & 0 & \dots & 0 \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & 0 & \alpha_{n-2} & 0 \\ 0 & \dots & \dots & \dots & 0 & \alpha_{n-1} \\ \alpha_n & 0 & \dots & \dots & \dots & 0 \end{bmatrix}.$$

3. Calcolare il determinante delle seguenti matrici:

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 2 & 2 & 3 \\ 8 & 5 & 6 \\ 14 & 8 & 1 \end{bmatrix}, \quad C = \begin{bmatrix} 4 & 2 & 3 \\ 10 & 5 & 6 \\ 8 & 8 & 1 \end{bmatrix}.$$

2.8 Matrici particolari

Definizione 2.10 ***Matrice trasposta*** A^T
Sia $A \in \mathbb{R}^{n\times m}$ *allora si definisce matrice trasposta la matrice* $A^T \in \mathbb{R}^{m\times n}$ *tale che*

$$(A^T)_{ij} = a_{ji},\ i = 1,\dots,m; j = 1,\dots,n$$

ossia

$$A = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{k1} & \dots & a_{kn} \\ \vdots & & \vdots \\ a_{m1} & \dots & a_{mn} \end{bmatrix} \text{ e } A^T = \begin{bmatrix} a_{11} & \dots & a_{k1} & \dots & a_{n1} \\ \vdots & & \vdots & & \vdots \\ a_{1m} & \dots & a_{km} & \dots & a_{nm} \end{bmatrix}$$

Definizione 2.11 ***Matrice aggiunta*** A^H
Sia $A \in \mathbb{C}^{n\times m}$ *allora si definisce matrice aggiunta la matrice* $A^H \in \mathbb{C}^{m\times n}$ *tale che*

$$(A^H)_{ij} = \overline{a}_{ji},\ i = 1,\dots,m; j = 1,\dots,n.$$

ossia

$$A = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{k1} & \dots & a_{kn} \\ \vdots & & \vdots \\ a_{m1} & \dots & a_{mn} \end{bmatrix} \text{ e } A^H = \begin{bmatrix} \overline{a}_{11} & \dots & \overline{a}_{k1} & \dots & \overline{a}_{n1} \\ \vdots & & \vdots & & \vdots \\ \overline{a}_{1m} & \dots & \overline{a}_{km} & \dots & \overline{a}_{nm} \end{bmatrix}$$

Proposizione 2.3 *Valgono le seguenti proprietà:*

1. $(AB)^T = B^T A^T$ *con* $A \in \mathbb{R}^{n\times m}, B \in \mathbb{R}^{m\times l}$;
2. $(AB)^H = B^H A^H$ *con* $A \in \mathbb{C}^{n\times m}, B \in \mathbb{C}^{m\times l}$;
3. $(AB)^{-1} = B^{-1}A^{-1}$ *con* $A, B \in \mathbb{C}^{n\times n}$ *invertibili;*
4. $(A^H)^{-1} = (A^{-1})^H$ *con* $A \in \mathbb{C}^{n\times n}$ *invertibile.*

Per ricordare le precedenti proprietà si tengano presenti le seguenti semplici considerazioni (che non hanno pretesa di dimostrazione).

Relativamente alla 2) (e alla 1)): siano $A \in \mathbb{C}^{n\times m}, B \in \mathbb{C}^{m\times l}$ allora $C = (AB)^H \in \mathbb{C}^{l\times n}$. D'altra parte $A^H \in \mathbb{C}^{m\times n}$ e $B^H \in \mathbb{C}^{l\times m}$, quindi, per consistenza, si può effettuare il prodotto $D = A^H B^H$ se e solo se $n = l$ (cosa che non sarà in generale vera) e in tal caso $D \in \mathbb{C}^{m\times m}$ che avrà le stesse dimensioni della matrice C se e solo se $m = n = l$ (e ovviamente non è detto che C coincida con D visto che in genere non vale la proprietà commutativa). In definitiva, basta pensare al caso di vere e proprie matrici rettangolari per accorgersi che occorre invertire l'ordine dei fattori.

Relativamente alla 3): siano $A, B \in \mathbb{R}^{n\times n}$, si ha

$$(AB)(AB)^{-1} = ABB^{-1}A^{-1} = AIA^{-1} = AA^{-1} = I$$

come deve essere per definizione di matrice inversa. D'altra parte se fosse $(AB)^{-1} = A^{-1}B^{-1}$ si avrebbe $(AB)(AB)^{-1} = ABA^{-1}B^{-1}$ e poichè non vale in genere la proprietà commutativa, non si riuscirebbe ad ottenere la matrice I.

Proposizione 2.4 *Valgono le seguenti proprietà del determinante:*

1. $\det(AB) = \det(A)\det(B)$ *(Teorema di Binet);*
2. $\det(A^T) = \det(A)$*;*
3. $\det(A^H) = \overline{\det(A)}$*;*
4. $\det(A^{-1}) = 1/\det(A)$.

Per ricordare le precedenti proprietà si tengano presenti le seguenti semplici considerazioni.

Relativamente alla 2) (e alla 3) per estensione): si applichi la Formula di Laplace alla prima colonna di A^T e si applichi la Formula di Laplace alla prima riga di A, tenendo conto che la proprietà è evidente nel caso di matrici 2×2. L'estensione al caso della 3) segue dal fatto che $\overline{z_1 + z_2} = \overline{z_1} + \overline{z_2}$ e $\overline{z_1 z_2} = \overline{z_1}\ \overline{z_2}$ per ogni $\overline{z_1}, \overline{z_2} \in \mathbb{C}$.

Relativamente alla 4): vale che $\det(I) = \det(AA^{-1}) =_{1)} \det(A)\det(A^{-1})$ ove $\det(I) = 1$ (si pensi all'applicazione ripetuta della Formula di Laplace).

Definizione 2.12 *Sia* $A \in \mathbb{R}^{n\times n}(\mathbb{C}^{n\times n})$

- $A \in \mathbb{R}^{n\times n}$ *matrice* ***diagonale*** *se* $a_{ij} = 0$ *per* $i \neq j$*, ossia*
$$A = \begin{bmatrix} a_{11} & 0 & \dots & \dots & \dots & 0 \\ 0 & a_{22} & 0 & \dots & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & 0 & a_{n-1n-1} & 0 \\ 0 & \dots & \dots & \dots & 0 & a_{nn} \end{bmatrix};$$
- $A \in \mathbb{R}^{n\times n}$ *matrice* ***tridiagonale*** *se* $a_{ij} = 0$ *per* $|i - j| > 1$*, ossia*
$$A = \begin{bmatrix} a_{11} & a_{12} & 0 & \dots & \dots & 0 \\ a_{21} & a_{22} & a_{23} & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & a_{n-1n-2} & a_{n-1n-1} & a_{n-1n} \\ 0 & \dots & \dots & 0 & a_{nn-1} & a_{nn} \end{bmatrix};$$
- $A \in \mathbb{R}^{n\times n}$ *matrice* ***triangolare superiore*** *se* $a_{ij} = 0$ *per* $i > j$*, ossia*
$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & \dots & \dots & a_{1n} \\ 0 & a_{22} & a_{23} & \dots & \dots & a_{2n} \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & 0 & a_{n-1n-1} & a_{n-1n} \\ 0 & \dots & \dots & \dots & 0 & a_{nn} \end{bmatrix};$$

- $A \in \mathbb{R}^{n\times n}$ *matrice* ***triangolare inferiore*** *se* $a_{ij} = 0$ *per* $i < j$, *ossia*
$$A = \begin{bmatrix} a_{11} & 0 & \dots & \dots & \dots & 0 \\ a_{12} & a_{22} & 0 & \dots & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ a_{n-11} & \dots & \dots & \dots & a_{n-1n-1} & 0 \\ a_{n1} & \dots & \dots & \dots & a_{nn-1} & a_{nn} \end{bmatrix};$$
- $A \in \mathbb{R}^{n\times n}$ ***simmetrica*** *se* $A^T = A$;
- $A \in \mathbb{C}^{n\times n}$ ***Hermitiana*** *se* $A^H = A$;
- $A \in \mathbb{C}^{n\times n}$ ***definita positiva*** *se* $A^H = A$ *e per ogni* $\underline{x} \in \mathbb{C}^n, \underline{x} \neq 0$ *vale che* $\underline{x}^H A\underline{x} > 0$;
- $A \in \mathbb{C}^{n\times n}$ ***semidefinita positiva*** *se* $A^H = A$ *e per ogni* $\underline{x} \in \mathbb{C}^n$, *vale che* $\underline{x}^H A\underline{x} \geq 0$;
- $A \in \mathbb{C}^{n\times n}$ ***definita negativa*** *se* $A^H = A$ *e per ogni* $\underline{x} \in \mathbb{C}^n, \underline{x} \neq 0$ *vale che* $\underline{x}^H A\underline{x} < 0$;
- $A \in \mathbb{C}^{n\times n}$ ***semidefinita negativa*** *se* $A^H = A$ *e per ogni* $\underline{x} \in \mathbb{C}^n$, *vale che* $\underline{x}^H A\underline{x} \leq 0$;
- $A \in \mathbb{R}^{n\times n}$ ***ortogonale*** *se* $A^T A = AA^T = I$;
- $A \in \mathbb{C}^{n\times n}$ ***unitaria*** *se* $A^H A = AA^H = I$.

Proposizione 2.5 *Valgono le seguenti proprietà*

1. *A diagonale allora* $\det(A) = \prod_{i=1}^n a_{ii}$;
2. *A triangolare superiore (inferiore) allora* $\det(A) = \prod_{i=1}^n a_{ii}$;
3. *A definita positiva (negativa) allora* $\det(A) > 0 (< 0)$.

Si osservi che le proprietà 1) e 2) si verificano semplicemente facendo uso della Formula di Laplace, opportunamente applicata.

2.9 Esercizi

1. Sia
$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}, \quad D = \begin{bmatrix} d_1 & 0 & 0 \\ 0 & d_2 & 0 \\ 0 & 0 & d_3 \end{bmatrix}.$$
Calcolare AD e DA e dedurne la regola generale per matrici di dimensione n.

2. Sia
$$L = \begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix}, \quad \tilde{L} = \begin{bmatrix} \tilde{l}_{11} & 0 & 0 \\ \tilde{l}_{21} & \tilde{l}_{22} & 0 \\ \tilde{l}_{31} & \tilde{l}_{32} & \tilde{l}_{33} \end{bmatrix}.$$
Calcolare $L\tilde{L}$ e $\tilde{L}L$ e dedurne la regola generale per matrici di dimensione n.

3. Sia
$$U = \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix}, \quad \tilde{U} = \begin{bmatrix} \tilde{u}_{11} & \tilde{u}_{12} & \tilde{u}_{13} \\ 0 & \tilde{u}_{22} & \tilde{u}_{23} \\ 0 & 0 & \tilde{u}_{33} \end{bmatrix}.$$
Calcolare $U\tilde{U}$ e $\tilde{U}U$ e dedurne la regola generale per matrici di dimensione n.

4. Verificare che $\{D \in \mathbb{R}^{n\times n}$ tali che D matrice diagonale$\}$ è un sottospazio vettoriale di $V = \{A$ tali che $A \in \mathbb{R}^{n\times n}\}$ che ha dimensione n e darne una base.

5. Verificare che $\{L \in \mathbb{R}^{n\times n}$ tali che L matrice triangolare inferiore$\}$ è un sottospazio vettoriale di $V = \{A$ tali che $A \in \mathbb{R}^{n\times n}\}$ che ha dimensione $n(n+1)/2$ e darne una base.

6. Verificare che $\{U \in \mathbb{R}^{n\times n}$ tali che U matrice triangolare superiore$\}$ è un sottospazio vettoriale di $V = \{A$ tali che $A \in \mathbb{R}^{n\times n}\}$ che ha dimensione $n(n+1)/2$ e darne una base.

7. Verificare che $\{S \in \mathbb{R}^{n\times n}$ tali che S matrice simmetrica$\}$ è un sottospazio vettoriale di $V = \{A$ tali che $A \in \mathbb{R}^{n\times n}\}$ che ha dimensione $n(n+1)/2$ e darne una base.

3 Sistemi lineari

3.1 Generalità

Si consideri il sistema a coefficienti reali di m equazioni lineari in n incognite

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = b_2 \\ \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n = b_m \end{cases}$$

ovvero, in forma matriciale,

$$A\underline{x} = \underline{b} \text{ con } \; A \in \mathbb{R}^{m\times n}, \underline{x} \in \mathbb{R}^n \text{ e } \underline{b} \in \mathbb{R}^m.$$

Si vuole determinare, se esiste, la soluzione (eventualmente le soluzioni) del sistema lineare, vale a dire

$$\underline{x}^* = [x_1^*, \ldots, x_n^*]^T \in \mathbb{R}^n \text{ tale che } \; A\underline{x}^* = \underline{b}.$$

Per poter dare delle condizioni sull'esistenza e sul numero di soluzioni $\underline{x}^*$ è necessario introdurre la seguente nozione di rango di una matrice.

Definizione 3.1 ***rango di una matrice***
Sia $A \in \mathbb{R}^{m\times n}$. Si dice che la matrice A ha ***rango*** *r se*

- *A contiene una sottomatrice quadrata di dimensione r con determinante non nullo;*
- *ogni sottomatrice quadrata di A di dimensione maggiore a r ha determinante nullo.*

Esempio 3.1
Sia

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

La matrice A è tale che $\text{rango}(A) = 1$, in quanto la seconda riga uguaglia la prima riga moltiplicata per 2.
Sia

$$B = \begin{bmatrix} 1 & 2 & -1 \\ 2 & 4 & 1 \end{bmatrix}.$$

La matrice B è tale che $\text{rango}(B) = 2$. Infatti, ad esempio, è

$$\det \begin{bmatrix} 2 & -1 \\ 4 & 1 \end{bmatrix} = 6 \neq 0.$$

L'esistenza e il numero di soluzioni di un sistema lineare sono regolate dal seguente Teorema.

Teorema 3.1 ***di Rouché-Capelli***
Sia $A\underline{x} = \underline{b}$ *con* $A \in \mathbb{R}^{m\times n}$, $\underline{x} \in \mathbb{R}^n$, $\underline{b} \in \mathbb{R}^m$. *Il sistema lineare ammette soluzione se e solo se rango*(A) =*rango*$([A|b])$, *ove la matrice*

$$[A|b] = \left[\begin{array}{ccc|c} a_{11} & \dots & a_{1n} & b_1 \\ \vdots & & \vdots & \vdots \\ a_{m1} & \dots & a_{mn} & b_m \end{array}\right]$$

è la matrice ottenuta orlando la matrice A *con il termine noto* b. *Di più, se il sistema lineare ammette soluzione allora si hanno due possibilità:*

- *se rango*$(A) = n$, *ovvero pari al numero di incognite,il sistema lineare ammette un'unica soluzione (nel caso di matrici quadrate rango*$(A) = n$ *se e solo se* $\det(A) \neq 0$*);*
- *se rango*$(A) = k < n$, *il sistema ammette un'infinità di soluzioni e più precisamente il sistema ammette* ∞^{n-k} *soluzioni.*

Il significato del Teorema di Rouché-Capelli è il seguente: la relazione $\underline{b} = A\underline{x}$ si può interpretare come

$$\underline{b} = \begin{bmatrix} a_{11} \\ \vdots \\ a_{m1} \end{bmatrix} x_1 + \begin{bmatrix} a_{12} \\ \vdots \\ a_{m2} \end{bmatrix} x_2 + \ldots + \begin{bmatrix} a_{1n} \\ \vdots \\ a_{mn} \end{bmatrix} x_n,$$

ossia $\underline{b}$ è combinazione lineare dei vettori $a_{j\star} = [a_{1j}, \ldots, a_{mj}]^T$, $j = 1, \ldots, n$ dati dalle colonne della matrice A, con coefficienti della combinazione lineare dati da $x_1, \ldots, x_n$. Se fosse rango$([A|b])$ >rango(A), allora vorrebbe dire che i vettori $a_{1\star}, \ldots, a_{n\star}, \underline{b}$ sono linearmente indipendenti, il che contraddirrebbe la scrittura precedente.

Esempio 3.2 *Sia* $A\underline{x} = \underline{b}$ *con*

- $$A = \begin{bmatrix} 1 & 1 & -1 \\ 2 & 2 & 1 \\ 1 & 0 & 0 \end{bmatrix}, \quad \underline{b} = \begin{bmatrix} 1 \\ 5 \\ 1 \end{bmatrix}.$$

  Poiché rango$(A) = 3$ *(e ovviamente rango*$([A|\underline{b}])$ *non può essere maggiore di* 3*), il sistema ammette soluzione, e più precisamente ne ammette una ed una sola. Infatti, risolvendo per sostituzione, si ha*

  $$\left\{\begin{array}{l} x_1 + x_2 - x_3 = 1 \\ 2x_1 + 2x_2 + x_3 = 5 \\ x_1 = 1 \end{array}\right. \quad \text{ossia} \quad \left\{\begin{array}{l} x_2 - x_3 = 0 \\ 2x_2 + x_3 = 3 \\ x_1 = 1 \end{array}\right. \quad \text{ossia} \quad \left\{\begin{array}{l} x_2 = x_3 \\ x_3 = 1 \\ x_1 = 1 \end{array}\right.$$

  Quindi si ha una ed una sola soluzione data da $x_1^* = x_2^* = x_3^* = 1$.

- $$A = \begin{bmatrix} 1 & 1 & -1 \\ 2 & 2 & -2 \\ 1 & 0 & 0 \end{bmatrix}, \quad \underline{b} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

Poiché $rango(A) = 2$ *e* $rango([A|\underline{b}]) = 3$*, il secondo sistema non ammette soluzioni. Infatti, risolvendo per sostituzione, si ha*

$$\begin{cases} x_1 + x_2 - x_3 = 1 \\ 2x_1 + 2x_2 - 2x_3 = 1 \\ x_1 = 1 \end{cases} \quad \text{ossia} \quad \begin{cases} x_2 - x_3 = 0 \\ 2x_2 - 2x_3 = -1 \\ x_1 = 1 \end{cases} \quad \text{ossia} \quad \begin{cases} x_2 = x_3 \\ 0 = -1 \\ x_1 = 1 \end{cases}$$

il che è impossibile!

- $$A = \begin{bmatrix} 1 & 1 & -1 \\ 2 & 2 & -2 \\ 1 & 0 & 0 \end{bmatrix}, \quad \underline{b} = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}.$$

  Poiché $rango(A) = rango([A|\underline{b}]) = 2 < n$*, il sistema ammette soluzione e più precisamente ne ammette* ∞^1*. Infatti, risolvendo per sostituzione, si ha*

  $$\begin{cases} x_1 + x_2 - x_3 = 1 \\ 2x_1 + 2x_2 - 2x_3 = 2 \\ x_1 = 1 \end{cases} \quad \text{ossia} \quad \begin{cases} x_2 - x_3 = 0 \\ x_2 - x_3 = 0 \\ x_1 = 1 \end{cases} \quad \text{ossia} \quad \begin{cases} x_2 = x_3 \\ x_1 = 1. \end{cases}$$

  Quindi si hanno ∞^1 *soluzioni del tipo* $x_1^* = 1, x_2^* = x_3^*$, $x_3^* \in \mathbb{R}$ *qualsiasi.*

- $$A = \begin{bmatrix} 1 & 1 & -1 \\ 2 & 2 & -2 \\ -3 & -3 & 3 \end{bmatrix}, \quad \underline{b} = \begin{bmatrix} 1 \\ 2 \\ -3 \end{bmatrix}.$$

  Poiché $rango(A) = rango([A|\underline{b}]) = 1 < n$*, il sistema ammette soluzione e più precisamente ne ammette* ∞^2*. Infatti, risolvendo per sostituzione, si ha*

  $$\begin{cases} x_1 + x_2 - x_3 = 1 \\ 2x_1 + 2x_2 - 2x_3 = 2 \\ 3x_1 + 3x_2 - 3x_3 = 3 \end{cases} \quad \text{ossia} \quad \begin{cases} x_1 + x_2 - x_3 = 1 \\ x_1 + x_2 - x_3 = 1 \\ x_1 + x_2 - x_3 = 1 \end{cases}$$

  Quindi si hanno ∞^2 *soluzioni del tipo* $x_3^* = x_1^* + x_2^* - 1$, $x_1^*, x_2^* \in \mathbb{R}$ *qualsiasi.*

Consideriamo, ora, il caso particolare dei sistemi lineari con termine noto pari al vettore nullo.

Definizione 3.2 ***Sistema lineare omogeneo***
Si dice sistema lineare omogeneo un sistema lineare del tipo

$$A\underline{x} = \underline{0} \text{ con } \ A \in \mathbb{R}^{m \times n}, \underline{x} \in \mathbb{R}^n \text{ e } \ \underline{0} \in \mathbb{R}^m.$$

Come semplice applicazione del Teorema di Rouchè-Capelli al caso di sistemi lineari omogenei si ha la seguente proposizione.

Proposizione 3.1
Sia

$$A\underline{x} = \underline{0} \text{ con } \ A \in \mathbb{R}^{m \times n}, \underline{x} \in \mathbb{R}^n, \underline{0} \in \mathbb{R}^m.$$

La soluzione banale $x_1^* = \ldots = x_n^* = 0$ *è sempre soluzione. Il sistema ammette anche soluzioni diverse da quella banale (e sono ovviamente infinite) se e solo se* $rango(A) < n$ *(nel caso* $A \in \mathbb{R}^{n \times n}$ *se e solo se* $\det(A) = 0$*).*

Esempi importanti di sistemi lineari omogenei verranno considerati nella successiva sezione 4.

3.2 Applicazioni lineari da $V = \mathbb{R}^m$ a $U = \mathbb{R}^n$

Il problema della risoluzione di un sistema lineare $A\underline{x} = \underline{b}$ con $A \in \mathbb{R}^{m\times n}, \underline{x} \in \mathbb{R}^n, \underline{b} \in \mathbb{R}^m$ ha un'interpretazione interessante quando inquadrato nella teoria delle applicazioni lineari fra spazi vettoriali. Per un'applicazione di rilievo di questa interpretazione si veda la sezione 4.

Definizione 3.3 ***Applicazione lineare fra spazi vettoriali***
Un'applicazione $F : V \to U$, ove $V = \mathbb{R}^m$ e $U = \mathcal{R}^n$, si dice lineare se

- *per ogni $\underline{x}_1, \underline{x}_2 \in V$ si ha $F(\underline{x}_1 + \underline{x}_2) = F(\underline{x}_1) + F(\underline{x}_2))$,*
- *per ogni $\alpha \in K = \mathbb{R}$ e per ogni $\underline{x} \in V$ si ha $F(\alpha\underline{x}) = \alpha F(\underline{x})$.*

Esempio 3.3 *Sia $V = U = \mathbb{R}^2$ e si consideri l'applicazione che associa ad un vettore del piano la sua proiezione sull'asse delle ascisse, ossia*

$$F\left(\begin{bmatrix} v_1 \\ v_2 \end{bmatrix}\right) = \begin{bmatrix} v_1 \\ 0 \end{bmatrix}$$

per ogni $\underline{v} = [v_1, v_2]^T$.
Tale applicazione è chiaramente un'applicazione lineare in quanto vale che

- *per ogni $\underline{v}, \underline{w} \in V$ si ha $F(\underline{v} + \underline{w} = F(\underline{v}) + F(\underline{w})) = [v_1 + w_1, 0]^T$,*
- *per ogni $\alpha \in K = \mathbb{R}$ e per ogni $\underline{v} \in V$ si ha $F(\alpha\underline{v}) = \alpha F(\underline{v}) = [\alpha v_1, 0]^T$.*

Ora consideriamo la base canonica di $V = \mathbb{R}^m$, ossia gli m vettori $\underline{e}_1 = [1\,0\ldots0]^T$, $\underline{e}_2 = [0\ 1\ 0\ldots0]^T$, ..., $\underline{e}_m = [0\ldots0\ 1]^T$. Grazie alla proprietà di linearità dell'applicazione F, per sapere come opera tale applicazione su un vettore generico

$$\underline{x} = x_1\underline{e}_1 + x_2\underline{e}_2 + \ldots + x_m\underline{e}_m \in \mathbb{R}^m$$

è sufficiente sapere come opera F su ciascuno dei vettori della base canonica di $V = \mathbb{R}^m$ in quanto

$$\begin{aligned} \underline{y} = F(\underline{x}) &= F(x_1\underline{e}_1 + x_2\underline{e}_2 + \ldots + x_m\underline{e}_m) \\ &= x_1F(\underline{e}_1) + x_2F(\underline{e}_2) + \ldots + x_mF(\underline{e}_m) \end{aligned}$$

Ora, poniamo, per ogni $k = 1, \ldots, m$

$$F(\underline{e}_k) = \begin{bmatrix} a_{1k} \\ a_{2k} \\ \vdots \\ a_{nk} \end{bmatrix}$$

ove il primo indice denota la componente del vettore $F(\underline{e}_k)$; mentre il secondo indice è fissato uguale a k ad indicare che si stà rappresentando il vettore $F(\underline{e}_k)$.

Quindi, costruiamo una matrice di n righe e m colonne accostando gli m vettori colonna $F(\underline{e}_k),\ k = 1, \ldots, m$, nell'ordine, ottenendo così

$$A = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1m} \\ a_{21} & a_{22} & \ldots & a_{2m} \\ \vdots & \vdots & \ldots & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nm} \end{bmatrix} \in \mathbb{R}^{n \times m}.$$

Ora, l'espressione della $i-$sima componente del vettore $\underline{y} \in \mathbb{R}^n$ è data da

$$\begin{aligned} \underline{y}_i = F(\underline{x})_i &= x_1 F(\underline{e}_1)_i + x_2 F(\underline{e}_2)_i + \ldots + x_m F(\underline{e}_m)_i \\ &= x_1 a_{i1} + x_2 a_{i2} + \ldots + x_m a_{im} \\ &= \sum_{j=1}^{m} a_{ij} x_j, \end{aligned}$$

e corrisponde a quella ottenuta considerando

$$\underline{y}_i = (A\underline{x})_i$$

con $\underline{y}$ vettore risultato del prodotto della matrice A per il vettore colonna $\underline{x}$. Pertanto, si può concludere che le matrici sono un modo compatto per rappresentare l'azione di un'applicazione lineare F su un qualsivoglia vettore.

Inoltre, il problema della risoluzione di un sistema lineare $A\underline{x} = \underline{b}$ con $A \in \mathbb{R}^{m \times n}, \underline{x} \in \mathbb{R}^n, \underline{b} \in \mathbb{R}^m$ risulta equivalente a quello di determinare i vettori $\underline{x}^*$, se esistono, che vengono trasformati dall'applicazione lineare F assegnata nel termine noto $\underline{b}$ assegnato. Ritorneremo sul significato di questa interpretazione nella sezione 4.

Esempio 3.4

1. *Si consideri l'applicazione lineare $F : \mathbb{R}^3 \to \mathbb{R}^3$ definita tramite le relazioni*

$$\begin{aligned} F([1\ -1\ 1]^T) &= [0\ 1+k\ k-1]^T \\ F([1\ 1\ 1]^T) &= [0\ 1+k\ 1+k]^T \\ F([0\ 0\ 1]^T) &= [1\ 0\ k]^T \end{aligned}$$

L'applicazione lineare è univocamente determinata in quanto i vettori $\underline{v}_1 = [1\ -1\ 1]^T$, $\underline{v}_2 = [1\ 1\ 1]^T$ e $\underline{v}_3 = [0\ 0\ 1]^T$ formano una base di $\mathbb{R}^3$. Infatti sono in numero di 3 e sono linearmente indipendenti essendo

$$\det \begin{bmatrix} 1 & 1 & 0 \\ -1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \neq 0$$

Si vuole determinare la matrice associata a F rispetto alla base canonica di $\mathbb{R}^3$. Per fare questo, occorre determinare le immagini degli elementi della base canonica di $\mathbb{R}^3$ tramite l'applicazione lineare F e scriverne le coordinate sempre rispetto a tale base.
Ora, per definizione di applicazione lineare, vale che se il generico vettore $\underline{v} \in \mathbb{R}^3$ ha coordinate $[a\ b\ c]^T$ rispetto alla base $\underline{v}_1, \underline{v}_2, \underline{v}_3$, ovvero $\underline{v} =$

$a\underline{v}_1 + b\underline{v}_2 + c\underline{v}_3$, allora $F(\underline{v}) = aF(\underline{v}_1) + bF\underline{v}_2) + cF\underline{v}_3)$.
Quindi, poiché vale che il primo vettore della base canonica si può scrivere come

$$\underline{e}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \frac{1}{2}\underline{v}_1 + \frac{1}{2}\underline{v}_2 - 1\underline{v}_3,$$

si ha che

$$\begin{aligned} F(\underline{e}_1) &= \frac{1}{2}F(\underline{v}_1) + \frac{1}{2}F(\underline{v}_2) - 1F(\underline{v}_3) \\ &= \frac{1}{2}\begin{bmatrix} 0 \\ 1+k \\ k-1 \end{bmatrix} + \frac{1}{2}\begin{bmatrix} 0 \\ 1+k \\ 1+k \end{bmatrix} - \begin{bmatrix} 1 \\ 0 \\ k \end{bmatrix} = \begin{bmatrix} -1 \\ 1+k \\ 0 \end{bmatrix}. \end{aligned}$$

Ora, poiché vale che il secondo vettore della base canonica si può scrivere come

$$\underline{e}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = -\frac{1}{2}\underline{v}_1 + \frac{1}{2}\underline{v}_2 + 0\underline{v}_3,$$

si ha che

$$\begin{aligned} F(\underline{e}_2) &= -\frac{1}{2}F(\underline{v}_1) + \frac{1}{2}F(\underline{v}_2) + 0F(\underline{v}_3) \\ &= -\frac{1}{2}\begin{bmatrix} 0 \\ 1+k \\ k-1 \end{bmatrix} + \frac{1}{2}\begin{bmatrix} 0 \\ 1+k \\ 1+k \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}. \end{aligned}$$

Ora, poiché il terzo vettore della base canonica è il terzo vettore rispetto a cui è definita l'applicazione lineare non occorre fare altri calcoli e si ottiene che la matrice A rappresentativa dell'applicazione lineare è

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

2. *Si consideri l'applicazione lineare $F : \mathbb{R}^4 \to \mathbb{R}^2$ definita tramite le relazioni*

$$\begin{aligned} F[1\ 0\ 1\ 0]^T) &= [1\ 0]^T \\ F([-1\ 0\ 1\ 0]^T) &= [-1\ 0]^T \\ F([0\ 0\ 1\ 1]^T) &= [0\ 2]^T \\ F([0\ 1\ 0\ -1]^T) &= [1\ 1]^T \end{aligned}$$

L'applicazione lineare è univocamente determinata in quanto i vettori $\underline{v}_1 = [1\ 0\ 1\ 0]^T$, $\underline{v}_2 = [-1\ 0\ 1\ 0]^T$, $\underline{v}_3 = [0\ 0\ 1\ 1]^T$ e $\underline{v}_4 = [0\ 1\ 0\ -1]^T$ formano una base di $\mathbb{R}^4$. Infatti sono in numero di 4 e sono linearmente indipendenti essendo

$$\det \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix} \neq 0$$

Si vuole determinare la matrice associata a F rispetto alle basi canoniche di $\mathbb{R}^4$ e $\mathbb{R}^2$. Per fare questo, occorre determinare le immagini degli elementi della base canonica di $\mathbb{R}^4$ tramite l'applicazione lineare F e scriverne le coordinate rispetto alla base canonica di $\mathbb{R}^2$.
Ora, vale che se il generico vettore $\underline{v} \in \mathbb{R}^4$ ha coordinate $[a\ b\ c\ d]^T$ rispetto alla base $\underline{v}_1, \underline{v}_2, \underline{v}_3, \underline{v}_4$, ovvero $\underline{v} = a\underline{v}_1 + b\underline{v}_2 + c\underline{v}_3 + d\underline{v}_4$, allora $F(\underline{v}) = aF(\underline{v}_1) + bF(\underline{v}_2) + cF(\underline{v}_3) + dF(\underline{v}_4)$.
Quindi, poiché vale che il primo vettore della base canonica si può scrivere come

$$\underline{e}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \frac{1}{2}\underline{v}_1 - \frac{1}{2}\underline{v}_2 + 0\underline{v}_3 + 0\underline{v}_4,$$

si ha che

$$\begin{aligned} F(\underline{e}_1) &= \frac{1}{2}F(\underline{v}_1) - \frac{1}{2}F(\underline{v}_2) + 0F(\underline{v}_3) + 0F(\underline{v}_4) \\ &= \frac{1}{2}\begin{bmatrix} 1 \\ 0 \end{bmatrix} - \frac{1}{2}\begin{bmatrix} -1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \end{aligned}$$

Ora, poiché vale che il secondo vettore della base canonica si può scrivere come

$$\underline{e}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} = -\frac{1}{2}\underline{v}_1 - \frac{1}{2}\underline{v}_2 + 1\underline{v}_3 + 1\underline{v}_4,$$

si ha che

$$\begin{aligned} F(\underline{e}_2) &= -\frac{1}{2}F(\underline{v}_1) - \frac{1}{2}F(\underline{v}_2) + 1F(\underline{v}_3) + 1F(\underline{v}_4) \\ &= -\frac{1}{2}\begin{bmatrix} 1 \\ 0 \end{bmatrix} - \frac{1}{2}\begin{bmatrix} -1 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 2 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \end{bmatrix}. \end{aligned}$$

Ripetendo il procedimento anche per i restanti due vettori si ottiene che la matrice A rappresentativa dell'applicazione lineare è

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

3.3 Un metodo di risoluzione per i sistemi lineari

Anziché considerare il classico metodo di Cramer (che fa uso della teoria del determinante e che è computazionalmente molto costoso (numero di operazioni dell'ordine di $n!$ con n dimensione del sistema lineare)), si considera un metodo di risoluzione alternativo che si basa sostanzialmente sulla nozione di sistemi lineari equivalenti.

Definizione 3.4 ***Sistemi lineari equivalenti***
Si considerino due sistemi lineari

$$\begin{aligned} A\underline{x} &= \underline{b} \text{ con } A \in \mathbb{R}^{m\times n}, \underline{x} \in \mathbb{R}^n, \underline{b} \in \mathbb{R}^m; \\ C\underline{y} &= \underline{d} \text{ con } C \in \mathbb{R}^{m\times n}, \underline{y} \in \mathbb{R}^n, \underline{d} \in \mathbb{R}^m. \end{aligned}$$

I due sistemi lineari si dicono equivalenti se ogni soluzione del primo sistema è soluzione del secondo sistema e viceversa.

Consideriamo ora due esempi di sistemi lineari equivalenti che sono alla base del metodo di risoluzione che si vuole qui di seguito introdurre.

Proposizione 3.2 *Si consideri il sistema lineare*

$$I)\begin{cases} R_1 = a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n - b_1 = 0 \\ \vdots \\ R_{i-1} = a_{i-11}x_1 + a_{i-12}x_2 + \ldots + a_{i-1n}x_n - b_{i-1} = 0 \\ R_i = a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{in}x_n - b_i = 0 \\ R_{i+1} = a_{i+11}x_1 + a_{i+12}x_2 + \ldots + a_{i+1n}x_n - b_{i+1} = 0 \\ \vdots \\ R_{j-1} = a_{j-11}x_1 + a_{j-12}x_2 + \ldots + a_{j-1n}x_n - b_{j-1} = 0 \\ R_j = a_{j1}x_1 + a_{j2}x_2 + \ldots + a_{jn}x_n - b_j = 0 \\ R_{j+1} = a_{j+11}x_1 + a_{j+12}x_2 + \ldots + a_{j+1n}x_n - b_{j+1} = 0 \\ \vdots \\ R_m = a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n - b_m = 0 \end{cases}$$

e il sistema lineare

$$II)\begin{cases} R_1 = a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n - b_1 = 0 \\ \vdots \\ R_{i-1} = a_{i-11}x_1 + a_{i-12}x_2 + \ldots + a_{i-1n}x_n - b_{i-1} = 0 \\ R_j = a_{j1}x_1 + a_{j2}x_2 + \ldots + a_{jn}x_n - b_j = 0 \\ R_{i+1} = a_{i+11}x_1 + a_{i+12}x_2 + \ldots + a_{i+1n}x_n - b_{i+1} = 0 \\ \vdots \\ R_{j-1} = a_{j-11}x_1 + a_{j-12}x_2 + \ldots + a_{j-1n}x_n - b_{j-1} = 0 \\ R_i = a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{in}x_n - b_i = 0 \\ R_{j+1} = a_{j+11}x_1 + a_{j+12}x_2 + \ldots + a_{j+1n}x_n - b_{j+1} = 0 \\ \vdots \\ R_m = a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n - b_m = 0 \end{cases}$$

che differisce dal sistema lineare I) per il solo fatto che la $j-$sima equazione $R_j = 0$ è stata scambiata con l'equazione $i-$sima $R_i = 0$. Il sistema lineare I) e il sistema lineare II) sono equivalenti.

Dimostrazione: l'affermazione è ovvia in quanto le soluzioni di un sistema lineare non dipendono dall'ordine delle equazioni. •

Proposizione 3.3 *Si consideri il sistema lineare*

$$I)\begin{cases} R_1 = a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n - b_1 = 0 \\ R_2 = a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n - b_2 = 0 \\ \vdots \\ R_{j-1} = a_{j-11}x_1 + a_{j-12}x_2 + \ldots + a_{j-1n}x_n - b_{j-1} = 0 \\ R_j = a_{j1}x_1 + a_{j2}x_2 + \ldots + a_{jn}x_n - b_j = 0 \\ R_{j+1} = a_{j+11}x_1 + a_{j+12}x_2 + \ldots + a_{j+1n}x_n - b_{j+1} = 0 \\ \vdots \\ R_m = a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n - b_m = 0 \end{cases}$$

e il sistema lineare

$$II)\begin{cases} R_1 = 0 \\ R_2 = 0 \\ \vdots \\ R_{j-1} = 0 \\ \tilde{R}_j = \alpha_1 R_1 + \alpha_2 R_2 + \ldots + \alpha_{j-1} R_{j-1} + \alpha_j R_j + \alpha_{j+1} R_{j+1} + \ldots + \alpha_m R_m = 0 \\ R_{j+1} = 0 \\ \vdots \\ R_m = 0 \end{cases}$$

che differisce dal sistema lineare I) per la sola $j-$sima equazione $\tilde{R}_j = 0$, che è una combinazione lineare delle m equazioni $R_1 = 0, \ldots, R_m = 0$ del sistema I) con coefficienti della combinazione lineare dati da $\alpha_1, \ldots, \alpha_m$, con $\alpha_j \neq 0$.
Il sistema lineare I) e il sistema lineare II) sono equivalenti.

Dimostrazione:
Si assume che $x_1^*, \ldots, x_n^*$ siano una soluzione del sistema $I)$ e si vuole verificare che $x_1^*, \ldots, x_n^*$ sono una soluzione del sistema $II)$.
Dalle ipotesi fatte si ha che

$$\begin{cases} R_1 = a_{11}x_1^* + a_{12}x_2^* + \ldots + a_{1n}x_n^* - b_1 \equiv 0 \\ \vdots \\ R_j = a_{j1}x_1^* + a_{j2}x_2^* + \ldots + a_{jn}x_n^* - b_j \equiv 0 \\ \vdots \\ R_m = a_{m1}x_1^* + a_{m2}x_2^* + \ldots + a_{mn}x_n^* - b_m \equiv 0 \end{cases}$$

Quindi, anche la $j-$sima equazione del sistema $II)$ è soddisfatta essendo

$$\tilde{R}_j = \alpha_1 0 + \ldots + \alpha_j 0 + \ldots + \alpha_m 0 \equiv 0$$

e ovviamente lo sono pure le rimanenti equazioni $R_1 = 0, \ldots, R_{j-1} = 0, R_{j+1} = 0, \ldots, R_m = 0$ essendo le medesime del sistema $I)$.
Viceversa, si assume che $x_1^*, \ldots, x_n^*$ siano una soluzione del sistema $II)$ e si vuole verificare che $x_1^*, \ldots, x_n^*$ sono una soluzione del sistema $I)$.
Dalle ipotesi fatte si ha che $x_1^*, \ldots, x_n^*$ soddisfano le equazioni

$$\begin{cases} R_1 = a_{11}x_1^* + a_{12}x_2^* + \ldots + a_{1n}x_n^* - b_1 \equiv 0 \\ \vdots \\ R_{j-1} = a_{j-11}x_1^* + a_{j-12}x_2^* + \ldots + a_{j-1n}x_n^* - b_{j-1} \equiv 0 \\ R_{j+1} = a_{j+11}x_1^* + a_{j+12}x_2^* + \ldots + a_{j+1n}x_n^* - b_{j+1} \equiv 0 \\ \vdots \\ R_m = a_{m1}x_1^* + a_{m2}x_2^* + \ldots + a_{mn}x_n^* - b_m \equiv 0 \end{cases}$$

ossia tutte, tranne la $j-$sima equazione, per entrambi i sistemi lineari. Ora sostituendo nella $j-$sima equazione del sistema $II)$ si ha

$$\tilde{R}_j = \alpha_1 0 + \ldots + \alpha_{j-1} 0 + \alpha_j R_j + \alpha_{j+1} 0 + \ldots + \alpha_m 0 = \alpha_j R_j \equiv 0$$

poiché $x_1^*, \ldots, x_n^*$ sono una soluzione del sistema II). Inoltre, poiché per ipotesi è $\alpha_j \neq 0$, deve essere pure $R_j \equiv 0$ ossia $x_1^*, \ldots, x_n^*$ soddisfano anche la $j-$sima equazione del sistema I).

Riassumendo, si è verificato che scambiare fra loro equazioni o sostituire ad un'equazione una combinazione lineare delle equazioni del sistema, con il solo vincolo $\alpha_j \neq 0$, permette di ottenere un sistema lineare equivalente, ossia con tutte e sole le soluzioni del sistema di partenza.
Ora, un'applicazione ripetuta e mirata di tale tecnica permette di trasformare un qualsivoglia sistema lineare assegnato in sistema lineare equivalente, ma di più facile risoluzione (anche rispetto alla risoluzione su calcolatore).
Vediamo il metodo su un esempio.

Esempio 3.5 *Sia* $A\underline{x} = \underline{b}$ *con*

$$A = \begin{bmatrix} 1 & -2 & 3 \\ 3 & -2 & 3 \\ 4 & -4 & 1 \end{bmatrix}, \quad \underline{b} = \begin{bmatrix} 2 \\ 4 \\ 1 \end{bmatrix},$$

ovvero

$$\begin{cases} R_1 = x_1 - 2x_2 + 3x_3 - 2 = 0 \\ R_2 = 3x_1 - 2x_2 + 3x_3 - 4 = 0 \\ R_3 = 4x_1 - 4x_2 + x_3 - 1 = 0 \end{cases}$$

del quale esiste unica la soluzione $\underline{x}^*$ *tale che* $A\underline{x}^* = \underline{b}$, *poiché* $\det(A) = -20 \neq 0$. *I* ***passo****: si sostituisce all'equazione* $R_2 = 0$ *la combinazione lineare* $R_2 - 3R_1 = 0$, *ove si è scelto come coefficiente della combinazione lineare* $3 = a_{21}/a_{11}$. *Analogamente si sostituisce all'equazione* $R_3 = 0$ *la combinazione lineare* $R_3 - 4R_1 = 0$, *ove si è scelto come coefficiente della combinazione lineare* $4 = a_{31}/a_{11}$. *Si considera quindi il sistema lineare equivalente*

$$\begin{cases} R_1 = x_1 - 2x_2 + 3x_3 - 2 = 0 \\ R_2 - 3R_1 = 3x_1 - 2x_2 + 3x_3 - 4 - 3(x_1 - 2x_2 + 3x_3 - 2) = 0 \\ R_3 - 4R_1 = 4x_1 - 4x_2 + x_3 - 1 - 4(x_1 - 2x_2 + 3x_3 - 2) = 0 \end{cases}$$

ossia

$$\begin{cases} R_1^{(1)} = R_1 = x_1 - 2x_2 + 3x_3 - 2 = 0 \\ R_2^{(1)} = 4x_2 - 6x_3 + 2 = 0 \\ R_3^{(1)} = 4x_2 - 11x_3 + 7 = 0 \end{cases}$$

II ***passo****: si sostituisce all'equazione* $R_3^{(1)} = 0$ *la combinazione lineare* $R_3^{(1)} - R_2^{(1)} = 0$, *ove si è scelto come coefficiente della combinazione lineare* $1 = a_{32}^{(1)}/a_{22}^{(1)}$. *Si ha quindi*

$$\begin{cases} R_1^{(1)} = x_1 - 2x_2 + 3x_3 - 2 = 0 \\ R_2^{(1)} = 4x_2 - 6x_3 + 2 = 0 \\ R_3^{(1)} - 2R_2^{(1)} = 4x_2 - 11x_3 + 7 - (4x_2 - 6x_3 + 2) = 0 \end{cases}$$

ossia

$$\begin{cases} R_1^{(2)} = R_1 = x_1 - 2x_2 + 3x_3 - 2 = 0 \\ R_2^{(2)} = R_2^{(1)} = 4x_2 - 6x_3 + 2 = 0 \\ R_3^{(2)} = -5x_3 + 5 = 0 \end{cases}$$

In definitiva, si è trasformato il sistema di partenza nel sistema lineare equivalente $\tilde{A}\underline{x} = \underline{\tilde{b}}$ *con*

$$\tilde{A} = \begin{bmatrix} 1 & -2 & 3 \\ 0 & 4 & -6 \\ 0 & 0 & -5 \end{bmatrix}, \quad \underline{\tilde{b}} = \begin{bmatrix} 2 \\ -2 \\ -5 \end{bmatrix}.$$

Ora, un sistema lineare, la cui matrice sia ***triangolare superiore****, si può risolvere facilmente (a mano o su calcolatore) con una procedura che prende il nome di* ***risoluzione Backward*** *(a ritroso).*
Infatti, dall'ultima equazione si ricava $x_3^* = 1$*; quindi sostituendo nella seconda equazione il valore trovato per* x_3^* *si ottiene*

$$x_2^* = (6x_3^* - 2)/4 = 1.$$

Infine, sostituendo nella prima equazione il valore trovato per x_2^* *e* x_3^* *si ottiene*

$$x_1^* = 2x_2^* - 3x_3^* + 2 = 1.$$

Qualora all'$k-$mo passo ci si trovasse ad avere che l'elemento in posizione k, k è nullo, è sufficiente scambiare l'$k-$sima equazione con una delle successive tale che il coefficiente relativo all'incognita x_k sia non nullo.
Si tenga presente che sotto l'ipotesi di A non singolare tale equazione alternativa esiste sempre.

3.4 Esercizi

1. In dipendenza di $k \in \mathbb{R}$, determinare il rango della matrice

$$A = \begin{bmatrix} 3 & 4 & 5 \\ 1 & 3 & 2 \\ k & k & k \end{bmatrix}$$

   e della matrice

$$B = \begin{bmatrix} 3 & -4 & 1 & 5 \\ 1 & -3 & 1 & 2 \\ k & k & k & k \end{bmatrix}.$$

2. Discutere in dipendenza del parametro $k \in \mathbb{R}$ la risolubilità del sistema lineare

$$\begin{cases} (k+1)y + 10z &= 8 \\ y + 4z &= 4 \\ x + 4y + 16z &= 9 \end{cases}$$

   e quando possibile determinarne le soluzioni.

3. Discutere in dipendenza del parametro $k \in \mathbb{R}$ la risolubilità del sistema lineare

$$\begin{cases} x + (k - k^2)z &= 2 - k \\ (k - k^2)z &= 1 - k \\ x - y &= 0 \end{cases}$$

   e quando possibile determinarne le soluzioni.

4. Discutere in dipendenza del parametro $k \in \mathbb{R}$ la risolubilità del sistema lineare
$$\begin{cases} 3x + 2y + kz + 4t &= k \\ 2x + y + kz + 3t &= 0 \\ x + y + 3t &= 1 \end{cases}$$
e per $k = 1$ determinare le soluzioni.

5. Discutere in dipendenza del parametro $k \in \mathbb{R}$ la risolubilità del sistema lineare
$$\begin{cases} x + y + kz + (k-1)w &= k+1 \\ ky + k(k+1)z + (k+1)w &= k+1 \\ x + y &= k \end{cases}$$
e quando possibile determinare le soluzioni.

6. Discutere in dipendenza del parametro $k \in \mathbb{R}$ la risolubilità del sistema lineare
$$\begin{cases} kx + y + z &= 0 \\ ky + z + 1 &= 0 \\ y + kz - 1 &= 0 \end{cases}$$
e per $k = 2$ determinare le soluzioni.

4 Autovettori e autovalori

4.1 Cambiamenti di base

Sia V uno spazio vettoriale tale che $\dim V = n$.
Si è visto in sezione 1.2 che uno spazio vettoriale ammette basi distinte, ma tutte con la medesima cardinalità. Approfondiamo ulteriormente questo argomento ponendoci la seguente domanda:

come variano le coordinate di un vettore $\underline{v}$ al variare della base rispetto a cui lo si rappresenta?

Siano $\underline{g}_1, \underline{g}_2, \ldots, \underline{g}_n$ e $\underline{h}_1, \underline{h}_2, \ldots, \underline{h}_n$ due distinte basi dello spazio vettoriale V in esame.
Il vettore $\underline{v} \in V$ si rappresenta rispetto alla prima base come combinazione lineare di coefficienti $x_1, x_2, \ldots, x_n$, vale a dire

$$\underline{v} = x_1\underline{g}_1 + x_2\underline{g}_2 + \ldots + x_n\underline{g}_n \tag{1}$$

(ossia le coordinate di $\underline{v}$ rispetto alla base $\underline{g}_1, \underline{g}_2, \ldots, \underline{g}_n$ sono $x_1, x_2, \ldots, x_n$) e si rappresenta rispetto alla seconda base come combinazione lineare di coefficienti $y_1, y_2, \ldots, y_n$, vale a dire

$$\underline{v} = y_1\underline{h}_1 + y_2\underline{h}_2 + \ldots + y_n\underline{h}_n \tag{2}$$

(ossia le coordinate di $\underline{v}$ rispetto alla base $\underline{h}_1, \underline{h}_2, \ldots, \underline{h}_n$ sono $y_1, y_2, \ldots, y_n$).
Per determinare il legame intercorrente tra le coordinate $x_1, x_2, \ldots, x_n$ e le coordinate $y_1, y_2, \ldots, y_n$ si ricorre alla proprietà di unicità di rappresentazione del Teorema 1.1: si può affermare che gli elementi della prima base si rappresentano in modo unico come combinazione lineare degli elementi della seconda base, ossia per ogni $k = 1, \ldots, n$ si ha

$$\underline{g}_k = \sum_{i=1}^{n} m_{ik}\underline{h}_i, \tag{3}$$

ove l'indice k di m_{ik} stà ad indicare che gli m_{ik} sono i coefficienti della combinazione lineare relativa al vettore $\underline{g}_k$.
Si può quindi definire una matrice

$$M = [m_{ik}]_{i=1,\ldots,n;k=1,\ldots,n} \in \mathcal{R}^{n\times n},$$

tale che la sua $k-$sima colonna esprime le coordinate del vettore $\underline{g}_k$ ($k-$simo vettore della prima base) rispetto alla seconda base.
Ora, sostituendo la (3) nella (1) si ha

$$\begin{aligned}
\underline{v} = \sum_{k=1}^{n} x_k\underline{g}_k \quad &=_{(3)} \quad \sum_{k=1}^{n} x_k \sum_{i=1}^{n} m_{ik}\underline{h}_i \\
&= \quad \sum_{i=1}^{n} \left(\sum_{k=1}^{n} m_{ik}x_k \right) \underline{h}_i \\
&=_{(2)} \quad \sum_{i=1}^{n} y_i\underline{h}_i.
\end{aligned}$$

Quindi, sempre per l'unicità di rappresentazione rispetto alla base $\underline{h}_1, \underline{h}_2, \ldots, \underline{h}_n$, deve essere

$$y_i = \sum_{k=1}^{n} m_{ik} x_k \quad i = 1, \ldots, n$$

ovvero, in forma compatta, posto $\underline{x} = [x_1, x_2, \ldots, x_n]^T$ e $\underline{y} = [y_1, y_2, \ldots, y_n]^T$, deve essere

$$\underline{y} = M\underline{x},$$

ove la matrice $M = [m_{ik}]_{i=1,\ldots,k;k=1,\ldots,n}$, definita precedentemente, viene detta matrice del cambiamento di base.

È importante sottolineare che la matrice M, così definita, è necessariamente invertibile (e quindi non singolare) in quanto le componenti di un vettore $\underline{v}$ rispetto alla prima base si possono esprimere in modo unico per mezzo di quelle della seconda base e viceversa. Infatti, ripetendo il ragionamento con ruoli rovesciati per le due basi si ottiene

$$\underline{x} = \tilde{M}\underline{y},$$

da cui

$$\underline{x} = \tilde{M}M\underline{x} \quad \text{e} \quad \underline{y} = M\tilde{M}\underline{y}$$

ossia

$$\tilde{M}M = M\tilde{M} = I,$$

ossia

$$\tilde{M} = M^{-1} \quad \text{e} \quad M = \tilde{M}^{-1}.$$

Esempio 4.1 *1. Sia $V = \mathbb{R}^4$. Si vuole determinare la matrice del cambiamento di base dalla base canonica*

$$\begin{aligned} \underline{e}_1 &= (1,0,0,0), \\ \underline{e}_2 &= (0,1,0,0), \\ \underline{e}_3 &= (0,0,1,0), \\ \underline{e}_4 &= (0,0,0,1) \end{aligned}$$

alla base a bandiera

$$\begin{aligned} \underline{v}_1 &= (1,0,0,0), \\ \underline{v}_2 &= (1,1,0,0), \\ \underline{v}_3 &= (1,1,1,0), \\ \underline{v}_4 &= (1,1,1,1) \end{aligned}$$

e viceversa. Si è già visto nell'esempio 1.3 di sezione 1.2 che ogni vettore $\underline{x} = (x_1, x_2, x_3, x_4)$ si può scrivere come combinazione lineare dei vettori $\underline{e}_1, \underline{e}_2, \underline{e}_3, \underline{e}_4$ nel modo seguente $\underline{x} = x_1\underline{e}_1 + x_2\underline{e}_2 + x_3\underline{e}_3 + x_4\underline{e}_4$ e come combinazione lineare dei vettori $\underline{v}_1, \underline{v}_2, \underline{v}_3, \underline{v}_4$ nel modo seguente $\underline{x} = (x_1 - x_2)\underline{v}_1 + (x_2 - x_3)\underline{v}_2 + (x_3 - x_4)\underline{v}_3 + x_4\underline{v}_4$.
Quindi i vettori della base canonica si scrivono rispetto alla base a bandiera

come

$$\begin{aligned}
\underline{e}_1 &= 1\underline{v}_1 + 0\underline{v}_2 + 0\underline{v}_3 + 0\underline{v}_4 \\
\underline{e}_2 &= -1\underline{v}_1 + 1\underline{v}_2 + 0\underline{v}_3 + 0\underline{v}_4 \\
\underline{e}_3 &= 0\underline{v}_1 - 1\underline{v}_2 + 1\underline{v}_3 + 0\underline{v}_4 \\
\underline{e}_4 &= 0\underline{v}_1 + 0\underline{v}_2 - 1\underline{v}_3 + 1\underline{v}_4
\end{aligned}$$

e la matrice del cambiamento di base è data da

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

Viceversa i vettori della base a bandiera si scrivono rispetto alla base canonica come

$$\begin{aligned}
\underline{v}_1 &= 1\underline{e}_1 + 0\underline{e}_2 + 0\underline{e}_3 + 0\underline{e}_4 \\
\underline{v}_2 &= 1\underline{e}_1 + 1\underline{e}_2 + 0\underline{e}_3 + 0\underline{e}_4 \\
\underline{v}_3 &= 1\underline{e}_1 + 1\underline{e}_2 + 1\underline{e}_3 + 0\underline{e}_4 \\
\underline{v}_4 &= 1\underline{e}_1 + 1\underline{e}_2 + 1\underline{e}_3 + 1\underline{e}_4
\end{aligned}$$

e la matrice del cambiamento di base è data da

$$\tilde{M} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Si può verificare che vale $\tilde{M}M = M\tilde{M} = I$. *Infatti*

$$\tilde{M}M = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 1 \end{bmatrix} = I$$

e

$$M\tilde{M} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} = I$$

2. *Sia* $V = \mathbb{P}^2$. *Si vuole determinare la matrice del cambiamento di base dalla base canonica* $1, x, x^2$ *alla base* $1, (x - x_0), (x - x_0)(x - x_1)$ *(con* x_0, x_1 *assegnati) e viceversa.*
Ora, ogni polinomio $p_2(x) = a_0 + a_1x + a_2x^2$ *si può scrivere rispetto alla base canonica come* $p_2(x) = a_0 \cdot 1 + a_1 \cdot x + a_2 \cdot x^2$ *(ossia ha coordinate* a_0, a_1, a_2*) e rispetto alla base* $1, (x - x_0), (x - x_0)(x - x_1)$ *come* $p_2(x) = (a_0 + a_1x_0 + a_2x_0^2) \cdot 1 + (a_1 + a_2(x_0 + x_1)) \cdot (x - x_0) + a_2 \cdot (x - x_0)(x - x_1)$

(ossia ha coordinate $a_0 + a_1x_0 + a_2x_0^2, a_1 + a_2(x_0 + x_1), a_2$).
Quindi i vettori della base canonica si scrivono rispetto alla seconda base come

$$\begin{aligned} 1 &= 1 \cdot 1 + 0 \cdot (x - x_0) + 0 \cdot (x - x_0)(x - x_1) \\ x &= x_0 \cdot 1 + 1 \cdot (x - x_0) + 0 \cdot (x - x_0)(x - x_1) \\ x^2 &= x_0^2 \cdot 1 + (x_0 + x_1)) \cdot (x - x_0) + 1 \cdot (x - x_0)(x - x_1) \end{aligned}$$

e la matrice del cambiamento di base è data da

$$M = \begin{bmatrix} 1 & x_0 & x_0^2 \\ 0 & 1 & (x_0 + x_1) \\ 0 & 0 & 1 \end{bmatrix}$$

Viceversa i vettori della seconda base si scrivono rispetto alla base canonica come

$$\begin{aligned} 1 &= 1 \cdot 1 + 0 \cdot x + 0 \cdot x^2 \\ (x - x_0) &= -x_0 \cdot 1 + 1 \cdot x + 0 \cdot x^2 \\ (x - x_0)(x - x_1) &= x_0x_1 \cdot 1 - (x_0 + x_1) \cdot x + 1 \cdot x^2 \end{aligned}$$

e la matrice del cambiamento di base è data da

$$\tilde{M} = \begin{bmatrix} 1 & -x_0 & x_0x_1 \\ 0 & 1 & -(x_0 + x_1) \\ 0 & 0 & 1 \end{bmatrix}$$

Si può verificare che vale $\tilde{M}M = M\tilde{M} = I$.

4.2 Definizione e calcolo di autovalori e autovettori

Si è visto nella sezione 3.2 come le applicazioni dello spazio vettoriale $\mathbb{R}^n$ in sé stesso si rappresentino tramite matrice quadrate $n \times n$. Analoga rappresentazione è possibile anche quando si consideri un generico spazio vettoriale V.
Il problema, di notevole rilevanza, che è naturale porsi è il seguente:

si può scegliere in V una base in modo che l'applicazione lineare da V in sé venga rappresentata in "forma particolarmente semplice"?

Chiaramente occorre innanzitutto precisare meglio cosa si intenda con "forma particolarmente semplice".
L'idea è quella di andare a cercare quei vettori non nulli dello spazio vettoriale V che vengono semplicemente dilatati (o contratti) dall'applicazione lineare F, ossia tali che

$$F(\underline{v}) = \lambda \underline{v}.$$

Più precisamente, si introduce la seguente definizione.

Definizione 4.1 ***Autovettore e autovalore***
Si dice autovettore dell'applicazione lineare F dello spazio vettoriale V in sé stesso ogni vettore $\underline{v} \in V$ non nullo tale che $F(\underline{v}) = \lambda \underline{v}$ e il relativo coefficiente λ di dilatazione (o contrazione) si dice autovalore relativo all'autovettore $\underline{v}$.

Ora, sia $A \in \mathbb{R}^{n\times n}$ la matrice rappresentativa dell'applicazione lineare F dello spazio vettoriale V in sé stesso rispetto ad una base fissata.
La ricerca di un vettore $\underline{v}$ tale che $F(\underline{v}) = \lambda\underline{v}$ si traduce nella risoluzione del sistema lineare

$$A\underline{x} = \lambda\underline{x} \tag{4}$$

ossia

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \dots a_{1n}x_n = \lambda x_1 \\ a_{21}x_1 + a_{22}x_2 + \dots a_{2n}x_n = \lambda x_2 \\ \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \dots a_{nn}x_n = \lambda x_n \end{cases}$$

ove $\underline{x}$ è il vettore delle coordinate del vettore $\underline{v}$ rispetto alla base fissata, ovvero rispetto a cui l'applicazione lineare F si rappresenta con la matrice A.
Ora, la (4) equivale a cercare le soluzioni non banali (ossia diverse dal vettore nullo) del sistema lineare

$$(A - \lambda I)\underline{x} = \underline{0} \text{ (con } I \text{ matrice identica)},$$

per quei valori di λ tali che

$$\det(A - \lambda I) = 0. \tag{5}$$

Infatti se la matrice $A-\lambda I$ fosse invertibile (ossia $\det(A-\lambda I) \neq 0$) allora l'unica soluzione possibile sarebbe la soluzione banale

$$\underline{x} = (A - \lambda I)^{-1}\underline{0} = \underline{0}.$$

Ora, tenuto conto dello sviluppo del determinante, si ha che

$$\det(A - \lambda I) = \begin{bmatrix} a_{11} - \lambda & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} - \lambda & \dots & a_{2n} \\ \vdots & \vdots & \dots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} - \lambda \end{bmatrix}$$

è un polinomio di grado n in λ, detto polinomio caratteristico della matrice A.

Esempio 4.2

- *Caso $n = 2$: si ha*

$$\begin{aligned} \det(A - \lambda I) &= \begin{bmatrix} a_{11} - \lambda & a_{12} \\ a_{21} & a_{22} - \lambda \end{bmatrix} = (a_{11} - \lambda)(a_{22} - \lambda) - a_{12}a_{21} \\ &= \lambda^2 - (a_{11} + a_{22})\lambda + a_{11}a_{22} - a_{12}a_{21} \end{aligned}$$

  che è un polinomio di grado 2 in λ.

- *Caso $n = 3$: usando la formula di Laplace con sviluppo rispetto alla prima riga, si ha*

$$\begin{aligned}
\det(A-\lambda I) &= \begin{bmatrix} a_{11}-\lambda & a_{12} & a_{13} \\ a_{21} & a_{22}-\lambda & a_{23} \\ a_{31} & a_{32} & a_{33}-\lambda \end{bmatrix} \\
&= (-)^{1+1}(a_{11}-\lambda)\det\begin{bmatrix} a_{22}-\lambda & a_{23} \\ a_{32} & a_{33}-\lambda \end{bmatrix} \\
&\quad +(-)^{1+2}a_{12}\det\begin{bmatrix} a_{21} & a_{23} \\ a_{31} & a_{33}-\lambda \end{bmatrix} \\
&\quad +(-)^{1+3}a_{13}\det\begin{bmatrix} a_{21} & a_{22}-\lambda \\ a_{31} & a_{32} \end{bmatrix} \\
&= (a_{11}-\lambda)((a_{22}-\lambda)(a_{33}-\lambda)-a_{23}a_{32}) \\
&\quad -a_{12}(a_{21}(a_{33}-\lambda)-a_{23}a_{31})+a_{13}(a_{21}a_{32}-(a_{22}-\lambda)a_{31})
\end{aligned}$$

che è un polinomio di grado 3 *in* λ.

Si noti che il contributo nel grado massimo e in quello successivo è dato unicamente dal primo minore (e questa osservazione è generalizzabile al caso di dimensione n generica).

Ora, dette $\lambda_1, \ldots, \lambda_n$ le n radici di tale polinomio (eventualmente contate tenendo conto della loro molteplicità), $\lambda_1, \ldots, \lambda_n$ sono gli n autovalori della matrice A.
In definitiva, il problema del calcolo degli autovalori di una assegnata matrice $A \in \mathbb{R}^{n\times n}$ è ricondotto a quello del calcolo delle radici di un'opportuno polinomio di grado n, detto polinomio caratteristico della matrice A.
Si tenga presente che tali radici potranno essere eventualmente numeri complessi anche se il polinomio ha coefficienti reali (si pensi al caso delle radici di un'equazione di secondo grado a coefficienti reali, ma con discriminate negativo).
Una volta calcolati gli autovalori della matrice A, si calcolano i corrispondenti autovettori risolvendo dei **sistemi lineari omogenei** del tipo $(A-\lambda I)\underline{x} = \underline{0}$ con λ fissato.
Chiaramente, trattandosi di un sistema lineare omogeneo con matrice singolare per definizione, si avranno infiniti vettori soluzione. Del resto è evidente che gli autovettori relativi ad un fissato autovalore sono determinati a meno di un coefficiente di proporzionalità. In effetti, se $\underline{x}$ è tale che $A\underline{x} = \lambda\underline{x}$ con λ fissato, allora $\underline{y} = \alpha\underline{x}$ è tale che

$$A\underline{y} = A(\alpha\underline{x}) = \alpha\lambda\underline{x} = \lambda\underline{y},$$

ossia $\underline{y} = \alpha\underline{x}$ è autovettore relativamente all'autovalore λ.

Esempio 4.3 *Sia*

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

Passo 1*: si calcolano gli autovalori, cercando le radici del polinomio caratteristico di secondo grado*

$$p_2(\lambda) = \det(A-\lambda I) = \begin{bmatrix} 1-\lambda & 1 \\ 2 & 1-\lambda \end{bmatrix} = (1-\lambda)^2 - 2 = \lambda^2 - 2\lambda - 1 = 0$$

Vale che $p_2(\lambda) = (\lambda - \lambda_1)(\lambda - \lambda_2)$ *con* $\lambda_1 = 1 + \sqrt{2}$ *e* $\lambda_2 = 1 - \sqrt{2}$*, ossia le due radici sono* $\lambda_1 = 1 + \sqrt{2}$ *e* $\lambda_2 = 1 - \sqrt{2}$*.*
Passo 2.a*: si calcolano gli autovettori relativamente all'autovalore* $\lambda_1 = 1+\sqrt{2}$*, risolvendo il sistema lineare omogeneo* $(A - \lambda_1 I)\underline{x} = \underline{0}$*, ossia*

$$\begin{cases} (1 - \lambda_1)x_1 + x_2 = 0 \\ 2x_1 + (1 - \lambda_1)x_2 = 0. \end{cases}$$

Poiché $rango(A - \lambda_1 I) = 1$ *(*$A - \lambda_1 I$ *non è la matrice nulla, quindi il rango è* ≥ 1 *e non può avere rango* 2 *in quanto vale* $\det(A - \lambda_1 I) = 0$*), il sistema ha infinite soluzioni del tipo*

$$x_2^* = -(1 - \lambda_1)x_1^* = \sqrt{2}x_1^*,\ x_1^* \in \mathbb{R},$$

ossia ogni vettore del tipo $[x_1^*, \sqrt{2}x_1^*]^T, x_1^* \in \mathbb{R}$ *è autovettore relativamente all'autovalore* $\lambda_1 = 1 + \sqrt{2}$*.*
Ad esempio si fissa $x_1^* = 1$ *e si considera* $\underline{x}^* = [1, \sqrt{2}]^T$ *come rappresentante di questo insieme infinito di autovettori.*
Si può facilmente verificare che

$$A\underline{x}^* = \begin{bmatrix} 1 & 1 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ \sqrt{2} \end{bmatrix} = \begin{bmatrix} 1 + \sqrt{2} \\ 2 + \sqrt{2} \end{bmatrix} = (1 + \sqrt{2}) \begin{bmatrix} 1 \\ \sqrt{2} \end{bmatrix} = \lambda_1 \underline{x}^*$$

Passo 2.b*: si calcolano gli autovettori relativamente all'autovalore* $\lambda_1 = 1-\sqrt{2}$*, risolvendo il sistema lineare* $(A - \lambda_2 I)\underline{x} = \underline{0}$*, ossia*

$$\begin{cases} (1 - \lambda_2)x_1 + x_2 = 0 \\ 2x_1 + (1 - \lambda_2)x_2 = 0. \end{cases}$$

Poiché $rango(A - \lambda_2 I) = 1$ *(came nel caso precedente* $A - \lambda_2 I$ *non è la matrice nulla, quindi il rango è* ≥ 1 *e non può avere rango* 2 *in quanto vale* $\det(A - \lambda_2 I) = 0$*), il sistema omogeneo ha infinite soluzioni del tipo*

$$x_2^* = -(1 - \lambda_2)x_1^* = -\sqrt{2}x_1^*,\ x_1^* \in \mathbb{R},$$

ossia ogni vettore del tipo $[x_1^*, -\sqrt{2}x_1^*]^T, x_1^* \in \mathbb{R}$ *è autovettore relativamente all'autovalore* λ_2*.*
Ad esempio si fissa $x_1^* = 1$ *e si considera* $\underline{x}^* = [1, -\sqrt{2}]^T$ *come rappresentante di questo insieme infinito di autovettori.*
Si può facilmente verificare che

$$A\underline{x}^* = \begin{bmatrix} 1 & 1 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -\sqrt{2} \end{bmatrix} = \begin{bmatrix} 1 - \sqrt{2} \\ 2 - \sqrt{2} \end{bmatrix} = (1 - \sqrt{2}) \begin{bmatrix} 1 \\ -\sqrt{2} \end{bmatrix} = \lambda_2 \underline{x}^*$$

4.3 Matrici diagonalizzabili e non

Supponiamo che esistano in V n autovettori $\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_n$ linearmente indipendenti con $n = \dim V$, ossia tali che $\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_n$ siano una base per V.
Rispetto a tale base i vettori $\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_n$ sono rappresentati dai vettori colonna di coordinate

$$\begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \ldots \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 0 \\ 1 \end{bmatrix},$$

in quanto le coordinate dei vettori di una base rispetto alla base stessa sono ovviamente i vettori della base canonica.
Inoltre, indicati con $\lambda_1, \lambda_2, \ldots, \lambda_n$ i relativi autovalori, allora i vettori $F(\underline{v}_1)$, $F(\underline{v}_2)$, $\ldots, F(\underline{v}_n)$, vale a dire i vettori $\lambda_1\underline{v}_1$, $\lambda_2\underline{v}_2, \ldots, \lambda_n\underline{v}_n$ (questo poiché i vettori $\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_n$ sono autovettori) sono espressi dai vettori colonna di coordinate

$$\begin{bmatrix} \lambda_1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ \lambda_2 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \ldots \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 0 \\ \lambda_n \end{bmatrix}.$$

Quindi, ricordando quanto visto nella sezione 3.2, si ha che rispetto alla base fissata, l'applicazione lineare F si rappresenta mediante la matrice

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \ldots & \ldots & 0 \\ 0 & \lambda_2 & 0 & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ldots & 0 & \lambda_{n-1} & 0 \\ 0 & \ldots & \ldots & 0 & \lambda_n \end{bmatrix}$$

che è una matrice diagonale ($\Lambda_{ij} = 0$ per ogni $i \neq j$), ossia la matrice di tipo più semplice possibile.
Infatti, indicato con $\underline{x} = [x_1, x_2, \ldots, x_n]^T$ il vettore colonna delle coordinate di un assegnato vettore $\underline{v} \in V$ rispetto alla base $\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_n$ degli autovettori e indicato con $\underline{y} = [y_1, y_2, \ldots, y_n]^T$ il vettore colonna delle coordinate del vettore trasformato $F(\underline{v})$, sempre rispetto a tale base, si ha che

$$\underline{y} = \Lambda\underline{x} \tag{6}$$

ovvero, per ogni $i = 1, \ldots, n$

$$y_i = \lambda_i x_i$$

ovvero le equazioni sono tutte disaccoppiate fra loro.

Infine, si supponga di considerare un'altra base dello spazio vettoriale V rispetto alla quale l'applicazione lineare F si rappresenta come

$$\underline{w} = A\underline{z}. \tag{7}$$

La domanda che è naturale porsi è la seguente:

qual è la relazione che sussiste tra le due matrici A e Λ, rappresentative della medesima applicazione lineare, ma rispetto a due basi diverse?

Dalla relazione che governa il cambiamento di base, vista nella sezione 4.1, si ricava

$$\underline{z} = M\underline{x} \quad \text{e} \quad \underline{w} = M\underline{y},$$

e sostituendo nell'equazione (7) si ha

$$M\underline{y} = AM\underline{x}$$

da cui, moltiplicando a sinistra per M^{-1} (le matrici non si dividono e il prodotto di matrici in genere non commuta!!!), ove M è invertibile come osservato precedentemente in sezione 4.1, si ottiene

$$\underline{y} = M^{-1}AM\underline{x}$$

ovvero, confrontando con l'equazione (6), si ha la relazione

$$\Lambda = M^{-1}AM,$$

ovvero le due matrici Λ e A sono simili in accordo alla seguente definizione.

Definizione 4.2 ***Matrici simili***
Siano $A, B \in \mathbb{R}^{n\times n}$. Si dice che A e B sono simili se esiste una matrice $S \in \mathbb{R}^{n\times n}$ invertibile tale che

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

La trasformazione che associa la matrice A alla matrice B viene detta trasformazione per similitudine.

Di particolare importanza è la seguente definizione.

Definizione 4.3 ***Matrice diagonalizzabile***
Sia $A \in \mathbb{R}^{n\times n}$. Si dice che A è diagonalizzabile se A è simile ad una matrice diagonale.

Ora l'analisi riportata sopra motiva una parte del seguente risultato.

Teorema 4.1 *Sia $A \in \mathbb{R}^{n\times n}$. La matrice A è diagonalizzabile se e solo se la matrice A ha n autovettori linearmente indipendenti. Inoltre, le colonne della matrice S, per cui $S^{-1}AS$ è diagonale, sono tali autovettori della matrice A.*

La domanda diviene quindi la seguente:

sotto quali condizioni una matrice A possiede n autovettori linearmente indipendenti?

Vale il seguente risultato.

Teorema 4.2 *Siano i vettori $\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_r$ autovettori relativi ad autovalori tutti distinti fra loro. Allora i vettori $\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_r$ sono linearmente indipendenti.*

Corollario 4.1 *Sia $A \in \mathbb{R}^{n\times n}$. Se la matrice A ha n autovalori distinti, allora la matrice A ha n autovettori linearmente indipendenti e quindi A è diagonalizzabile.*

Tuttavia, è importante tenere presente che è possibile avere n autovettori linearmente indipendenti anche quando non si hanno n autovalori distinti. Basta pensare alla matrice identica

$$I = \begin{bmatrix} 1 & 0 & \ldots & \ldots & \ldots & 0 \\ 0 & 1 & 0 & \ldots & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \ldots & \ldots & 0 & 1 & 0 \\ 0 & \ldots & \ldots & \ldots & 0 & 1 \end{bmatrix}$$

che ha tutti gli autovalori uguali a 1 e per la quale gli n vettori della base canonica sono autovettori relativamente a tale autovalore.

Si tenga infine presente che, a differenza di quanto in genere accade, la proprietà di diagonalizzabilità, pur essendo una proprietà buona, è una proprietà verificata dalla maggioranza delle matrici.

Esempio 4.4

- *La matrice*
$$A = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 2 & 1 \\ 0 & 0 & 3 \end{bmatrix}$$
è diagonalizzabile.
La matrice ha tre autovalori distinti $\lambda_1 = 1$, $\lambda_2 = 2$, $\lambda_3 = 3$ e quindi risulta diagonalizzabile (ad autovalori distinti corrispondono autovettori linearmente indipendenti fra loro).
Più in dettaglio, la matrice ha autovettori del tipo:
  - $[x_1\ 0\ 0]^T$, *con* $x_1 \in \mathbb{R}$ *relativamente all'autovalore* $\lambda_1 = 1$, *ad esempio il vettore* $[1\ 0\ 0]^T$;
  - $[x_1\ x_1\ 0]^T$, *con* $x_1 \in \mathbb{R}$ *relativamente all'autovalore* $\lambda_2 = 2$, *ad esempio il vettore* $[1\ 1\ 0]^T$;
  - $[x_1\ x_1\ x_1]^T$, *con* $x_1 \in \mathbb{R}$ *relativamente all'autovalore* $\lambda_3 = 3$, *ad esempio il vettore* $[1\ 1\ 1]^T$.

- *La matrice*
$$A = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$
è diagonalizzabile.
L'affermazione è ovvia, in quanto la matrice è già in forma diagonale.
Di più, essa ha un'unico autovalore $\lambda = 2$ contato tre volte e rispetto a tale autovalore il tre vettori $\underline{e}_1, \underline{e}_2, \underline{e}_3$ sono autovettori linearmente indipendenti fra loro.

- *La matrice*
$$A = \begin{bmatrix} 2 & 1 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$
non è diagonalizzabile.
Infatti, essa ha un'unico autovalore $\lambda = 2$ contato tre volte. Rispetto a tale autovalore gli autovettori sono del tipo $x_2 = 0$ e $x_1, x_3 \in \mathbb{R}$ qualsiasi; quindi, ad esempio, i vettori $[1\ 0\ 0]^T$ e $[0\ 0\ 1]^T$ sono linearmente indipendenti, ma non è possibile trovarne un terzo, così da formare una base.

- *La matrice*
$$A = \begin{bmatrix} 2 & 1 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 2 \end{bmatrix}$$

non è diagonalizzabile.
Infatti, essa ha un'unico autovalore $\lambda = 2$ contato tre volte. Rispetto a tale autovalore gli autovettori sono del tipo $x_2 = x_3 = 0$ e $x_1 \in \mathbb{R}$ qualsiasi; quindi, ad esempio, il vettore $[1\ 0\ 0]^T$, ma non è possibile trovarne altri due, così da formare una base.

- *La matrice*
$$A = \begin{bmatrix} 2 & 1 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$
non è diagonalizzabile.
Infatti, essa ha un'autovalore $\lambda_1 = 1$ e un'autovalore $\lambda_2 = 2$ contato due volte. Rispetto all'autovalore $\lambda_1 = 1$ gli autovettori sono del tipo $x_1 = -x_2, x_3 = -x_2$ e $x_2 \in \mathbb{R}$ qualsiasi; quindi, ad esempio, il vettore $[1\ -1\ 1]^T$. Rispetto all'autovalore $\lambda_2 = 2$ gli autovettori sono del tipo $x_2 = x_3 = 0$ e $x_1 \in \mathbb{R}$ qualsiasi; quindi, ad esempio, il vettore $[1\ 0\ 0]^T$. Tuttavia, non è possibile trovarne un'altro autovettore relativamente all'autovalore λ_2, così da formare una base insieme a $[1\ -1\ 1]^T$ e $[1\ 0\ 0]^T$.

4.4 Esercizi

1. Determinare la matrice rappresentativa rispetto alla base canonica di $\mathbb{R}^3$ dell'applicazione lineare $F : \mathbb{R}^3 \to \mathbb{R}^3$ tale che $[111]^T$ è autovettore relativamente all'autovalore 1, $[120]^T$ è autovettore relativamente all'autovalore 0, e $F([101]^T) = [201]^T$

2. Sono date le due matrici
$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 3 & 4 & 5 \end{bmatrix} \quad \text{e} \quad B = \begin{bmatrix} 1 & 2 & 5 \\ 0 & 2 & 0 \\ 0 & 0 & k \end{bmatrix}$$
Calcolare autovalori e autovettori della matrice A.
Determinare i valori del parametro k per cui la matrice A è simile alla matrice B.

3. È data la matrice
$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 3 & 4 & 5 \end{bmatrix}$$
Calcolare autovalori e autovettori della matrice A e dire se è diagonalizzabile.

4. È data la matrice
$$A = \begin{bmatrix} 0 & -2 & 0 & 0 \\ 1 & 3 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -2 & 3 \end{bmatrix}$$
Calcolare autovalori e autovettori della matrice A e dire se è diagonalizzabile.

5 Un'applicazione: le matrici di rotazione

5.1 Rotazioni nel piano di un angolo ϑ

Si vuole considerare il caso della rotazione nel piano di un vettore di $\mathbb{R}^2$ di un angolo ϑ assegnato.
Chiaramente si tratta di un'applicazione lineare (si veda la definizione 3.3), in quanto la rotazione nel piano di un'angolo ϑ del vettore somma di due vettori assegnati è equivalente al vettore somma dei due vettori precedentemente ruotati. Inoltre, il fatto che il vettore venga dilatato (o contratto) non modifica l'angolo di rotazione e quindi è del tutto indifferente farlo prima o dopo.
In definitiva, essendo la rotazione nel piano di un angolo ϑ un'applicazione lineare, essa può essere rappresentata da una matrice, ottenuta nel modo seguente (si faccia riferimento sempre alla sezione 3.2).
Si considera la base canonica

$$\begin{array}{rcl} \underline{e}_1 = [1,0]^T & \rightarrow & F(\underline{e}_1) = [\cos\vartheta, \sin\vartheta]^T \\ \underline{e}_2 = [0,1]^T & \rightarrow & F(\underline{e}_2) = [\cos(\pi/2+\vartheta), \sin(\pi/2+\vartheta)]^T \\ \text{vettori della base} & & \text{immagini dei vettori della base} \end{array}$$

Ma, poiché vale $\cos(\alpha+\beta) = \cos\alpha\cos\beta - \sin\alpha\sin\beta$ e $\sin(\alpha+\beta) = \sin\alpha\cos\beta + \cos\alpha\sin\beta$, si ha $\cos(\pi/2+\beta) = \cos\pi/2\cos\vartheta - \sin\pi/2\sin\vartheta = -\sin\vartheta$ e $\sin(\pi/2+\beta) = \sin\pi/2\cos\vartheta + \cos\pi/2\sin\vartheta = \cos\vartheta$.
In definitiva, l'applicazione lineare da $\mathbb{R}^2$ a $\mathbb{R}^2$ corrispondente alla rotazione di un angolo ϑ è rappresentata dalla matrice

$$A = A_\vartheta = [F(\underline{e}_1), F(\underline{e}_2)] = \begin{bmatrix} \cos\vartheta & -\sin\vartheta \\ \sin\vartheta & \cos\vartheta \end{bmatrix}$$

Per verifica, si consideri un vettore generico $\underline{v} = [\rho\cos\alpha, \rho\sin\alpha]^T,\ \rho \geq 0$, allora

$$\begin{aligned} F(\underline{v}) &= A\underline{v} = \begin{bmatrix} \cos\vartheta & -\sin\vartheta \\ \sin\vartheta & \cos\vartheta \end{bmatrix} \begin{bmatrix} \rho\cos\alpha \\ \rho\sin\alpha \end{bmatrix} \\ &= [\rho\cos\alpha\cos\vartheta - \rho\sin\alpha\sin\vartheta, \rho\cos\alpha\sin\vartheta + \rho\sin\alpha\cos\vartheta]^T \\ &= [\rho\cos(\alpha+\vartheta), \rho\sin(\alpha+\vartheta)]^T \end{aligned}$$

È interessante notare che la matrice A è ortogonale, ossia vale la proprietà $A^TA = AA^T = I$. Infatti è

$$A^T = \begin{bmatrix} \cos\vartheta & \sin\vartheta \\ -\sin\vartheta & \cos\vartheta \end{bmatrix}$$

e quindi

$$\begin{aligned} A^TA &= \begin{bmatrix} \cos\vartheta & -\sin\vartheta \\ \sin\vartheta & \cos\vartheta \end{bmatrix} \begin{bmatrix} \cos\vartheta & \sin\vartheta \\ -\sin\vartheta & \cos\vartheta \end{bmatrix} \\ &= \begin{bmatrix} \cos^2\vartheta + \sin^2\vartheta & \cos\vartheta\sin\vartheta - \cos\vartheta\sin\vartheta \\ \cos\vartheta\sin\vartheta - \cos\vartheta\sin\vartheta & \sin^2\vartheta + \cos^2\vartheta \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \end{aligned}$$

e

$$\begin{aligned} AA^T &= \begin{bmatrix} \cos\vartheta & \sin\vartheta \\ -\sin\vartheta & \cos\vartheta \end{bmatrix} \begin{bmatrix} \cos\vartheta & -\sin\vartheta \\ \sin\vartheta & \cos\vartheta \end{bmatrix} \\ &= \begin{bmatrix} \cos^2\vartheta + \sin^2\vartheta & -\cos\vartheta\sin\vartheta + \cos\vartheta\sin\vartheta \\ -\cos\vartheta\sin\vartheta + \cos\vartheta\sin\vartheta & \sin^2\vartheta + \cos^2\vartheta \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \end{aligned}$$

Si procede ora al calcolo degli autovalori e autovettori della matrice A.
Passo 1: Calcolo degli autovalori. Vale

$$\begin{aligned}\det(A-\lambda I) &= \det\begin{bmatrix}\cos\vartheta-\lambda & -\sin\vartheta\\ \sin\vartheta & \cos\vartheta-\lambda\end{bmatrix}\\ &= (\cos\vartheta-\lambda)^2+\sin^2\vartheta=\cos^2\vartheta+\lambda^2-2\cos\vartheta\lambda+\sin^2\vartheta\\ &= \lambda^2-2\cos\vartheta\lambda+1\end{aligned}$$

da cui

$$\begin{aligned}\lambda &= \frac{2\cos\vartheta\pm\sqrt{4\cos^2\vartheta-4}}{2}=\cos\vartheta\pm\sqrt{-\sin^2\vartheta}=\cos\vartheta\pm i|\sin\vartheta|\\ &= \cos\vartheta\pm i\sin\vartheta=e^{\pm i\vartheta}\end{aligned}$$

Si noti che $|\lambda|=1$. Questo fatto non è un caso: vale la proprietà che autovalori di matrici ortogonali (e più in generale unitarie) sono di modulo unitario.
Passo 2.a: sia $\lambda_1=\cos\vartheta+i\sin\vartheta$, si considera il sistema omogeneo

$$A\underline{z}_1=\lambda_1\underline{z}_1$$

ossia

$$\begin{bmatrix}\cos\vartheta-(\cos\vartheta+i\sin\vartheta) & -\sin\vartheta\\ \sin\vartheta & \cos\vartheta-(\cos\vartheta+i\sin\vartheta)\end{bmatrix}\underline{z}_1=\underline{0},$$

ossia

$$\begin{bmatrix}-i\sin\vartheta & -\sin\vartheta\\ \sin\vartheta & -i\sin\vartheta\end{bmatrix}\begin{bmatrix}x_1\\ y_1\end{bmatrix}=\underline{0},\quad\text{con}\quad\underline{z}_1=\begin{bmatrix}x_1\\ y_1\end{bmatrix}.$$

Ora, $\text{rango}(A-\lambda_1 I)=1$ se $\sin\vartheta\neq 0$, ossia $\vartheta\neq 0;\pi$.
Quindi per $\vartheta\neq 0;\pi$, si considera

$$-i\sin\vartheta x_1-\sin\vartheta y_1=0,$$

ossia

$$y_1=-ix_1;$$

mentre l'altra equazione risulterà automaticamente verificata.
Quindi, posto ad esempio $x_1=1$ si ha $y_1=-i$ e, normalizzando in norma 2 (si veda sezione 6.1), ($\|[1,-i]^T\|_2=\sqrt{2}$) si ottiene

$$\underline{z}_1=\frac{1}{\sqrt{2}}\begin{bmatrix}1\\ -i\end{bmatrix}$$

Passo 2.b: sia $\lambda_2=\cos\vartheta-i\sin\vartheta$, si considera il sistema

$$A\underline{z}_2=\lambda_2\underline{z}_2$$

ossia

$$\begin{bmatrix}\cos\vartheta-(\cos\vartheta-i\sin\vartheta) & -\sin\vartheta\\ \sin\vartheta & \cos\vartheta-(\cos\vartheta-i\sin\vartheta)\end{bmatrix}\underline{z}_2=\underline{0},$$

ossia

$$\begin{bmatrix}i\sin\vartheta & -\sin\vartheta\\ \sin\vartheta & i\sin\vartheta\end{bmatrix}\begin{bmatrix}x_2\\ y_2\end{bmatrix}=\underline{0},\quad\text{con}\quad\underline{z}_2=\begin{bmatrix}x_2\\ y_2\end{bmatrix}.$$

Ora, rango$(A - \lambda_2 I) = 1$ se $\sin\vartheta \neq 0$, ossia $\vartheta \neq 0; \pi$.
Quindi per $\vartheta \neq 0; \pi$, si considera

$$i \sin\vartheta x_2 - \sin\vartheta y_2 = 0,$$

ossia

$$y_2 = ix_2;$$

mentre l'altra equazione risulterà automaticamente verificata.
Quindi, posto ad esempio $x_2 = 1$ si ha $y_2 = i$ e normalizzando in norma 2 (si veda sezione 6.1) ($\|[1, i]^T\|_2 = \sqrt{2}$) si ottiene

$$\underline{z}_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ i \end{bmatrix}$$

Pertanto, per se $\vartheta \neq 0; \pi$ la matrice è sicuramente diagonalizzabile , ossia vale

$$A = S\Lambda S^{-}1$$

con

$$\begin{aligned} \Lambda &= \text{diag}(\lambda_1, \lambda_2) \\ S &= [\underline{z}_1, \underline{z}_2] = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -i & i \end{bmatrix}. \end{aligned}$$

Si può verificare la seguente proprietà: le matrici normali (ossia tali che $AA^H = A^H A$) si diagonalizzano tramite matrici unitarie. Infatti, la matrice $S = [\underline{z}_1, \underline{z}_2]$ è unitaria, ossia $SS^H = S^H S = I$, o meglio $\underline{z}_1$ è ortogonale, anzi ortonormale, a $\underline{z}_2$, come qui di seguito verificato.

$$S^H = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & i \\ 1 & -i \end{bmatrix}$$

$$S^H S = \frac{1}{2} \begin{bmatrix} 1 & i \\ 1 & -i \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -i & i \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 1 - i^2 & 1 + i^2 \\ 1 + i^2 & 1 - i^2 \end{bmatrix} \underbrace{=}_{i^2 = -1} \frac{1}{2} \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} = I$$

$$SS^H = \frac{1}{2} \begin{bmatrix} 1 & 1 \\ -i & i \end{bmatrix} \begin{bmatrix} 1 & i \\ 1 & -i \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 2 & i - i \\ -i + i & -i^2 - i^2 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} = I$$

Quindi si può affermare che

$$S^{-1} = S^H$$

e

$$A = S\Lambda S^H = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -i & i \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & i \\ 1 & -i \end{bmatrix}$$

Si tenga presente che i vettori $\underline{z}_1, \underline{z}_2$ formano una base ortonormale per $\mathbb{R}^2$: sono in numero di 2 e sono linearmente indipendenti, essendo

$$\det S = (\frac{1}{\sqrt{2}})^2 \begin{vmatrix} 1 & 1 \\ -i & i \end{vmatrix} = \frac{1}{2}(i + i) = i \neq 0.$$

Casi particolari: $\boldsymbol{\vartheta = 0, \pi}$ **(**$\boldsymbol{\sin\vartheta = 0}$**)**.
Per $\vartheta = 0$ si ha

$$A = A_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

matrice diagonale con autovalori coincidenti $\lambda_1 = \lambda_2 = 1$ (autovettori $\underline{e}_1, \underline{e}_2$).
Per $\vartheta = \pi$ si ha

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

matrice diagonale con autovalori coincidenti $\lambda_{1,2} = -1$ (autovettori $\underline{e}_1, \underline{e}_2$).
In entrambi i casi le matrici sono diagonalizzabili, in quanto già diagonali.

5.2 Rotazioni nello spazio di un angolo ϑ intorno all'asse z

Si vuole considerare il caso della rotazione nello spazio intorno all'asse z di un vettore di $\mathbb{R}^3$ di un angolo ϑ assegnato.
Come nel caso precedente, si tratta chiaramente di un'applicazione lineare, quindi si procede analogamente.
Si considera la base canonica

$$\begin{array}{rcl} \underline{e}_1 = [1,0,0]^T & \rightarrow & F(\underline{e}_1) = [\cos\vartheta, \sin\vartheta, 0]^T \\ \underline{e}_2 = [0,1,0]^T & \rightarrow & F(\underline{e}_2) = [-\sin\vartheta, \cos\vartheta, 0]^T \\ \underline{e}_3 = [0,0,1]^T & \rightarrow & F(\underline{e}_1) = [0,0,1]^T \\ \text{vettori della base} & & \text{immagini dei vettori della base} \end{array}$$

In definitiva, tale applicazione lineare da $\mathbb{R}^3$ a $\mathbb{R}^3$ è rappresentata dalla matrice

$$A = A_\vartheta = [F(\underline{e}_1), F(\underline{e}_2), F(\underline{e}_3)] = \begin{bmatrix} \cos\vartheta & -\sin\vartheta & 0 \\ \sin\vartheta & \cos\vartheta & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

È interessante notare che la matrice A è ortogonale, ossia vale $A^T A = AA^T = I$.
Infatti è

$$A^T = \begin{bmatrix} \cos\vartheta & \sin\vartheta & 0 \\ -\sin\vartheta & \cos\vartheta & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

e quindi

$$\begin{aligned} A^T A &= \begin{bmatrix} \cos\vartheta & -\sin\vartheta & 0 \\ \sin\vartheta & \cos\vartheta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\vartheta & \sin\vartheta & 0 \\ -\sin\vartheta & \cos\vartheta & 0 \\ 0 & 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} \cos^2\vartheta + \sin^2\vartheta & \cos\vartheta\sin\vartheta - \cos\vartheta\sin\vartheta & 0 \\ \cos\vartheta\sin\vartheta - \cos\vartheta\sin\vartheta & \sin^2\vartheta + \cos^2\vartheta & 0 \\ 0 & 0 & 1 \end{bmatrix} = I \end{aligned}$$

e

$$\begin{aligned} AA^T &= \begin{bmatrix} \cos\vartheta & \sin\vartheta & 0 \\ -\sin\vartheta & \cos\vartheta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\vartheta & -\sin\vartheta & 0 \\ \sin\vartheta & \cos\vartheta & 0 \\ 0 & 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} \cos^2\vartheta + \sin^2\vartheta & -\cos\vartheta\sin\vartheta + \cos\vartheta\sin\vartheta & 0 \\ -\cos\vartheta\sin\vartheta + \cos\vartheta\sin\vartheta & \sin^2\vartheta + \cos^2\vartheta & 0 \\ 0 & 0 & 1 \end{bmatrix} = I \end{aligned}$$

Si procede ora al calcolo degli autovalori e autovettori della matrice A.
Passo 1: Calcolo degli autovalori. Vale

$$\begin{aligned}\det(A-\lambda I) &= \det\begin{bmatrix}\cos\vartheta-\lambda & -\sin\vartheta & 0\\ \sin\vartheta & \cos\vartheta-\lambda & 0\\ 0 & 0 & 1-\lambda\end{bmatrix}\\ &= (1-\lambda)((\cos\vartheta-\lambda)^2+\sin^2\vartheta)\\ &= (1-\lambda)(\cos^2\vartheta+\lambda^2-2\cos\vartheta\lambda+\sin^2\vartheta)\\ &= (1-\lambda)(\lambda^2-2\cos\vartheta\lambda+1)\end{aligned}$$

da cui

$$\begin{aligned}\lambda_{1,2} &= \cos\vartheta\pm i\sin\vartheta = e^{\pm i\vartheta},\\ \lambda_3 &= 1.\end{aligned}$$

Si noti che il secondo fattore del polinomio caratteristico è il polinomio caratteristico della matrice di rotazione nel piano di sezione 5.1.
Passo 2.a: sia $\lambda_1=\cos\vartheta+i\sin\vartheta$, si considera il sistema

$$A\underline{w}_1=\lambda_1\underline{w}_1$$

ossia

$$\begin{bmatrix}\cos\vartheta-(\cos\vartheta+i\sin\vartheta) & -\sin\vartheta & 0\\ \sin\vartheta & \cos\vartheta-(\cos\vartheta+i\sin\vartheta) & 0\\ 0 & 0 & 1-(\cos\vartheta+i\sin\vartheta)\end{bmatrix}\underline{w}_1=\underline{0},$$

ossia, posto $\underline{w}_1=[x_1,y_1,z_1]^T$,

$$\begin{bmatrix}-i\sin\vartheta & -\sin\vartheta & 0\\ \sin\vartheta & \boxed{\begin{matrix}-i\sin\vartheta & 0\\ 0 & 1-(\cos\vartheta+i\sin\vartheta)\end{matrix}}\\ 0 & \end{bmatrix}\begin{bmatrix}x_1\\ y_1\\ z_1\end{bmatrix}=\underline{0}.$$

Ora, rango$(A-\lambda_1 I)=2$ se $\sin\vartheta\neq 0$ e $\lambda_1\neq 1$, ossia $\vartheta\neq 0;\pi$. Per affermare questo si è considerata la sottomatrice 2×2 riquadrata.Si noti tra l'altro che ogni altra sottomatrice 2×2 è singolare.
Quindi per $\vartheta\neq 0;\pi$, si considera

$$\begin{cases} y_1 = -ix_1\\ z_1 = 0\end{cases}$$

Si noti che si stanno considerando solo le equazioni relative alla sottomatrice riquadrata: ogni altra equazione risulterà automaticamente verificata dalla soluzione di tali equazioni.
Quindi, posto ad esempio $x_1=1$ si ha $y_1=-i, z_1=0$ e normalizzando in norma 2 (si veda sezione 6.1) ($\|[1,-i,0]^T\|_2=\sqrt{2}$) si ottiene

$$\underline{w}_1=\frac{1}{\sqrt{2}}\begin{bmatrix}1\\ -i\\ 0\end{bmatrix}$$

Si tenga comunque presente che, nel caso esistesse più di una sottomatrice quadrata 2×2 con determinante diverso da zero, si andrebbe a scegliere quella più semplice (con più zeri o con coefficienti più semplici), in quanto questo semplifica il sistema lineare da risolvere.
Passo 2.b: sia $\lambda_2 = \cos\vartheta - i\sin\vartheta$, si considera il sistema

$$A\underline{w}_2 = \lambda_2\underline{w}_2$$

ossia

$$\left[\begin{array}{ccc} i\sin\vartheta & -\sin\vartheta & 0 \\ \sin\vartheta & \boxed{\begin{array}{cc} i\sin\vartheta & 0 \\ 0 & 1-(\cos\vartheta - i\sin\vartheta) \end{array}} & \\ 0 & & \end{array}\right] \left[\begin{array}{c} x_2 \\ y_2 \\ z_2 \end{array}\right] = \underline{0}, \text{ con } \underline{w}_2 = \left[\begin{array}{c} x_2 \\ y_2 z_2 \end{array}\right].$$

Ora, rango$(A - \lambda_2 I) = 2$ se $\sin\vartheta \neq 0$ e $\lambda_2 \neq 1$, ossia $\vartheta \neq 0; \pi$. Per affermare questo si è considerata la sottomatrice 2×2 riquadrata.
Quindi per $\vartheta \neq 0; \pi$, si considera

$$\left\{\begin{array}{rcl} y_2 & = & ix_2 \\ z_2 & = & 0 \end{array}\right.$$

Quindi, posto ad esempio $x_2 = 1$ si ha $y_2 = i, z_2 = 0$ e normalizzando in norma 2 (si veda sezione 6.1) ($\|[1, i, 0]^T\|_2 = \sqrt{2}$) si ottiene

$$\underline{w}_2 = \frac{1}{\sqrt{2}} \left[\begin{array}{c} 1 \\ i \\ 0 \end{array}\right]$$

Passo 2.c: sia $\lambda_3 = 1$, si considera il sistema

$$A\underline{w}_3 = \lambda_3\underline{w}_3$$

ossia

$$\left[\begin{array}{ccc} \boxed{\begin{array}{cc} \cos\vartheta - 1 & -\sin\vartheta \\ \sin\vartheta & \cos\vartheta - 1 \end{array}} & & 0 \\ & & 0 \\ 0 \quad 0 & & 0 \end{array}\right] \left[\begin{array}{c} x_3 \\ y_3 \\ z_3 \end{array}\right] = \underline{0}, \text{ con } \underline{w}_3 = \left[\begin{array}{c} x_3 \\ y_3 \\ z_3 \end{array}\right].$$

Ora, rango$(A - \lambda_3 I) = 2$ se $\vartheta \neq 0$. Per affermare questo si è considerata la sottomatrice 2×2 riquadrata.
Quindi per $\vartheta \neq 0$, si ha il sistema omogeneo di due equazioni in due incognite

$$\left\{\begin{array}{rcl} x_3(\cos\vartheta - 1) - \sin\vartheta y_3 & = & 0 \\ \sin\vartheta x_3 + (\cos\vartheta - 1)y_3 & = & 0 \end{array}\right.$$

con matrice non singolare, che ha come unica soluzione possibile la soluzione banale $x_3 = y_3 = 0$. Quindi, posto ad esempio $z_3 = 1$ si ha

$$\underline{w}_3 = \left[\begin{array}{c} 0 \\ 0 \\ 1 \end{array}\right]$$

Pertanto, per se $\vartheta \neq 0; \pi$ la matrice è sicuramente diagonalizzabile , ossia vale

$$A = S\Lambda S^{-}1$$

con

$$\begin{aligned}\Lambda &= \mathrm{diag}(\lambda_1,\lambda_2,\lambda_3)\\ S &= [\underline{w}_1,\underline{w}_2,\underline{w}_3] = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0\\ -\frac{i}{\sqrt{2}} & \frac{i}{\sqrt{2}} & 0\\ 0 & 0 & 1\end{bmatrix}.\end{aligned}$$

Anche in questo caso la matrice $S = [\underline{z}_1, \underline{z}_2]$ è unitaria, ossia $SS^H = S^H S = I$. Quindi, si ha $S^{-1} = S^H$, da cui

$$A = S\Lambda S^H$$

Si tenga presente che i vettori $\underline{w}_1, \underline{w}_2, \underline{w}_3$ formano una base ortonormale per $\mathbb{R}^3$: sono in numero di 3 e sono linearmente indipendenti, essendo $\det S = i \neq 0$.

Casi particolari $\vartheta = 0, \pi$
Per $\vartheta = 0$ si ha

$$A = A_0 = \begin{bmatrix} 1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}$$

matrice diagonale con autovalori coincidenti $\lambda_1 = \lambda_2 = \lambda_3 = 1$ e con tre autovettori linearmente indipendenti $\underline{e}_1, \underline{e}_2, \underline{e}_3$.
Infatti, il numero di autovettori linearmente indipendenti associati all'autovalore λ_1 è dato da $n-\mathrm{rango}(A - \lambda_1 I) = n = 3$, essendo $\mathrm{rango}(A - \lambda_1 I) = 0$ poiché

$$(A - \lambda_1 I)\underline{w_1} = \begin{bmatrix} 0 & 0 & 0\\ 0 & 0 & 0\\ 0 & 0 & 0\end{bmatrix} = \underline{0}\ (\textit{matrice nulla}).$$

Per $\vartheta = \pi$ si ha

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

matrice diagonale con autovalori $\lambda_{1,2} = -1$,$\lambda_3 = 1$.
Si noti che

$$(A - \lambda_1 I)\underline{w_1} = \begin{bmatrix} 0 & 0 & 0\\ 0 & 0 & 0\\ 0 & 0 & \boxed{2}\end{bmatrix} \underline{w_1} = \underline{0}$$

con x_1, y_1 qualsiasi e $z_1 = 0$. Essendo $\mathrm{rango}(A - \lambda_1 I) = 1$, si ha che il numero di autovettori linearmente indipendenti associati a λ_1 è $n-\mathrm{rango}(A - \lambda_1 I) = n - 1 = 2$ e, ad esempio, $\underline{e}_1, \underline{e}_2$ sono autovettori.
Inoltre

$$(A - \lambda_3 I)\underline{w_3} = \begin{bmatrix} \boxed{\begin{matrix} -2 & 0\\ 0 & -2\end{matrix}} & \begin{matrix}0\\0\end{matrix}\\ \begin{matrix}0 & 0\end{matrix} & 0\end{bmatrix} \underline{w_3} = \underline{0}$$

con z_3 qualsiasi e $x_3 = y_3 = 0$. Infatti, $n-\mathrm{rango}(A - \lambda_1 I) = n - 2 = 1$, essendo $\mathrm{rango}(A - \lambda_1 I) = 2$, e, ad esempio, $\underline{e}_3$ è autovettore.

6 Norme vettoriali e matriciali

6.1 Norme vettoriali: definizione ed esempi

In molte applicazioni è conveniente associare ad ogni vettore uno scalare non negativo che ne misuri in qualche senso la grandezza.
In prima istanza, il concetto di norma è una generalizzazione del concetto lunghezza di un vettore.
L'obiettivo che ci si prefigge è quello di misurare delle distanze o di quantificare degli errori.

Definizione 6.1 ***Norma vettoriale*** *Si definisce norma nello spazio vettoriale* $\mathbb{C}^n$ $(K = \mathbb{C})$ *una qualsiasi funzione* $\|\cdot\| : \mathbb{C}^n \to \mathbb{R}$ *tale che valgano le seguenti proprietà:*

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

Esempio 6.1

- *Norma 2:* $\|\underline{x}\|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2}$*;*
- *Norma 1:* $\|\underline{x}\|_1 = \sum_{i=1}^{n} |x_i|$*;*
- *Norma infinito:* $\|\underline{x}\|_\infty = \max_{i=1,\ldots,n} |x_i|$

Proprietà 6.1

1. *Una norma vettoriale è una funzione (uniformemente) continua, ossia per ogni* $\varepsilon > 0$ *esiste* $\delta = \delta(\varepsilon)$ *tale che per ogni* $\underline{x}, \underline{y} \in \mathbb{C}^n$ *con* $\|x - y\| \leq \delta$ *allora vale che* $|\|\underline{x}\| - \|\underline{y}\|| \leq \varepsilon$*.*
2. *Le norme vettoriali di* $\mathbb{C}^n$ *(con* n *fissato e finito) sono topologicamente equivalenti, ossia per ogni* $\|\cdot\|_\star, \|\cdot\|_{\star\star} : \mathbb{C}^n \to \mathbb{R}$ *norme vettoriali assegnate esistono due costanti* $\alpha, \beta \in \mathbb{R}$ *con* $0 < \alpha \leq \beta$ *tali che per ogni* $\underline{x} \in \mathbb{C}^n$

$$\alpha\|\underline{x}\|_\star \leq \|\underline{x}\|_{\star\star} \leq \beta\|\underline{x}\|_\star.$$

Il significato delle due precedenti prorprietà è il seguente:

1. vettori "vicini" fra loro hanno norme "vicine" fra loro;
2. la differenza tra scegliere una norma od un'altra per effettuare le misurazioni è puramente in una costante moltiplicativa.

Con riferimento alla seconda proprietà, nel caso delle norme date negli esempi 6.1 le relazioni sono le seguenti.

Proposizione 6.1 *Per ogni $\underline{x} \in \mathbb{C}^n$ vale che*

$$\begin{aligned} \|\underline{x}\|_\infty &\leq \|\underline{x}\|_2 \leq \sqrt{n}\ \|\underline{x}\|_\infty \\ \|\underline{x}\|_2 &\leq \|\underline{x}\|_1 \leq \sqrt{n}\ \|\underline{x}\|_2 \\ \|\underline{x}\|_\infty &\leq \|\underline{x}\|_1 \leq n\ \|\underline{x}\|_\infty \end{aligned}$$

6.2 Norme matriciali: definizione ed esempi

Come già nel caso dei vettori, in molte applicazioni è conveniente associare ad ogni matrice uno scalare non negativo che ne misuri in qualche senso la grandezza.
L'obiettivo è quello di misurare delle distanze o di quantificare degli errori.

Definizione 6.2 ***Norma matriciale***
Si definisce norma nello spazio vettoriale $\mathbb{C}^{n\times n}$ $(K = \mathbb{C})$ una qualsiasi funzione $\|\cdot\| : \mathbb{C}^{n\times n} \to \mathbb{R}$ tale che valgano le seguenti proprietà:

1. *$\|A\| \geq 0$ per ogni $A \in \mathbb{C}^{n\times n}$ e $\|A\| = 0$ se e solo se $A = 0$ (matrice nulla);*
2. *$\|\alpha A\| = |\alpha|\|A\|$ per ogni $A \in \mathbb{C}^{n\times n}$ e per ogni $\alpha \in K = \mathbb{C}$;*
3. *$\|A + B\| \leq \|A\| + \|B\|$ per ogni $A, B \in \mathbb{C}^{n\times n}$;*
4. *$\|AB\| \leq \|A\|\|B\|$ per ogni $A, B \in \mathbb{C}^{n\times n}$ (proprietà submoltiplicativa);*

Si noti che la condizione 4) nella precedente definizione è motivata dal fatto che esistono matrici diverse dalla matrice nulla il cui prodotto è la matrice nulla. Ad esempio, si considerino

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} \text{ e } B = \begin{bmatrix} -6 & -8 \\ 3 & 4 \end{bmatrix}.$$

Per la 1) si ha che $\|A\| > 0$ e $\|B\| > 0$, ma

$$C = AB = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

e quindi, sempre per la 1), si ha $\|C\| = 0$. Questo esempio giustifica la necessità di mettere $\leq$ nella 4). Poiché le condizioni 1)-3) sono le medesime già imposte nella definizione di norma vettoriale, anche nel caso della norma matriciale valgono le seguenti due proprietà.

Proprietà 6.2

1. *Una norma matriciale è una funzione (uniformemente) continua, ossia per ogni $\varepsilon > 0$ esiste $\delta = \delta(\varepsilon)$ tale che per ogni $A, B \in \mathbb{C}^{n\times n}$ con $\|A - B\| \leq \delta$ allora vale che $|\|A\| - \|B\|| \leq \varepsilon$.*
2. *Le norme matriciali di $\mathbb{C}^{n\times n}$ (con n fissato e finito) sono topologicamente equivalenti ossia per ogni $\|\cdot\|_\star, \|\cdot\|_{\star\star} : \mathbb{C}^{n\times n} \to \mathbb{R}$ norme matriciali assegnate esistono due costanti $\alpha, \beta \in \mathbb{R}$ con $0 < \alpha \leq \beta$ tali che*

   $$\alpha\|A\|_\star \leq \|A\|_{\star\star} \leq \beta\|A\|_\star$$

   per ogni $A \in \mathbb{C}^{n\times n}$.

Come già nel caso delle norme vettoriali, il significato delle due precedenti prorprietà è il seguente:

1. matrici "vicine" fra loro hanno norme "vicine" fra loro;

2. la differenza tra scegliere una norma od un'altra per effettuare le misurazioni è puramente in una costante moltiplicativa.

È interessante indagare ulteriormente le connessioni fra norme vettoriali e norme matriciali.

Definizione 6.3 ***Norma matriciale compatibile***
Sia $\|\cdot\|_\star$ una norma vettoriale e sia $\|\cdot\|_{\star\star}$ una norma matriciale. La $\|\cdot\|_{\star\star}$ è una norma matriciale compatibile con la norma vettoriale $\|\cdot\|_\star$ se per ogni $A \in \mathbb{C}^{n\times n}$ e per ogni $\underline{x} \in \mathbb{C}^n$ si ha

$$\|Ax\|_\star \leq \|A\|_{\star\star}\|x\|_\star.$$

In particolare si ha la seguente definizione.

Definizione 6.4 ***Norma matriciale indotta***
Sia $\|\cdot\|$ una norma vettoriale. Si dice norma matriciale indotta dalla norma vettoriale $\|\cdot\|$ la funzione

$$\|A\| = \max_{\|\underline{x}\|=1} \|A\underline{x}\| = \sup_{\|\underline{x}\|\neq\underline{0}} \frac{\|Ax\|}{\|x\|}$$

per ogni $A \in \mathbb{C}^{n\times n}$.

Si tenga presente che una norma matriciale indotta da una norma vettoriale è una norma matriciale compatibile, anzi è la più piccola tra le norme matriciali compatibili con la fissata norma vettoriale.

Esempio 6.2
Si dimostra che le seguenti norme matriciali sono le norme matriciali indotte dalle corrispondenti norme vettoriali precedentemente definite.

- *Norma 1:* $\|A\|_1 = \max_{j=1,\dots,n} \sum_{i=1}^{n} |a_{ij}|$;
- *Norma infinito:* $\|A\|_\infty = \max_{i=1,\dots,n} \sum_{j=1}^{n} |a_{ij}|$;
- *Norma 2:* $\|A\|_2 = \sqrt{\rho(A^H A)}$ *ove* $\rho(B) = \max_{i=1,\dots,n}(|\lambda_i(B)|)$ *viene detto raggio spettrale della matrice B.*

Proposizione 6.2 *Per ogni $\|\cdot\|$ norma matriciale indotta vale che per ogni $A \in \mathbb{C}^{n\times n}$*

$$\rho(A) \leq \|A\|.$$

Proposizione 6.3 *Nel caso particolare in cui la matrice A sia simmetrica (Hermitiana), ossia $A = A^H$ si ha che*

$$\begin{aligned} \|A\|_1 &= \|A\|_\infty, \\ \|A\|_2 &= \rho(A). \end{aligned}$$

La prima uguaglianza è ovvia. Per la seconda vale che

$$\|A\|_2 = \sqrt{\rho(A^H A)} = \sqrt{\rho(A^2)} = \sqrt{\rho^2(A)} = |\rho(A)| = \rho(A),$$

in quanto se $Ax = \lambda x$ allora $A^2 x = A(Ax) = A(\lambda x) = \lambda^2 x$.

Parte II
Algebra lineare numerica

7 Rappresentazione dei numeri e teoria dell'errore

7.1 Rappresentazione dei numeri

La rappresentazione dei numeri usata dai calcolatori moderni è quella in base $B = 2$ e in virgola mobile normalizzata.
Inizialmente, per fissare le idee, consideriamo il caso più semplice di rappresentazione in base $B = 10$. Inoltre, per meglio apprezzare i vantaggi di tale rappresentazione in virgola mobile normalizzata, analizziamo preliminarmente il caso della rappresentazione in virgola fissa.

Sia $\alpha \in \mathbb{R}$.
Il numero α viene convenzionalmente scritto come

$$\alpha = \underbrace{\text{sign}(\alpha)}_{\text{segno}} \underbrace{\alpha_n \alpha_{n-1} \dots \alpha_1 \alpha_0}_{\text{parte intera}} . \underbrace{\alpha_{-1} \alpha_{-2} \dots \alpha_{-m} \dots}_{\text{parte decimale o mantissa}}$$

con $\alpha_i \in \{0, \dots, 9\}$, $\alpha_n \neq 0$.
Come ben noto, questa scrittura sta a significare

$$\begin{aligned} \alpha &= \text{sign}(\alpha)\ \alpha_n 10^n + \alpha_{n-1} 10^{n-1} + \dots + \alpha_1 10^1 + \alpha_0 10^0 \\ &\quad + \alpha_{-1} 10^{-1} + \alpha_{-2} 10^{-2} + \dots + \alpha_{-m} 10^{-m} + \dots \end{aligned}$$

ovvero, in forma compatta,

$$\alpha = \text{sign}(\alpha) \sum_{k=-\infty}^{n} \alpha_k B^k$$

Esempio 7.1

$$\begin{aligned} \alpha &= -1234.56 \\ &= -1 \cdot 10^3 + 2 \cdot 10^2 + 3 \cdot 10^1 + 4 \cdot 10^0 + 5 \cdot 10^{-1} + 6 \cdot 10^{-2} \end{aligned}$$

Ora, in vista dell'applicazione su calcolatore, è importante garantire che tale rappresentazione del numero non sia ambigua, vale a dire occorre garantirne l'unicità.
Tale proprietà è verificata a patto di richiedere che non esista $\overline{k}$ tale che per ogni $k \leq \overline{k}$ sia $a_k = B - 1$, vale a dire che, ad esempio nel caso della base $B = 10$, i coefficienti della parte decimale non siano definitivamente uguali a 9.

Esempio 7.2 *Per illustrare la necessità dell'ipotesi sopra formulate, si consideri il numero*

$$\alpha = +0.9999 \dots 9 \dots$$

tale che per ogni $k \leq \overline{k} = -1$ *è* $\alpha_k = B - 1 = 9$.
Si ha che

$$\alpha = + \sum_{k=-\infty}^{-1} (B-1) B^k$$

$$
\begin{aligned}
&= \sum_{k=1}^{+\infty}(B-1)B^{-k} \\
&= (B-1)B^{-1}\sum_{k=1}^{+\infty}B^{-k+1} \\
&= (B-1)B^{-1}\sum_{k=0}^{+\infty}B^{-k} \quad \textit{(serie geometrica di ragione } B^{-1} \\
&\qquad\qquad\qquad\qquad\qquad\qquad \textit{con } |B^{-1}|<1\textit{)} \\
&= (B-1)B^{-1}\frac{1}{1-B^{-1}} \\
&= (B-1)B^{-1}\frac{B}{B-1} = 1,
\end{aligned}
$$

ovvero una rappresentazione differente dello stesso numero.

Si può quindi enunciare il seguente teorema di rappresentazione.

Teorema 7.1 *Per ogni $\alpha \in \mathbb{R}$ esiste unica la rappresentazione in virgola fissa del numero rispetto alla base B come*

$$\alpha = \text{sign}(\alpha)\sum_{k=-\infty}^{n}\alpha_k B^k$$

con $\alpha_k \in \{0,1,\ldots,B-1\}$, $\alpha_n \neq 0$, a patto che non esista $\overline{k}$ tale che per ogni $k \leq \overline{k}$ sia $a_k = B-1$.

La rappresentazione in virgola fissa ha come vantaggio la sua semplicità, caratteristica che la rende la rappresentazione convenzionale, ma ha come svantaggio la sua "uniformità".
Infatti, poiché su calcolatore si ha a disposizione uno spazio limitato di memoria per la memorizzazione del numero, di tutti i numeri reali è possibile rappresentarne solo una parte discreta.
Supposto di aver fissato un numero di cifre massimo per la parte intera e un numero di cifre massimo per la mantissa, i numeri rappresentabili risulteranno equispaziati. Pertanto, avendo fissato tali due numeri di cifre massime in modo adatto per numeri molto grandi (per i quali ha molta importanza la parte intera), risulterà impossibile apprezzare le differenze fra numeri molto piccoli (per i quali ha molta importanza la parte decimale) e viceversa.

Consideriamo, ora, la cosiddetta **rappresentazione in virgola mobile normalizzata**. Essa è una rappresentazione equivalente del numero con la caratteristica di metterne esplicitamente in evidenza l'ordine di grandezza.
Più precisamente, si può riscrivere il numero $\alpha \in \mathbb{R}$ come

$$
\begin{aligned}
\alpha &= \text{sign}(\alpha)\sum_{k=-\infty}^{n}\alpha_k B^k, \quad \alpha_n \neq 0 \\
&= \text{sign}(\alpha)B^{n+1}\sum_{k=-\infty}^{n}\alpha_k B^{k-(n+1)}
\end{aligned}
$$

$$
\begin{aligned}
&= \operatorname{sign}(\alpha) B^{n+1} \sum_{s=-\infty}^{-1} \alpha_{s+n+1} B^{s} \\
&= \operatorname{sign}(\alpha) B^{n+1} \sum_{k=1}^{+\infty} \alpha_{n+1-k} B^{-k},
\end{aligned}
$$

ovvero

$$
\alpha = \operatorname{sign}(\alpha) B^{n+1} 0.\alpha_n \alpha_{n-1} \ldots \alpha_1 \alpha_0 \alpha_{-1} \ldots \alpha_{-m} \ldots, \quad \alpha_n \neq 0
$$

o meglio, con una notazione più semplice rispetto agli indici,

$$
\alpha = \operatorname{sign}(\alpha) B^{n+1} \sum_{k=1}^{+\infty} \tilde{\alpha}_k B^{-k}, \quad \tilde{\alpha}_1 \neq 0
$$

(la corrispondenza con gli indici della rappresentazione in virgola fissa è data da $\tilde{\alpha}_k = \alpha_{n+1-k}$). Tale rappresentazione del numero viene detta rappresentazione in virgola mobile normalizzata.
Si noti che l'ipotesi $\tilde{\alpha}_1 \neq 0$, ossia che la rappresentazione sia normalizzata, serve a garantire l'unicità di rappresentazione. Infatti, se così non fosse $\alpha = 10^{-2}$ potrebbe essere scritto, ad esempio, sia come $0.1{*}10^{-1}$ (che è la rappresentazione normalizzata), sia come $0.01 * 10^{0}$.

Ora, poiché su calcolatore si ha a disposizione uno spazio limitato di memoria per la memorizzazione del numero, di tutti i numeri reali è possibile rappresentarne solo una parte discreta; fissato pari a t il numero di cifre significative, il modello di rappresentazione in virgola mobile normalizzata fa si che l'insieme dei numeri rappresentabili su calcolatore sia dato dall'insieme

$$
\begin{array}{c}
\mathbb{F}(B,t,L,U) = \{0\} \cup \{\alpha \in \mathbb{R} \text{ tali che } \alpha = \operatorname{sign}(\alpha) B^{p} \sum_{i=1}^{t} \alpha_i B^{-i} \\
= \operatorname{sign}(\alpha) B^{p} . \alpha_1 \alpha_2 \ldots \alpha_t \\
\text{con } \alpha_i \in \{0, 1, \ldots, B-1\}, \alpha_1 \neq 0 \text{ e } L \leq p \leq U (\text{range esponente})\}
\end{array}
$$

Si ha quindi che il più piccolo e il più grande numero positivo rappresentabile sono rispettivamente pari a

$$
\begin{aligned}
m &= \min_{\alpha > 0} \{\alpha \in \mathbb{F}(B,t,L,U)\} = B^{L-1} \\
M &= \max_{\alpha > 0} \{\alpha \in \mathbb{F}(B,t,L,U)\} = B^{U} \sum_{k=1}^{t} (B-1) B^{-k} = \ldots = B^{U}(1 - B^{t})
\end{aligned}
$$

ove nel primo caso si considera il minimo esponente possibile e delle t cifre significative la prima pari ad 1 e le rimanenti tutte nulle; mentre nel secondo caso si considera il massimo esponente possibile e le t cifre significative pari ad $B - 1$.
Il numero m individua la cosiddetta barriera di underflow (numeri troppo piccoli per essere rappresentati) e il numero M individua la cosiddetta barriera di overflow (numeri troppo grandi per essere rappresentati).

Di più, i numeri rappresentabili si infittiscono vicino alle barriere di underflow, dove ha più importanza la parte decimale, e si diradano vicino alle barriere di overflow, dove ha più importanza la parte intera.
Inoltre, fra due successive potenze della base la suddivisione è equispaziata.

Esempio 7.3 *Per fissare meglio le idee consideriamo il seguente modello di rappresentazione in base $B = 10$, con $t = 2$ cifre significative per la mantissa e con $L = -1$ e $U = 1$ esponenti minimo e massimo.*
L'insieme dei numeri rappresentabili su calcolatore è dato da

$$\mathbb{F}(B,t,L,U) = \{0\} \cup \{\alpha \in \mathbb{R} \text{ tali che } \alpha = \text{sign}(\alpha)0.\alpha_1\alpha_2 B^p \\ \text{con } \alpha_i \in \{0,1,\ldots,9\}, \alpha_1 \neq 0 \text{ e } -1 \leq p \leq 1\}$$

Il più piccolo e più grande numero positivo sono dati da

$$\begin{aligned} m &= 0.10 \cdot 10^{-1} = 10^{-2} \\ M &= 0.99 \cdot 10^{1} = 9.9 \end{aligned}$$

I numeri successivi a m con la medesima potenza della base, ossia 10^{-1}, sono

$$\begin{array}{llll} 0.11 \cdot 10^{-1}, & 0.12 \cdot 10^{-1}, & \ldots, & 0.19 \cdot 10^{-1}, \\ 0.20 \cdot 10^{-1}, & 0.21 \cdot 10^{-1}, & \ldots, & 0.29 \cdot 10^{-1}, \\ \vdots & & & \\ 0.90 \cdot 10^{-1}, & 0.91 \cdot 10^{-1}, & \ldots, & 0.99 \cdot 10^{-1}. \end{array}$$

Quindi la suddivisione tra le due successive potenze della base 10^{-2} e 10^{-1} è equispaziata con passo $1/1000$. Tale numero di suddivisioni dipende esclusivamente dal numero di cifre t fissato per la mantissa.
Ora, la suddivisione tra le potenze della base 10^{-1} e 10^{0} è sempre equispaziata, ma con passo $1/100$ e quella tra le potenze della base 10^{0} e 10^{1} è equispaziata, ma con passo $1/10$.

Operativamente, ogni qualvolta il numero α non risulta rappresentabile esattamente in quanto non appartenente ad $\mathbb{F}$, se ne considera un'approssimazione detta rounding, vale a dire

$$\alpha \quad = \quad = \text{sign}(\alpha)B^p.\alpha_1\alpha_2\ldots\hat{\alpha}_t \text{ con } \hat{\alpha}_t = \begin{cases} \alpha_t & \text{se } \alpha_{t+1} \leq B/2 \\ (\alpha_t) + 1 & \text{se } \alpha_{t+1} > B/2 \end{cases}$$

In numeri $\alpha \in \mathbb{R}$ risultano quindi suddivisi in numeri macchina, ossia numeri rappresentabili esattamente su calcolatore, e in numeri non di macchina, che vengono rappresentati tramite il numero di macchina floating(α), secondo la convenzione del rounding sopra specificata.
Introduciamo, ora, le seguenti definizioni di errore assoluto, relativo e percentuale.

Definizione 7.1
Sia x il valore esatto e sia $\tilde{x}$ il valore con cui x viene approssimato per una qualche ragione. Si definiscono le quantità

$$\begin{array}{ll} e_a = |x - \tilde{x}| & \text{Errore assoluto} \\ e_r = |x - \tilde{x}|/|x| & \text{Errore relativo} \\ e_p = e_r * 100\% & \text{Errore percentuale} \end{array}$$

Esempio 7.4
Vediamo su degli esempi la differenza di informazione fornita dai differenti tipi di errore.
Sia $x = 0.1$ *e* $\tilde{x} = 0.99$. *Si ha che*

Errore assoluto	$e_a = \lvert x - \tilde{x}\rvert$	$= 10^{-2}$,
Errore relativo	$e_r = \lvert x - \tilde{x}\rvert / \lvert x\rvert$	$= 10^{-1}$,
Errore percentuale	$e_p = e_r * 100\%$	$= 10\%$.

Sia $x = 1000$ *e* $\tilde{x} = 999$. *Si ha che*

Errore assoluto	$e_a = \lvert x - \tilde{x}\rvert$	$= 1$,
Errore relativo	$e_r = \lvert x - \tilde{x}\rvert / \lvert x\rvert$	$= 10^{-3}$,
Errore percentuale	$e_p = e_r * 100\%$	$= 0.1\%$.

Ora, l'errore assoluto è più piccolo nel primo esempio e molto più grande nel secondo. Tuttavia, tale informazione è in un certo senso fuorviante in quanto non tiene conto dell'ordine di grandezza dei numeri in esame. In effetti, l'importanza dell'errore commesso è molto maggiore nel primo caso in cui si stanno misurando la differenza fra due numeri piccoli, che nel secondo in cui si sta misurando la differenza tra due numeri grandi (è sulla terza cifra significativa). Tale fatto è invece ben indicato dall'errore relativo e, a maggior ragione, dall'errore percentuale.

Ora, avendo già garantito la non ambiguità del modello di rappresentazione, occorre garantirne la consistenza, vale a dire occorre garantire che esso sia un buon modello di rappresentazione. In effetti, tale proprietà di consistenza è garantita dal fatto che per ogni $\varepsilon > 0$ e per ogni $\alpha \in \mathbb{R}$ esiste β a rappresentazione finita tale che $|\alpha - \beta| < \varepsilon$.

Di più, la domanda che è naturale porsi è la seguente:

qual è l'entità dell'errore commesso nell'approssimazione di $\alpha \in \mathbb{R}$ con $fl(\alpha)$?

La risposta è formalizzata nel seguente teorema.

Teorema 7.2 *L'errore relativo*

$$e_r = |\alpha - fl(\alpha)|/|\alpha| \leq \varepsilon_M = B^{1-t}/2$$

ove ε_M *viene detta* ***precisione di macchina****; essa dipende solo dalla base* B *e dal numero di cifre* t *usate per la rappresentazione della mantissa.*

Una caratterizzazione equivalente è la seguente: la precisione di macchina ε_M è il più piccolo numero positivo tale che

$$fl(1 + \varepsilon_M) > 1,$$

ossia sommare ad 1 un numero più piccolo della precisione di macchina equivale a sommare 0.

7.2 Teoria dell'errore

7.2.1 Premesse

È possibile interpretare il generico problema di calcolare un risultato y univocamente determinato da un numero finito di dati $x_1, \ldots, x_n$ come quello del calcolo del valore di una funzione

$$F : \mathbb{R}^n \longrightarrow \mathbb{R}$$
$$y = F(x_1, \ldots, x_n)$$

Volendo dare una prima descrizione sommaria, il valore di y effettivamente calcolato può essere affetto da:

- **Errore analitico**, ossia l'errore indotto sul risultato dall'approssimazione della funzione F con una funzione G razionale, ossia una funzione calcolabile con un numero finito di operazioni aritmetiche elementari $(+, -, *, /)$. Tale tipo di errore è connesso alla questione della **"convergenza della soluzione discreta alla soluzione continua"**: ad esempio, supponiamo di voler calcolare la derivata della funzione $F(x) = \sin(x)$ in $x = x_0$ e di non conoscerne la derivazione formale; si può allora approssimare $F'(x)$ con il rapporto incrementale $(F(x_0 + h) - F(x_0))/h$ con un errore analitico pari a $O(h)$ con h fissato.

- **Errore inerente**, ossia l'errore indotto sul risultato dall'errore di rappresentazione sui dati con i numeri macchina $fl(x_i)$ anziché x_i, $i = 1, \ldots, n$. Tale tipo di errore è connesso al cosiddetto **condizionamento di un problema**, che definiremo in sezione 7.2.3.

- **Errore algoritmico**, ossia l'errore indotto sul risultato dall'uso dell'aritmetica finita. In effetti, anche un'operazione fra due numeri macchina può avere come risultato un numero non di macchina; tale numero risultante necessiterà di essere approssimato con un numero macchina e tale approssimazione avrà ovviamente delle conseguenze sui calcoli successivi. Tale tipo di errore è connesso alla cosiddetta **stabilità del metodo** di calcolo, che definiremo 7.2.5.

Volendo schematizzare, posto $\underline{x} = (x_1, \ldots, x_n)$ e $fl(\underline{x}) = (fl(x_1), \ldots, fl(x_n))$, si ha

F non razionale	$\rightsquigarrow$	G razionale	Errore analitico
$\underline{x}$	$\rightsquigarrow$	$fl(\underline{x})$	Errore inerente
$G(fl(\underline{x}))$	$\rightsquigarrow$	$\Phi(fl(\underline{x}))$	Errore algoritmico

ove Φ è la funzione effettivamente calcolata al posto della funzione G a causa delle operazioni in aritmetica finita.
In definitiva

$y = F(\underline{x})$	$\rightsquigarrow$	$\Phi(fl(\underline{x}))$
con $\underline{x} = (x_1, \ldots, x_n)$	$\rightsquigarrow$	con $fl(\underline{x}) = (fl(x_1), \ldots, fl(x_n))$
valore che si vorrebbe calcolare		valore effettivamente calcolato

e vale la seguente corrispondenza

Errore analitico	$\rightsquigarrow$	Convergenza
Errore inerente	$\rightsquigarrow$	Condizionamento
Errore algoritmico	$\rightsquigarrow$	Stabilità

7.2.2 Caso di F funzione razionale

Consideriamo per primo il caso più semplice in cui F sia una funzione razionale, ossia una funzione calcolabile con un numero finito di operazioni aritmetiche elementari, ossia

$$y = F(x_1, \ldots, x_n) \rightsquigarrow \Phi(fl(x_1), \ldots, fl(x_n))$$

con $G = F$.
Posto $\underline{x} = (x_1, \ldots, x_n)$ e $fl(\underline{x}) = (fl(x_1), \ldots, fl(x_n))$, si definisce

$$\text{Errore totale} \qquad e_{TOT} = \frac{\Phi(fl(\underline{x})) - F(\underline{x})}{F(\underline{x})},$$

$$\text{Errore inerente} \qquad e_{IN} = \frac{F(fl(\underline{x})) - F(\underline{x})}{F(\underline{x})},$$

$$\text{Errore algoritmico} \qquad e_{AL} = \frac{\Phi(fl(\underline{x})) - F(fl(\underline{x}))}{F(fl(\underline{x}))}.$$

Si noti che nell'espressione dell'errore inerente e_{IN} non compare Φ in quanto si vogliono solo misurare le conseguenze sul risultato dell'uso di $fl(\underline{x})$ al posto di $\underline{x}$ e che nell'espressione dell'errore algoritmico e_{AL} non compare $\underline{x}$ in quanto si vogliono solo misurare le conseguenze sul risultato dell'uso dell'aritmetica finita, assumendo che il dato iniziale sia $fl(\underline{x})$.
Sotto l'ipotesi di funzione F "sufficientemente" regolare si ha che

$$e_{TOT} \doteq e_{IN} + e_{AL} \quad \text{(uguaglianza al primo ordine)}$$

ossia i due tipi di errore si assommano con contributi separati in modo lineare. Infatti

$$\begin{aligned} e_{TOT} &= \frac{\Phi(fl(\underline{x})) - F(\underline{x})}{F(\underline{x})} \\ &= \frac{\Phi(fl(\underline{x}))}{F(fl(\underline{x}))} \frac{F(fl(\underline{x}))}{F(\underline{x})} - 1 \\ &= (e_{AL} + 1)(e_{IN} + 1) - 1 \\ &= e_{AL} + e_{IN} + e_{AL} e_{IN} + 1 - 1 \\ &\doteq e_{IN} + e_{AL} \quad \text{(uguaglianza al primo ordine)} \end{aligned}$$

7.2.3 Problemi ben/mal condizionati

L'errore inerente e_{IN} misura il cosiddetto condizionamento di un problema definito come calcolo di una funzione $F(\underline{x})$. Più precisamente, si dice che un

problema è ben condizionato se a piccole perturbazioni sui dati $\underline{x}$ corrispondono piccole variazioni sul risultati $F(\underline{x})$.
Sotto l'ipotesi di funzione F "sufficientemente" regolare si ha che

$$e_{IN} = \sum_{i=1}^{n} c_{x_i} \varepsilon_{x_i}$$

ove $\varepsilon_{x_i} = (fl(x_i) - x_i)/x_i)$ è l'errore relativo di rappresentazione del dato x_i e

$$c_{x_i} = \frac{x_i}{F(x)} \frac{\partial F(\underline{z})}{\partial x_i}_{|\underline{z}=\underline{x}}$$

è il coefficiente di amplificazione dell'errore sul dato ε_{x_i}.
L'espressione dell'errore inerente sopra riportata ha la seguente motivazione. Nel caso $n = 1$ si ha

$$\begin{aligned} e_{IN} &= \frac{F(fl(\underline{x})) - F(\underline{x})}{F(\underline{x})} \\ &= \frac{F'(\underline{x})(fl(\underline{x}) - \underline{x}) + o(fl(\underline{x}) - \underline{x})}{F(\underline{x})} \\ &= \frac{\underline{x}F'(\underline{x})}{F(\underline{x})} \frac{fl(\underline{x}) - \underline{x}}{\underline{x}} + o(fl(\underline{x}) - \underline{x}) \end{aligned}$$

e nel caso $n > 1$ si ha

$$\begin{aligned} e_{IN} &= \frac{F(fl(\underline{x})) - F(\underline{x})}{F(\underline{x})} \\ &= \frac{1}{F(\underline{x})} \left(\sum_{i=1}^{n} \frac{\partial F(\underline{z})}{\partial x_i}_{|\underline{z}=\underline{x}} (fl(x_i) - x_i) + o(\|fl(\underline{x}) - \underline{x}\|) \right) \\ &= \sum_{i=1}^{n} \frac{x_i}{F(\underline{x})} \frac{\partial F(\underline{z})}{\partial x_i}_{|\underline{z}=\underline{x}} \frac{fl(x_i) - x_i}{x_i} + o(\|fl(\underline{x}) - \underline{x}\|) \end{aligned}$$

Si noti che i coefficienti c_{x_i} sono dei veri e propri coefficienti di amplificazione: se i c_i sono piccoli, si avrà un e_{IN} piccolo, essendo l'errore di rappresentazione dei dati al più pari alla precisione di macchina. Viceversa, se i c_i sono grandi, si avrà un e_{IN} grande e quindi un grande errore sul risultato del calcolo di $F(x)$, per quanto gli ε_i siano piccoli. In tale caso si dice che il problema è un problema malcondizionato.
Si noti che il condizionamento è una caratteristica intrinseca del problema rappresentato dal metodo scelto per il calcolo di F. Pertanto, l'unica possibilità a disposizione è quella di cambiare la funzione F, intendendo con ciò un'eventuale rimodellazione del problema che porti al calcolo di una diversa funzione.

Esempio 7.5 *La soluzione $y(t)$ del problema di Cauchy*

$$\begin{cases} y' = h(t, y(t)) \\ y(t_0) = y_0 \end{cases}$$

è soluzione dell'equazione integrale

$$y(t) = y(t_0) + \int_{t_0}^{t} h(s, y(s))ds$$

e viceversa. La modellazione è tuttavia di natura completamente diversa.

D'altro canto, l'errore algoritmico e_{AL} misura invece la "stabilità" del metodo scelto per il calcolo di F. Poiché occorre prima analizzare in dettaglio il caso delle operazioni aritmetiche elementari, rimandiamo alla sezione 7.2.5 un'analisi più approfondita delle origini e delle conseguenze dell'errore algoritmico.
Si tenga tuttavia presente che se il problema di calcolo è malcondizionato non ha senso porsi il quesito della stabilità del metodo in quanto nell'errore totale e_{TOT} prevale l'errore inerente e_{IN}.

7.2.4 Errore nelle operazioni di macchina

Innanzitutto analizziamo l'errore totale commesso nell'effettuare le operazioni aritmetiche elementari, vale a dire nel calcolare

$$F(x, y) = x \circ y$$

ove $\circ$ denota una delle quattro operazioni aritmetiche elementari $+,\ -,\ *,\ /$.
Si ha

$$\begin{array}{ccc} x & \rightsquigarrow & fl(x) \\ y & \rightsquigarrow & fl(y) \end{array} \longrightarrow fl(x) \circ fl(y) \rightsquigarrow fl(fl(x) \circ fl(y)),$$

ossia i due dati x e y vengono approssimati con i rispettivi floating $fl(x)$ e $fl(y)$, si effettua poi l'operazione $\circ$ e, poiché non è detto che l'operazione $\circ$ su due numeri macchina dia come risultato un numero macchina, occorre in genere fare il floating del risultato.
Quindi,

$$e_{TOT} \doteq e_{IN} + e_{AL},$$

ove

$$\begin{array}{rcl} e_{IN} & = & c_x\varepsilon_x + c_y\varepsilon_y, \\ e_{AL} & = & \varepsilon \text{ errore locale generato nella singola operazione con } |\varepsilon| < \varepsilon_M, \\ e_{AN} & = & 0 \text{ poiché } F \text{ è una funzione razionale.} \end{array}$$

Andiamo ora a considerare in dettaglio le quattro operazioni aritmetiche elementari.

1. $\mathbf{F(x, y) = x + y}$. Si ha che

$$\begin{array}{rcl} c_x & = & \dfrac{x}{F(x,y)}\dfrac{\partial F}{\partial x} = \dfrac{x}{x+y} \cdot 1 = \dfrac{x}{x+y} \\ c_y & = & \dfrac{y}{F(x,y)}\dfrac{\partial F}{\partial y} = \dfrac{y}{x+y} \cdot 1 = \dfrac{y}{x+y} \end{array}$$

da cui

$$e_{TOT} = \underbrace{\frac{x}{x+y}\varepsilon_x + \frac{y}{x+y}\varepsilon_y}_{\text{errore inerente}} + \underbrace{\varepsilon}_{\text{errore algoritmico}}$$

2. $\mathbf{F}(\mathbf{x}, \mathbf{y}) = \mathbf{x} - \mathbf{y}$. Si ha che

$$\begin{aligned} c_x &= \frac{x}{F(x,y)}\frac{\partial F}{\partial x} = \frac{x}{x-y} \cdot 1 = \frac{x}{x-y} \\ c_y &= \frac{y}{F(x,y)}\frac{\partial F}{\partial y} = \frac{y}{x-y} \cdot (-1) = -\frac{y}{x-y} \end{aligned}$$

da cui

$$e_{TOT} = \underbrace{\frac{x}{x-y}\varepsilon_x - \frac{y}{x-y}\varepsilon_y}_{\text{errore inerente}} + \underbrace{\varepsilon}_{\text{errore algoritmico}}$$

3. $\mathbf{F}(\mathbf{x}, \mathbf{y}) = \mathbf{x} * \mathbf{y}$. Si ha che

$$\begin{aligned} c_x &= \frac{x}{F(x,y)}\frac{\partial F}{\partial x} = \frac{x}{xy} \cdot y = 1 \\ c_y &= \frac{y}{F(x,y)}\frac{\partial F}{\partial y} = \frac{y}{xy} \cdot x = 1 \end{aligned}$$

da cui

$$e_{TOT} = \underbrace{\varepsilon_x + \varepsilon_y}_{\text{errore inerente}} + \underbrace{\varepsilon}_{\text{errore algoritmico}}$$

4. $\mathbf{F}(\mathbf{x}, \mathbf{y}) = \mathbf{x}/\mathbf{y}$. Si ha che

$$\begin{aligned} c_x &= \frac{x}{F(x,y)}\frac{\partial F}{\partial x} = \frac{x}{x/y} \cdot \frac{1}{y} = 1 \\ c_y &= \frac{y}{F(x,y)}\frac{\partial F}{\partial y} = \frac{y}{x/y} \cdot \frac{-x}{y^2} = -1 \end{aligned}$$

da cui

$$e_{TOT} = \underbrace{\varepsilon_x - \varepsilon_y}_{\text{errore inerente}} + \underbrace{\varepsilon}_{\text{errore algoritmico}}$$

Nel terzo e nel quarto caso, vale a dire nel caso delle operazioni $*$ e $/$, il problema è sempre ben condizionato indipendentemente dai valori di x e y.
Infatti,

$$\begin{aligned} |\varepsilon_{TOT}| &= |\varepsilon_x \pm \varepsilon_y + \varepsilon| \\ &\leq |\varepsilon_x| + |\varepsilon_y| + |\varepsilon| \\ &\leq 3\varepsilon_M \end{aligned}$$

in quanto $|\varepsilon_x|$, $|\varepsilon_y|$ e $|\varepsilon|$ sono maggiorati da ε_M.
Di converso, nel primo e nel secondo caso, vale a dire nel caso delle operazioni

+ e −, il problema può essere ben condizionato o mal condizionato a seconda dei valori di x e y. Più precisamente, se la quantità $|x \pm y|$ è piccola allora i coefficienti di amplificazione c_x e c_y esplodono e si ha un'errore inerente elevato per quanto ε_x e ε_y siano piccoli. Tale fenomeno viene anche detto **Errore di cancellazione**.

Esempio 7.6 *Per semplicità consideriamo $B = 10$, $t = 3$ e la procedura di rounding.*
Sia $x = 0.1236$ e $y = -0.1234$. Vale che

$$x + y = 0.0002 = 0.2 \cdot 10^{-3} \text{ risultato esatto}$$

Su calcolatore si ha

$$\begin{array}{lll} x & \rightsquigarrow & fl(x) = 0.124 \cdot 10^0 \\ y & \rightsquigarrow & fl(y) = -0.123 \cdot 10^0 \end{array} \longrightarrow fl(x) + fl(y) = 0.001 \cdot 10^0 = 0.1 \cdot 10^{-2},$$

(il risultato è già un numero macchina e quindi non occorre farne il floating).
Si noti che non si ha nessuna cifra esatta; in effetti, andando a calcolare l'errore relativo e percentuale si ha che

$$\varepsilon_r = \frac{0.1 \cdot 10^{-2} - 0.2 \cdot 10^{-3}}{0.2 \cdot 10^{-3}} = 4 \quad e \quad \varepsilon_p = \varepsilon_r * 100\% = 400\%.$$

Viceversa se la quantità $|x \pm y|$ non è piccola allora i coefficienti di amplificazione c_x e c_y soddisfano

$$|c_x| < 1 \quad \text{e} \quad |c_y| < 1,$$

ovvero il problema è ben condizionato e vale

$$\begin{aligned} |e_{TOT}| &= |c_x\varepsilon_x + c_y\varepsilon_y + \varepsilon| \\ &\leq |c_x||\varepsilon_x| + |c_y||\varepsilon_y| + |\varepsilon| \\ &< 1 \cdot |\varepsilon_x| + 1 \cdot |\varepsilon_y| + |\varepsilon| \\ &\leq 3\varepsilon_M \end{aligned}$$

7.2.5 Stabilità di un metodo di calcolo

Per fissare le idee si pensi di voler calcolare la funzione

$$F(x, y) = x^2 - y^2,$$

tenendo conto che vale pure

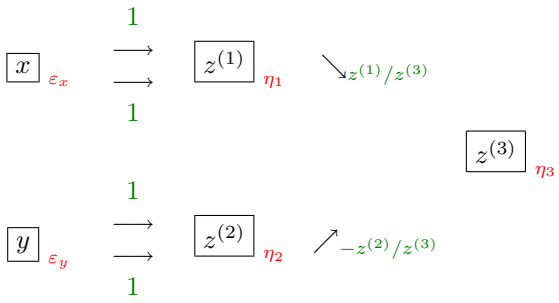
$$F(x, y) = (x - y)(x + y).$$

Infatti, queste due espressioni sono algebricamente equivalenti, ossia forniscono lo stesso risultato in aritmetica esatta, ma non lo sono necessariamente in aritmetica finita.
L'errore algoritmico misura la sensibilità del metodo di calcolo scelto agli errori locali che vengono generati nella sequenza delle operazioni elementari. Ora, se e_{AL} è piccolo si dice che il metodo è numericamente stabile, altrimenti numericamente instabile.

Scriviamo in dettaglio l'algoritmo relativo alla prima procedura di calcolo di F come $F(x,y) = x^2 - y^2$.

$$\begin{aligned} z^{(1)} &= x * x \\ z^{(2)} &= y * y \\ z^{(3)} &= z^{(1)} - z^{(2)} \end{aligned}$$

Uno strumento molto comodo per l'analisi dell'errore e in particolare dell'errore algoritmico è dato dai grafi; più precisamente all'algoritmo sopra indicato possiamo associare questo grafo:



ove

- $\varepsilon_x = (fl(x) - x)/x$ e $\varepsilon_y = (fl(y) - y)/y$ sono gli errori di rappresentazione sui dati (e vale $|\varepsilon_x|, |\varepsilon_x| < \varepsilon_M$ precisione di macchina)
- η_1, η_2 e η_3 sono gli errori locali generati nelle singole operazioni di macchina (e vale $|\eta_1|, |\eta_2|, |\eta_3| < \varepsilon_M$)
- e $c_x(*) = 1$ e $c_y(*) = 1$ sono i coefficienti di amplificazione dell'operazione di prodotto e $c_x(-) = x/(x-y)$ e $c_y(-) = -y/(x-y)$ sono i coefficienti di amplificazione dell'operazione di sottrazione.

Il grafo così costruito viene letto da destra a sinistra sommando all'errore locale generato nell'ultima operazione di macchina il coefficiente di amplificazione relativo al primo arco moltiplicato per "l'errore precedente relativo al primo dato" e il coefficiente di amplificazione relativo al secondo arco moltiplicato per "l'errore precedente relativo al secondo dato". L'"errore precedente relativo al dato" si costruisce ripetendo opportunamente la stessa procedura. Quindi, si ha che

$$\begin{aligned} e_{TOT} &= \eta_3 + \frac{z^{(1)}}{z^{(3)}}(\eta_1 + 1\varepsilon_x + 1\varepsilon_x) - \frac{z^{(2)}}{z^{(3)}}(\eta_2 + 1\varepsilon_y + 1\varepsilon_y) \\ &= \underbrace{\frac{2x^2}{x^2-y^2}\varepsilon_x - \frac{2y^2}{x^2-y^2}\varepsilon_y}_{\text{errore inerente}} + \underbrace{\eta_3 - \frac{y^2}{x^2-y^2}\eta_2 + \frac{x^2}{x^2-y^2}\eta_1}_{\text{errore algoritmico}} \end{aligned}$$

in quanto l'errore inerente vale $c_x(F)\varepsilon_x + c_y(F)\varepsilon_y$ con

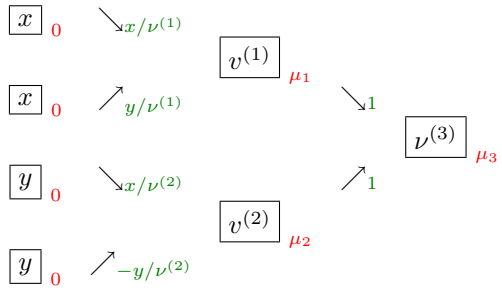
$$c_x(F) = \frac{2x^2}{x^2-y^2} \quad \text{e} \quad c_y(F) = -\frac{2y^2}{x^2-y^2}.$$

Si noti, comunque, che per $|x^2 - y^2|$ piccolo il calcolo di F è un problema mal condizionato.

Ora, scriviamo in dettaglio l'algoritmo relativo alla seconda procedura di calcolo di F come $F(x, y) = (x + y)(x - y)$.

$$\begin{aligned} w^{(1)} &= x + y \\ w^{(2)} &= x - y \\ w^{(3)} &= w^{(1)} * w^{(2)} \end{aligned}$$

In maniera analoga a quanto visto precedentemente, all'algoritmo sopra indicato possiamo associare il grafo:



ove, poiché abbiamo visto che l'errore inerente è indipendente dall'algoritmo scelto e ne conosciamo già l'espressione, usiamo il grafo per calcolare il solo errore algoritmico, ossia poniamo uguali a zero gli errori di rappresentazione dei dati ε_x e ε_y.
Quindi, si ha che

$$\begin{aligned} e_{AL} &= \mu_3 + 1(\mu_1 + 0\frac{x}{\nu^{(1)}} + 0\frac{y}{\nu^{(1)}}) + 1(\mu_2 + 0\frac{x}{\nu^{(2)}} - 0\frac{y}{\nu^{(2)}}) \\ &= \mu_3 + \mu_2 + \mu_1 \end{aligned}$$

Volendo infine confrontare i due algoritmi abbiamo che per il primo

$$|e_{AL}| < \left(1 + \frac{x^2 + y^2}{|x^2 - y^2|}\right)\varepsilon_M;$$

mentre per il secondo

$$|e_{AL}| < 3\varepsilon_M;$$

si può quindi concludere che il secondo algoritmo è più stabile per $|x^2 - y^2|$ "piccolo del primo algoritmo.

Più in generale, un'analisi al primo ordine permette di esprimere l'errore algoritmico e_{AL} come un'opportuna combinazione lineare degli errori locali generati nelle singole operazioni elementari $(+, -, *, /)$, i quali sono di modulo inferiore alla precisione di macchina ε_M. Più precisamente si può affermare che vale la seguente maggiorazione

$$|e_{AL}| < \vartheta(\underline{x})\varepsilon_M + O(\varepsilon_M^2)$$

ove $\vartheta(\underline{x})$ è indipendente da ε_M.
Ora, se e_{AL} è piccolo, ovvero se $\vartheta(\underline{x})$ è piccolo, si dice che il metodo è numericamente stabile, altrimenti numericamente instabile.

7.2.6 Caso di F funzione non razionale

Si può infine affrontare il caso generale di una funzione F non razionale, ossia una funzione non calcolabile con un numero finito di operazioni aritmetiche elementari e che pertanto viene approssimata su calcolatore con una funzione razionale G.
Si ha

$$y = F(x_1, \ldots, x_n) \rightsquigarrow G(x_1, \ldots, x_n) \rightsquigarrow \Phi(fl(x_1), \ldots, fl(x_n))$$

con G funzione razionale che approssima la funzione non razionale F e Φ funzione effettivamente calcolata al posto della funzione G a causa delle operazioni in aritmetica finita.
Si definisce

$$\text{Errore totale} \qquad e_{TOT} = \frac{\Phi(fl(\underline{x})) - F(\underline{x})}{F(\underline{x})},$$

$$\text{Errore inerente} \qquad e_{IN} = \frac{F(fl(\underline{x})) - F(\underline{x})}{F(\underline{x})},$$

$$\text{Errore analitico} \qquad e_{ANL} = \frac{G(fl(\underline{x})) - F(fl(\underline{x}))}{F(fl(\underline{x}))},$$

$$\text{Errore algoritmico} \qquad e_{AL} = \frac{\Phi(fl(\underline{x})) - G(fl(\underline{x}))}{G(fl(\underline{x}))}.$$

Si noti che nell'espressione dell'errore inerente e_{IN} non compaiono G e Φ in quanto si vogliono misurare solo le conseguenze sul risultato dell'uso di $fl(\underline{x})$ al posto di $\underline{x}$. Nell'espressione dell'errore analitico e_{AN} non compaiono x e Φ in quanto si vogliono misurare solo le conseguenze sul risultato dell'uso di G al posto di F, assumendo che il dato iniziale sia $fl(\underline{x})$ e che G venga calcolata esattamente. Nell'espressione dell'errore algoritmico e_{AL} non compaiono x e F in quanto si vogliono misurare solo le conseguenze sul risultato dell'uso dell'aritmetica finita, assumendo che il dato iniziale sia $fl(\underline{x})$ e che la funzione da calcolare sia G.

Sotto l'ipotesi di funzione F "sufficientemente" regolare si ha che

$$e_{TOT} \doteq e_{IN} + e_{AL} + e_{AN} \quad \text{(uguaglianza al primo ordine)}$$

ossia i tre tipi di errore si assommano con contributi separati in modo lineare. Il risultato è la naturale estensione del caso analizzato precedentemente di F funzione razionale.

8 Metodi iterativi per la risoluzione di sistemi lineari

È dato il sistema lineare

$$A\underline{x} = \underline{b} \quad \text{con} \quad A \in \mathbb{R}^{n\times n} \quad \text{e} \quad \underline{x}, \underline{b} \in \mathbb{R}^n,$$

con $\det(A) \neq 0$. Si vogliono individuare dei metodi per determinarne su calcolatore la soluzione, vale a dire per determinare il vettore $\underline{x}^* \in \mathbb{R}^n$ tale che

$$A\underline{x}^* = \underline{b}.$$

Si tenga presente che, essendo $\det(A) \neq 0$, ossia A non singolare, per il Teorema di Rouché-Capelli esiste ed è unica la soluzione $\underline{x}^* \in \mathbb{R}^n$ e è data da

$$\underline{x}^* = A^{-1}\underline{b}.$$

Tuttavia, poiché su calcolatore il calcolo della matrice A^{-1} è costoso e in genere malcondizionato, si cercano altri metodi per determinare la soluzione, i quali non richiedano il calcolo esplicito della matrice inversa A^{-1}.

Nel caso dei metodi iterativi, l'idea è quella di andare a costruire un'opportuna successione di vettori $\{\underline{x}^{(k)}\}_k$ tale che

$$\lim_{k\to+\infty} \underline{x}^{(k)} = \underline{x}^*,$$

ossia, equivalentemente,

$$\lim_{k\to+\infty} x_i^{(k)} = x_i^* \text{ per ogni } \; i = 1, \ldots, n.$$

Nel determinare $\underline{x}^*$ su calcolatore, si avrà evidentemente **errore analitico** in quanto, non avendo a disposizione un tempo infinito di calcolo, occorre troncare la successione $\{\underline{x}^{(k)}\}_k$ ad un "opportuno" valore $\overline{k}$, con criteri che vedremo nel seguito. Di converso, si avrà in genere una minor sensibilità all'**errore algoritmico** proprio grazie alla natura iterativa del metodo. Si ricordi, infine, che l'entità dell'**errore inerente** è una caratteristica del problema $A\underline{x} = \underline{b}$ e non dipende dal metodo scelto per la risoluzione.

Vediamo ora alcuni esempi di metodi.

Metodo di Jacobi
Per semplicità, sia $A \in \mathbb{R}^{3\times 3}$ con $a_{ii} \neq 0$ per ogni $i = 1, \ldots, 3$, cioé il sistema di equazioni

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 = b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 = b_3. \end{cases}$$

Tale sistema può essere riscritto come

$$\begin{cases} x_1 = (b_1 - a_{12}x_2 - a_{13}x_3)/a_{11} \\ x_2 = (b_2 - a_{21}x_1 - a_{23}x_3)/a_{22} \\ x_3 = (b_3 - a_{31}x_1 - a_{32}x_2)/a_{33} \end{cases}$$

ove la prima equazione è stata esplicitata rispetto all'incognita x_1, la seconda equazione rispetto alla'incognita x_2 e la terza equazione rispetto all'incognita x_3. Tale riscrittura suggerisce direttamente la formulazione di un metodo iterativo nel modo seguente

$$\begin{cases} x_1^{(k+1)} = \left(b_1 - a_{12}x_2^{(k)} - a_{13}x_3^{(k)}\right)/a_{11} \\ x_2^{(k+1)} = \left(b_2 - a_{21}x_1^{(k)} - a_{23}x_3^{(k)}\right)/a_{22} \\ x_3^{(k+1)} = \left(b_3 - a_{31}x_1^{(k)} - a_{32}x_2^{(k)}\right)/a_{33} \end{cases}$$

ove si assume che venga assegnato un vettore $\underline{x}^{(0)}$, detto vettore di innesco, a partire dal quale viene generata la successione.
Per un sistema di generica dimensione n, l'i-sima equazione può essere riscritta come

$$x_i = \left(b_i - \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij}x_j \right)/a_{ii}, \text{ per } i = 1, \ldots, n.$$

e quindi il metodo di Jacobi ha la seguente formulazione

$$x_i^{(k+1)} = \left(b_i - \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij}x_j^{(k)} \right)/a_{ii}, \text{ per } i = 1, \ldots, n.$$

Il costo in termini di operazioni di tipo moltiplicativo per calcolare una nuova iterazione è n^2 (n operazioni di tipo moltiplicativo per ciascuna componente moltiplicato per il numero n di componenti). Chiaramente tale costo dovrà essere a sua volta moltiplicato per il numero di iterazioni effettuate.
Infine, si tenga presente che se l'ipotesi $a_{ii} \neq 0$ per ogni $i = 1, \ldots, n$ non è verificata si possono scambiare fra loro equazioni del sistema cosı` che risulti soddisfatta.

Metodo di Gauss-Seidel
Una variante del metodo precedente può essere semplicemente ottenuta utilizzando le componenti della soluzione già aggiornate. Più precisamente, nel caso $n = 3$ il metodo di Gauss-Seidel è dato da

$$\begin{cases} x_1^{(k+1)} = \left(b_1 - a_{12}x_2^{(k)} - a_{13}x_3^{(k)}\right)/a_{11} \\ x_2^{(k+1)} = \left(b_2 - a_{21}x_1^{(k+1)} - a_{23}x_3^{(k)}\right)/a_{22} \\ x_3^{(k+1)} = \left(b_3 - a_{31}x_1^{(k+1)} - a_{32}x_2^{(k+1)}\right)/a_{33} \end{cases}$$

Infatti, nella seconda equazione si utilizza il valore $x_1^{(k+1)}$ appena calcolato al posto del valore $x_1^{(k)}$ e nella terza equazione si utilizzano il valori $x_1^{(k+1)}$ e $x_2^{(k+1)}$ appena calcolati al posto dei valori $x_1^{(k)}$ e $x_2^{(k)}$.
Per un sistema di generica dimensione n il metodo di Gauss-Seidel ha la seguente formulazione

$$x_i^{(k+1)} = \left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)} \right)/a_{ii}, \ i = 1, \ldots, n$$

Come nel caso del metodo precedente, il costo in termini di operazioni di tipo moltiplicativo per calcolare una nuova iterazione è n^2 (n operazioni di tipo moltiplicativo per ciascuna componente moltiplicato per il numero n di componenti, in quanto la sostituzione di $\sum_{j=1}^{i-1} a_{ij}x_j^{(k)}$ con $\sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)}$ non ne influenza il numero). Chiaramente tale costo dovrà essere moltiplicato per il numero di iterazioni effettuate e questo potrà risultare diverso dal numero di iterazioni del metodo precedente.

Si tenga presente che se intuitivamente si potrebbe pensare che il metodo di Gauss-Seidel sia migliore del metodo di Jacobi, in quanto utilizza nel calcolo anche le componenti più aggiornate, questo non è sempre vero.
La domanda che è quindi naturale porsi è da che cosa dipenda la convergenza o meno del metodo iterativo alla soluzione $\underline{x}^*$ e, in caso di convergenza, da che cosa dipenda la sua velocità.
Prima di affrontare questa questione, introduciamo un'ultimo metodo che può essere pensato come un'ulteriore miglioramento del metodo di Gauss-Seidel
Metodo SOR
Tale metodo considera una combinazione con parametro ω dell'iterata precedente $x_i^{(k)}$ e dell'iterata ottenuta applicando il metodo di Gauss-Seidel.
Più precisamente, per un sistema di generica dimensione n il metodo SOR ha la seguente formulazione

$$x_i^{(k+1)} = (1-\omega)x_i^{(k)} + \omega\left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)}\right)/a_{ii},\ i = 1,\dots,n$$

ove $\omega \in (0,2)$.
Teoria della convergenza
I metodi di Jacobi, di Gauss-Seidel e SOR si possono scrivere in forma compatta come

$$\underline{x}^{(k+1)} = P\underline{x}^{(k)} + \underline{c},$$

ove la matrice P viene detta matrice di iterazione.
Più precisamente, posto

$$D = \begin{bmatrix} a_{11} & 0 & \dots & \dots & \dots & 0 \\ 0 & a_{22} & 0 & \dots & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & 0 & a_{n-1n-1} & 0 \\ 0 & \dots & \dots & \dots & 0 & a_{nn} \end{bmatrix}$$

matrice diagonale con diagonale data dalla diagonale principale di A,

$$U = \begin{bmatrix} 0 & a_{12} & \dots & \dots & \dots & a_{1n} \\ 0 & 0 & a_{23} & \dots & \dots & a_{2n} \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & 0 & 0 & a_{n-1n} \\ 0 & \dots & \dots & \dots & 0 & 0 \end{bmatrix}$$

matrice triangolare superiore stretta data dalla triangolare superiore stretta di A e

$$L = \begin{bmatrix} 0 & 0 & \dots & \dots & \dots & 0 \\ a_{12} & 0 & 0 & \dots & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ a_{n-11} & \dots & \dots & \dots & 0 & 0 \\ a_{n1} & \dots & \dots & \dots & a_{nn-1} & 0 \end{bmatrix}$$

matrice triangolare inferiore stretta data dalla triangolare inferiore stretta di A, si ha per il metodo di Jacobi

$$P_J = -D^{-1}(L+U) \quad \text{e} \quad \underline{c}_J = D^{-1}\underline{b},$$

per il metodo di Gauss-Seidel

$$P_{GS} = -(D+L)^{-1}U \quad \text{e} \quad \underline{c}_{GS} = (D+L)^{-1}\underline{b},$$

e per il metodo di SOR

$$P_{SOR} = -(D+\omega L)^{-1}((1-\omega)D - \omega U) \quad e \quad \underline{c}_{SOR} = \omega(D+\omega L)^{-1}\underline{b}.$$

Più in generale, per formulare un metodo iterativo si può considerare la decomposizione della matrice A in due matrici M e N tali che

$$A = M - N$$

e ove si assume che M sia invertibile.
Si ha allora che il sistema $A\underline{x} = \underline{b}$ si può scrivere come

$$M\underline{x} = N\underline{x} + \underline{b},$$

ovvero, poiché M è per definizione invertibile,

$$\underline{x} = M^{-1}N\underline{x} + M^{-1}\underline{b},$$

ovvero

$$\underline{x} = P\underline{x} + \underline{c},$$

cosicché la matrice di iterazione precedentemente introdotta è data da $P = M^{-1}N$ e $\underline{c} = M^{-1}\underline{b}$.
Tale riscrittura fa si che il generico metodo iterativo possa essere scritto come

$$\underline{x}^{(k+1)} = P\underline{x}^{(k)} + c. \tag{8}$$

con $\underline{x}^{(0)}$ vettore di innesco assegnato.
Ora, evidentemente, si ha pure che

$$\underline{x}^* = P\underline{x}^* + c. \tag{9}$$

in quanto questa è una semplice riscrittura secondo la procedura precedente della relazione $A\underline{x}^* = \underline{b}$ e quindi, posto $\underline{e}^{(k+1)} = \underline{x}^* - \underline{x}^{(k+1)}$, errore assoluto al passo $k+1$, sottraendo membro a membro la (8) e la (9), si ha

$$\begin{aligned} \underline{e}^{(k+1)} &= \underline{x}^* - \underline{x}^{(k+1)} \\ &= P\underline{x}^* + c - (P\underline{x}^{(k+1)} + c) \\ &= P(\underline{x}^* - \underline{x}^{(k+1)}) \\ &= P\underline{e}^{(k)}. \end{aligned}$$

Pertanto, la relazione che governa la trasformazione dell'errore assoluto dal passo k al passo $k+1$ è data da

$$\underline{e}^{(k+1)} = P\underline{e}^{(k)}.$$

Poiché tale relazione vale per ogni passo k si può ulteriormente riscrivere

$$\underline{e}^{(k)} = P\underline{e}^{(k-1)} = PP\underline{e}^{(k-2)} = P^2 e^{(k-2)} = \ldots = P^k \underline{e}^{(0)}$$

ove $\underline{e}^{(0)} = \underline{x}^* - \underline{x}^{(0)}$ è l'errore iniziale, il quale dipende esclusivamente dal vettore di innesco $x^{(0)}$ scelto per l'applicazione del metodo.

Si può quindi introdurre il seguente risultato di convergenza.

Teorema 8.1 ***Condizione necessaria e sufficiente di convergenza***
Il metodo iterativo converge alla soluzione $\underline{x}^*$, ovvero

$$\lim_{k\to+\infty} e^{(k)} = 0,$$

se e solo se

$$\lim_{k\to+\infty} P^{(k)} = 0 \text{ (matrice nulla)}$$

ovvero se e solo se

$$\rho(P) < 1$$

ove $\rho(P) = \max_{i=1,\ldots,n} |\lambda_i(P)|$ denota il raggio spettrale della matrice di iterazione.

A parziale dimostrazione di questo teorema, possiamo considerare il caso in cui la matrice P sia diagonalizzabile, ossia nel caso in cui esista una matrice S tale che valga la relazione

$$P = S\Lambda_P S^{-1}$$

con $\Lambda_P = \text{diag}(\lambda_1(P), \ldots, \lambda_n(P))$ matrice diagonale con diagonale data dagli autovalori della matrice P.
Si ha che

$$\begin{aligned} P^k &= \overbrace{(S\Lambda_P S^{-1})(S\Lambda_P S^{-1})(S\Lambda_P S^{-1})\ldots(S\Lambda_P S^{-1})}^{k \text{ volte}} \\ &= S\Lambda_P \underbrace{(S^{-1}S)}_{=I} \Lambda_P \underbrace{(S^{-1}S)}_{=I} \Lambda_P \underbrace{(S^{-1}S)}_{=I} \ldots \underbrace{(S^{-1}S)}_{=I} \Lambda_P S^{-1} \\ &= S\Lambda_P^k S^{-1} \end{aligned}$$

ove

$$\Lambda_P^k = \text{diag}(\lambda_1^k(P), \ldots, \lambda_n^k(P)).$$

Quindi

$$\lim_{k\to+\infty} P^k = \lim_{k\to+\infty} S\Lambda_P^k S^{-1} = S(\lim_{k\to+\infty} \Lambda_P^k)S^{-1}$$

e chiaramente

$$\lim_{k\to+\infty} \Lambda_P^k = 0$$

se e solo se

$$\lim_{k\to+\infty} \lambda_i(P)^k = 0 \text{ per ogni } i = 1, \ldots, n$$

ovvero se e solo se

$$|\lambda_i(P)| < 1 \text{ per ogni } i = 1, \ldots, n$$

ovvero se e solo se

$$\rho(P) = \max_{i=1,\ldots,n} |\lambda_i(P)| < 1$$

da cui la tesi.

Poiché per ogni P e per ogni $\|\cdot\|$ norma matriciale indotta vale che

$$\rho(P) \leq \|P\|$$

vale pure la seguente condizione sufficiente di convergenza.

Teorema 8.2 ***Condizione sufficiente di convergenza***
Se esiste $\|\cdot\|$ norma matriciale indotta tale che

$$\|P\| < 1,$$

allora il metodo iterativo converge.

Ora, il calcolo esplicito della matrice di iterazione P può essere estremamente "scomodo". In effetti esistono classi di matrici, di interesse nelle applicazioni, per cui è possibile garantire la convergenza di un metodo, senza calcolarne esplicitamente la matrice di iterazione.

Teorema 8.3 ***Condizione sufficiente di convergenza***
Se la matrice A è diagonalmente dominante in senso stretto, ossia per ogni $i = 1, \ldots, n$

$$|a_{ii}| > \sum_{j=1, j\neq i}^{n} |a_{ij}|$$

allora i metodi di Jacobi e di Gauss-Seidel convergono.

Esempio 8.1 *La matrice*

$$A = \begin{bmatrix} 4 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 4 \end{bmatrix}$$

è diagonalmente dominante in senso stretto.

Teorema 8.4 ***Condizione sufficiente di convergenza***
Se la matrice A è simmetrica definita positiva allora il metodo di Gauss-Seidel converge e il metodo SOR converge per ogni $\omega \in (0, 2)$.

Test di arresto
Affrontiamo infine la questione relativa alla scelta dell'iterazione $\overline{k}$ a cui arrestare la generazione della successione su calcolatore, ossia la scelta del cosiddetto test di arresto.
La scelta più naturale per il test di arresto a prima vista potrebbe sembrare la seguente. Posto $\underline{r}^{(k)} = \underline{b} - A\underline{x}^{(k)}$, residuo al passo k, si richiede

$$\|\underline{r}^{(k)}\| \leq \varepsilon$$

Tale criterio di arresto viene detto criterio di arresto del residuo, ossia ci si chiede di quanto l'iterata $\underline{x}^{(k)}$ al passo k non soddisfi la relazione $\underline{b} - A\underline{x} = \underline{0}$ e trattandosi di un vettore, se ne considera la norma (si ricordi che la scelta della norma non ha una particolare rilevanza in virtù della proprietà di equivalenza topologica delle norme).
Si può dimostrare che tale criterio di arresto del residuo controlla la norma dell'errore assoluto in accordo alla seguente disuguaglianza

$$\|\underline{e}^{(k)}\| \leq \frac{K(A)}{\|A\|}\|\underline{r}^{(k)}\|,$$

ove $K(A) = \|A\|\|A^{-1}\|$ è detto numero di condizionamento della matrice A.
Evidentemente tale criterio fornisce un'informazione corretta quando $K(A)$ è piccolo, in quanto a norma di residuo piccolo corrisponde in buona sostanza una norma di errore assoluto piccolo. Di converso, non è un buon criterio di arresto nel caso in cui la matrice A sia mal condizionata, in quanto se il fattore di amplificazione $K(A)$ è elevato, un residuo piccolo non garantisce un errore assoluto piccolo.
Su calcolatore, un test più naturale è in realtà il seguente

$$\|\underline{x}^{(k+1)} - \underline{x}^{(k)}\| \leq \varepsilon$$

detto criterio di arresto dell'incremento. Misurando la differenza fra due successive iterate, il test dà l'indicazione di quante cifre coincidono nelle due successive approssimazioni e quindi quante cifre esatte sono già state trovate della soluzione.
Tale criterio di arresto controlla la norma dell'errore assoluto in accordo alla seguente disuguaglianza

$$\|\underline{e}^{(k)}\| \leq \frac{\|P\|}{1 - \|P\|}\|\underline{x}^{(k+1)} - \underline{x}^{(k)}\|,$$

ove P è la matrice di iterazione del metodo considerato.
Si tenga presente che anche questo criterio può non essere un buon criterio di arresto. Infatti, nel caso in cui la norma della matrice P sia prossima ad 1, essendo il fattore $\|P\|/(1 - \|P\|)\|$ elevato, un incremento di norma piccola non garantisce necessariamente un errore assoluto di norma piccola.

9 Metodi diretti per la risoluzione di sistemi lineari: fattorizzazione $PA = LU$

9.1 Il metodo di Gauss

Come si è visto nella sezione 3.3, per la risoluzione di un sistema lineare si può considerare al posto del metodo di Cramer, troppo costoso dal punto di vista computazionale, il metodo di Gauss.
Tale metodo si basa sostanzialmente sulla nozione di sistemi lineari equivalenti (si veda definizione 3.4) e invoca la proprietà enunciata nella Proposizione 3.3: sostituendo alla $j-$sima equazione una combinazione lineare delle equazioni del sistema, con il solo vincolo che il coefficiente corrispondente nella combinazione sia diverso da 0, si ottiene un sistema lineare equivalente, ossia con tutte e sole le soluzioni del sistema di partenza.
Ora, un'applicazione ripetuta e mirata di tale proprietà permette di trasformare un generico sistema lineare

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

con $A \in \mathbb{R}^{n\times n}, x, b \in \mathbb{R}^n$ e tale che $\det(A) \neq 0$, in un sistema lineare equivalente del tipo

$$U\underline{x} = \underline{c}$$

con U matrice triangolare superiore. Tale trasformazione è particolarmente vantaggiosa in quanto il sistema

$$U\underline{x} = \underline{c}$$

può facilmente essere risolto su calcolatore con la cosiddetta procedura di risoluzione a ritroso (backward)

$$x_i = \left(c_i - \sum_{j=i+1}^{n} u_{ij}x_j \right) / u_{ii}, \quad i = n, n-1, \ldots, 2, 1.$$

Avendo già visto in sezione 3.3 il metodo applicato ad un esempio, vediamo ora di formalizzarlo su un generico sistema lineare con $n = 4$.
Sia

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} \text{ e } \underline{b} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix},$$

I passo: si vuole eliminare l'incognita x_1 nella seconda, terza e quarta equazione, ossia si vuole azzerare tutta la prima sottocolonna della matrice A a partire dall'elemento in posizione $(2, 1)$. Supposto che sia $a_{11} \neq 0$, per ottenere questo risultato è sufficiente sottrarre alla seconda equazione la prima equazione moltiplicata per $m_{21} = a_{21}/a_{11}$, alla terza equazione la prima equazione moltiplicata per $m_{31} = a_{31}/a_{11}$ e alla quarta equazione la prima equazione moltiplicata per $m_{41} = a_{41}/a_{11}$.
Infatti vale

$$a_{21} - m_{21}a_{11} = 0, \quad a_{31} - m_{31}a_{11} = 0, \quad a_{41} - m_{41}a_{11} = 0.$$

Chiaramente si modificheranno pure i restanti coefficienti di tali equazioni e indicando con l'apice $^{(1)}$ i coefficienti modificati con questo primo passo, si ha che per la seconda equazione vale

$$a_{22}^{(1)} = a_{22} - m_{21}a_{12}, \quad a_{23}^{(1)} = a_{23} - m_{21}a_{13}, \quad a_{24}^{(1)} = a_{24} - m_{21}a_{14}$$

e

$$b_2^{(1)} = b_2 - m_{21}b_1,$$

ovvero, in forma compatta, con l'indice j che varia sulle incognite delle equazioni,

$$a_{2j}^{(1)} = a_{2j} - m_{21}a_{1j} \; j = 2, \ldots, 4, \quad b_2^{(1)} = b_2 - m_{21}b_1.$$

Analogamente, per i coefficienti delle rimanenti due equazioni vale

$$\begin{aligned} a_{3j}^{(1)} &= a_{3j} - m_{31}a_{1j} \; j = 2, \ldots, 4, \quad b_3^{(1)} = b_3 - m_{31}b_1 \\ a_{4j}^{(1)} &= a_{4j} - m_{41}a_{1j} \; j = 2, \ldots, 4, \quad b_4^{(1)} = b_4 - m_{41}b_1. \end{aligned}$$

Riassumendo, il primo passo di fattorizzazione comporta le seguenti trasformazioni

$$\left[\begin{array}{l} \text{per } i = 2, 4 \quad \text{indice equazione} \\ \quad m_{i1} = a_{i1}/a_{11} \\ \quad \left[\begin{array}{l} \text{per } j = 2, 4 \quad \text{indice incognita} \\ \quad\quad a_{ij}^{(1)} = a_{ij} - m_{i1}a_{1j} \end{array} \right. \\ \quad b_i^{(1)} = b_i - m_{i1}b_1, \end{array} \right.$$

mentre la prima equazione rimane invariata.
Si ha quindi il sistema lineare equivalente $A^{(1)}\underline{x} = \underline{b}^{(1)}$, con

$$A^{(1)} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & a_{24}^{(1)} \\ 0 & a_{32}^{(1)} & a_{33}^{(1)} & a_{34}^{(1)} \\ 0 & a_{42}^{(1)} & a_{43}^{(1)} & a_{44}^{(1)} \end{bmatrix} \text{ e } \underline{b}^{(1)} = \begin{bmatrix} b_1 \\ b_2^{(1)} \\ b_3^{(1)} \\ b_4^{(1)} \end{bmatrix}$$

II passo: si vuole eliminare l'incognita x_2 nella terza e quarta equazione, ossia si vuole azzerare tutta la seconda sottocolonna della matrice $A^{(1)}$ a partire dall'elemento in posizione $(3, 2)$. Supposto che sia $a_{22}^{(1)} \neq 0$, per ottenere questo risultato è sufficiente sottrarre alla terza equazione la seconda equazione moltiplicata per $m_{32} = a_{32}^{(1)}/a_{22}^{(1)}$ e alla quarta equazione la seconda equazione moltiplicata per $m_{42} = a_{42}^{(1)}/a_{22}^{(1)}$.
Infatti vale

$$a_{32}^{(1)} - m_{32}a_{22}^{(1)} = 0 \text{ e } a_{42}^{(1)} - m_{42}a_{22}^{(1)} = 0$$

Quindi, il secondo passo di fattorizzazione comporta le seguenti trasformazioni sulla terza e quarta equazione

$$\left[\begin{array}{l} \text{per } i = 3, 4 \quad \text{indice equazione} \\ \quad m_{i2} = a_{i2}^{(1)}/a_{22}^{(1)} \\ \quad \left[\begin{array}{l} \text{per } j = 3, 4 \quad \text{indice incognita} \\ \quad\quad a_{ij}^{(2)} = a_{ij}^{(2)} - m_{i2}a_{2j}^{(1)} \end{array} \right. \\ \quad b_i^{(2)} = b_i^{(1)} - m_{i2}b_2^{(1)}, \end{array} \right.$$

mentre la prime due equazioni rimangono invariate.
Si ha quindi il sistema lineare equivalente con

$$A^{(2)} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & a_{24}^{(1)} \\ 0 & 0 & a_{33}^{(2)} & a_{34}^{(2)} \\ 0 & 0 & a_{43}^{(2)} & a_{44}^{(2)} \end{bmatrix} \text{ e } \underline{b}^{(2)} = \begin{bmatrix} b_1 \\ b_2^{(1)} \\ b_3^{(2)} \\ b_4^{(2)} \end{bmatrix}$$

III passo: si vuole eliminare l'incognita x_3 nella quarta equazione, ossia si vuole azzerare tutta la terza sottocolonna della matrice $A^{(2)}$ a partire dall'elemento in posizione $(4,3)$. Supposto che sia $a_{33}^{(2)} \neq 0$, per ottenere questo risultato è sufficiente sottrarre alla quarta equazione la terza equazione moltiplicata per $m_{43} = a_{43}^{(2)}/a_{33}^{(2)}$.
Infatti vale

$$a_{43}^{(2)} - m_{43} a_{33}^{(2)} = 0$$

e il terzo passo di fattorizzazione comporta le seguenti trasformazioni

$$\left[\begin{array}{l} \text{per } i = 4 \quad \text{indice equazione} \\ \quad m_{i3} = a_{i3}^{(2)}/a_{33}^{(2)} \\ \quad \left[\begin{array}{l} \text{per } j = 4 \quad \text{indice incognita} \\ \quad\quad a_{ij}^{(3)} = a_{ij}^{(2)} - m_{i3} a_{3j}^{(2)} \end{array} \right. \\ \quad b_i^{(3)} = b_i^{(2)} - m_{i3} b_3^{(2)}, \end{array} \right.$$

mentre le prime tre equazioni rimangono invariate.
Si ha quindi il sistema lineare equivalente con

$$A^{(3)} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & a_{24}^{(1)} \\ 0 & 0 & a_{33}^{(2)} & a_{34}^{(2)} \\ 0 & 0 & 0 & a_{44}^{(3)} \end{bmatrix} \text{ e } \underline{b}^{(3)} = \begin{bmatrix} b_1 \\ b_2^{(1)} \\ b_3^{(2)} \\ b_4^{(3)} \end{bmatrix}$$

ove vale, evidentemente, $U = A^{(3)}$ e $\underline{c} = \underline{b}^{(3)}$.

Ora, detto k il passo di eliminazione, si può osservare che nei tre gruppi di trasformazioni sopra riportate, i numeri in rosso valgono esattamente k (indicando il passo o l'equazione di riferimento nelle combinazioni lineari), i numeri in verde valgono esattamente $k+1$ (indicando l'indice dell'equazione o dell'incognita da cui si parte nelle trasformazioni) e i numeri in blu valgono esattamente $k-1$ (indicando gli elementi al passo precedente). Inoltre, i passi di fattorizzazione sono $n-1$, ove n è la dimensione del sistema poiché dopo il passo $n-1$ l'ultima equazione contiene solo l'incognita x_n. Quindi, posto $A^{(0)} = A$, si può riscrivere in forma compatta

$$\left[\begin{array}{l} \text{per } k = 1, n-1 \quad \text{passo di eliminazione} \\ \quad \left[\begin{array}{l} \text{per } i = k+1, n \quad \text{indice equazione} \\ \quad\quad m_{ik} = a_{ik}^{(k-1)}/a_{kk}^{(k-1)} \\ \quad\quad \left[\begin{array}{l} \text{per } j = k+1, n \quad \text{indice incognita} \\ \quad\quad\quad a_{ij}^{(k)} = a_{ij}^{(k-1)} - m_{ik} a_{kj}^{(k-1)} \end{array} \right. \\ \quad\quad b_i^{(k)} = b_i^{(k-1)} - m_{ik} b_k^{(k-1)} \end{array} \right. \end{array} \right.$$

ottenendo cosı` l'algoritmo del metodo di Gauss.
I costo in termini di operazioni moltiplicative di tale algoritmo è dato da

$$\sum_{k=1}^{n-1}\sum_{i=k+1}^{n}\left(1+\left(\sum_{j=k+1}^{n}1\right)+1\right)$$

ed è dell'ordine di $n^3/3$ operazioni, cui si aggiungono $n^2/2$ operazioni di tipo moltiplicativo per la procedura di risoluzione backward.

Infine, occorre sottolineare che nulla garantisce che l'assunzione $a_{kk}^{(k)} \neq 0$ per ogni $k = 1, \ldots, n-1$ sia verificata; rimandiamo alla sezione 9.3 la trattazione di questa circostanza.

9.2 La fattorizzazione $A = LU$

Vediamo ora una riformulazione del metodo precedente che comporta un vantaggio significativo nel caso si debbano risolvere più sistemi lineari con la stessa matrice A e differenti termini noti (un esempio significativo di applicazione è dato dal metodo della potenza inversa descritto in sezione 12).
L'idea è quella di separare il momento della trasformazione di A in U da quello della trasformazione del termine noto $\underline{b}$ in $\underline{c}$. Focalizziamo l'attenzione sulla matrice A.
Sia

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix}.$$

I passo: si vuole azzerare tutta la prima sottocolonna della matrice A a partire dall'elemento in posizione $(2,1)$. Supposto che sia $a_{11} \neq 0$, per ottenere questo risultato è sufficiente, posto $m_{21} = a_{21}/a_{11}$, $m_{31} = a_{31}/a_{11}$ e $m_{41} = a_{41}/a_{11}$, esattamente come nel metodo di Gauss, moltiplicare per la matrice $M^{(1)}$, triangolare inferiore con elementi sulla diagonale principale tutti uguali a 1, data da

$$M^{(1)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -m_{21} & 1 & 0 & 0 \\ -m_{31} & 0 & 1 & 0 \\ -m_{41} & 0 & 0 & 1 \end{bmatrix}$$

Si ha quindi

$$\begin{aligned} M^{(1)}A &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ -m_{21} & 1 & 0 & 0 \\ -m_{31} & 0 & 1 & 0 \\ -m_{41} & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} \\ &= \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & a_{24}^{(1)} \\ 0 & a_{32}^{(1)} & a_{33}^{(1)} & a_{34}^{(1)} \\ 0 & a_{42}^{(1)} & a_{43}^{(1)} & a_{44}^{(1)} \end{bmatrix} = A^{(1)} \end{aligned}$$

ove l'espressione del coefficienti $a_{ij}^{(1)},\ i,j=2,\ldots,4$ è la stessa riportata nella sezione 9.1 relativamente al metodo di Gauss.

II passo: si vuole azzerare tutta la seconda sottocolonna della matrice $A^{(1)}$ a partire dall'elemento in posizione $(3,2)$. Supposto che sia $a_{22}^{(1)} \neq 0$, per ottenere questo risultato è sufficiente, posto $m_{32} = a_{32}^{(1)}/a_{22}^{(1)}$ e $m_{42} = a_{42}^{(1)}/a_{22}^{(1)}$, esattamente come nel metodo di Gauss, moltiplicare per la matrice $M^{(2)}$, triangolare inferiore con elementi sulla diagonale principale tutti uguali a 1, data da

$$M^{(2)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -m_{32} & 1 & 0 \\ 0 & -m_{42} & 0 & 1 \end{bmatrix}$$

Si ha quindi

$$\begin{aligned} M^{(2)}A^{(1)} &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -m_{32} & 1 & 0 \\ 0 & -m_{42} & 0 & 1 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & a_{24}^{(1)} \\ 0 & a_{32}^{(1)} & a_{33}^{(1)} & a_{34}^{(1)} \\ 0 & a_{42}^{(1)} & a_{43}^{(1)} & a_{44}^{(1)} \end{bmatrix} \\ &= \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & a_{24}^{(1)} \\ 0 & 0 & a_{33}^{(2)} & a_{34}^{(2)} \\ 0 & 0 & a_{43}^{(2)} & a_{44}^{(2)} \end{bmatrix} = A^{(2)} \end{aligned}$$

ove l'espressione del coefficienti $a_{ij}^{(2)},\ i,j=3;4$ è la stessa riportata nella sezione 9.1 relativamente al metodo di Gauss.

III passo: si vuole azzerare tutta la terza sottocolonna della matrice $A^{(2)}$ a partire dall'elemento in posizione $(4,3)$. Supposto che sia $a_{33}^{(2)} \neq 0$, per ottenere questo risultato è sufficiente, posto $m_{43} = a_{43}^{(2)}/a_{33}^{(2)}$, esattamente come nel metodo di Gauss, moltiplicare per la matrice $M^{(3)}$, triangolare inferiore con elementi sulla diagonale principale tutti uguali a 1, data da

$$M^{(3)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -m_{43} & 1 \end{bmatrix}$$

Si ha quindi

$$\begin{aligned} M^{(3)}A^{(2)} &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -m_{43} & 1 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & a_{24}^{(1)} \\ 0 & 0 & a_{33}^{(2)} & a_{34}^{(2)} \\ 0 & 0 & a_{43}^{(2)} & a_{44}^{(2)} \end{bmatrix} \\ &= \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & a_{24}^{(1)} \\ 0 & 0 & a_{33}^{(2)} & a_{34}^{(2)} \\ 0 & 0 & 0 & a_{44}^{(3)} \end{bmatrix} = A^{(3)} = U \end{aligned}$$

ove l'espressione del coefficiente $a_{44}^{(3)}$ è la stessa riportata nella sezione 9.1 relativamente al metodo di Gauss.

Ora, chiaramente vale che

$$U = A^{(3)}$$

in quanto le trasformazioni effettuate sono le medesime del metodo di Gauss; di più, ripercorrendo a ritroso le trasformazioni effettuate, si ha che

$$M^{(3)}A^{(2)} = M^{(3)}M^{(2)}A^{(1)} = M^{(3)}M^{(2)}M^{(1)}A = U$$

ossia

$$M^{(3)}M^{(2)}M^{(1)}A = U$$

da cui

$$\begin{aligned} A &= \left(M^{(3)}M^{(2)}M^{(1)}\right)^{-1} U \\ &= \left(M^{(1)}\right)^{-1}\left(M^{(2)}\right)^{-1}\left(M^{(3)}\right)^{-1} U \end{aligned}$$

ove le matrici $M^{(k)}$ sono invertibili in quanto $\det(M^{(k)}) = 1 \neq 0$.
In virtù della particolare struttura delle matrici $M^{(k)}$, vale che

$$\left(M^{(1)}\right)^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ m_{21} & 1 & 0 & 0 \\ m_{31} & 0 & 1 & 0 \\ m_{41} & 0 & 0 & 1 \end{bmatrix}$$

$$\left(M^{(2)}\right)^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & m_{32} & 1 & 0 \\ 0 & m_{42} & 0 & 1 \end{bmatrix}$$

$$\left(M^{(3)}\right)^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & m_{43} & 1 \end{bmatrix}$$

ossia per ottenere le matrici inverse è sufficiente cambiare di segno ai coefficienti della triangolare inferiore stretta. Quindi, si ha che

$$\left(M^{(1)}\right)^{-1}\left(M^{(2)}\right)^{-1}\left(M^{(3)}\right)^{-1} =$$

$$= \begin{bmatrix} 1 & 0 & 0 & 0 \\ m_{21} & 1 & 0 & 0 \\ m_{31} & 0 & 1 & 0 \\ m_{41} & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & m_{32} & 1 & 0 \\ 0 & m_{42} & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & m_{43} & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & 0 & 0 \\ m_{21} & 1 & 0 & 0 \\ m_{31} & 0 & 1 & 0 \\ m_{41} & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & m_{32} & 1 & 0 \\ 0 & m_{42} & m_{43} & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & 0 & 0 \\ m_{21} & 1 & 0 & 0 \\ m_{31} & m_{32} & 1 & 0 \\ m_{41} & m_{42} & m_{43} & 1 \end{bmatrix} = L$$

ossia la matrice prodotto risulta essere una matrice triangolare inferiore con tutti gli elementi della diagonale principale pari a 1 e si ottiene semplicemente giustapponendo nell'ordine le colonne "significative" delle matrici $M^{(k)}$.
Pertanto, tutte le informazioni necessarie a trasformare la matrice A nella matrice U sono contenute nella matrice L e sono le medesime trasformazioni necessarie per trasformare il termine noto $\underline{b}$ in $\underline{c}$. Infatti, cosı` come vale

$$U = L^{-1}A,$$

vale pure

$$\underline{c} = L^{-1}\underline{b},$$

in quanto basta pensare di applicare la procedura sopra riportata non alla sola matrice A, ma al sistema di equazioni $A\underline{x} = \underline{b}$, per ottenere

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

e quindi le relazioni di cui sopra.
Risulta cosı` dimostrata l'equivalenza del metodo di Gauss con il metodo della fattorizzazione LU, che quindi consiste nel calcolare la fattorizzazione $A = LU$, nel calcolare il termine noto trasformato $\underline{c}$ risolvendo il sistema

$$L\underline{c} = \underline{b}$$

e calcolare la soluzione $\underline{x}$ risolvendo il sistema

$$U\underline{x} = \underline{c}.$$

Il primo sistema lineare con matrice triangolare inferiore si risolve con la procedura di risoluzione in avanti (forward)

$$c_i = \left(b_i - \sum_{j=1}^{i-1} l_{ij}c_j \right) / l_{ii}, \quad i = 1, 2, \ldots, n-1, n,$$

tenendo presente che $l_{ii} = 1$ per ogni $i = 1, \ldots, n$ e il secondo sistema lineare con matrice triangolare superiore si risolve con la procedura di risoluzione all'indietro (backward)

$$x_i = \left(c_i - \sum_{j=i+1}^{n} u_{ij}x_j \right) / u_{ii}, \quad i = n, n-1, \ldots, 2, 1.$$

Il costo computazionale in termini di operazioni di tipo moltiplicativo è dell'ordine di $n^2/2$ per ciascuna delle due procedure di risoluzione.
In una prospettiva diversa, legata esclusivamente all'idea della fattorizzazione $A = LU$, si ha che, una volta nota la fattorizzazione, la si applica al sistema lineare $A\underline{x} = \underline{b}$ ottenendo

$$LU\underline{x} = \underline{b},$$

e, avendo posto $\underline{c} = U\underline{x}$, si risolvono in sequenza i due sistemi

$$\begin{aligned} L\underline{c} &= \underline{b} \\ U\underline{x} &= \underline{c} \end{aligned}$$

9.3 La fattorizzazione $PA = LU$

In ultimo, affrontiamo la questione dell'assunzione $a_{kk}^{(k)} \neq 0$ per ogni $k = 1, \ldots, n-1$. La questione non è trascurabile perché l'ipotesi $\det(A) \neq 0$ non è sufficiente a garantirla.
Ora, in linea teorica, per poter procedere nella fattorizzazione sarebbe sufficiente considerare un qualsivoglia scambio di righe che porti a soddisfare l'ipotesi richiesta. Tuttavia, nella pratica, si ricorre alla seguente strategia detta **strategia di pivot parziale per righe**.
Ad ogni passo di fattorizzazione, prima di procedere alle trasformazioni prescritte, si scambia la riga k con la riga $ipiv$ tale che

$$|a_{ipiv\ k}^{(k-1)}| = \max_{i=k,\ldots,n} |a_{ik}^{(k-1)}|$$

ossia, al passo k, si cerca l'elemento di modulo massimo nella sottocolonna a partire dalla posizione (k, k) e si scambia tale riga con la riga $k-$sima.
Il sistema ottenuto è equivalente in virtù della proprietà enunciata nella Proposizione 3.2, in quanto lo scambiare due equazioni fra loro permette di ottenere un sistema lineare equivalente, ossia con tutte e sole le soluzioni del sistema di partenza.
Si tenga inoltre presente che nella pratica tale strategia vine applicata comunque ad ogni passo e non solo in caso di necessità.
Ora, per poter tenere conto in modo formale di tali scambi nel procedimento della fattorizzazione LU occorre introdurre la nozione di matrice di permutazione.

Definizione 9.1 *Matrice di permutazione*
Una matrice

$$P^{(ij)} = \begin{bmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & 1 & & & & & & \\ & & & 0 & & 1 & & & \leftarrow \text{riga } i \\ & & & & 1 & & & & \\ & & & & & \ddots & & & \\ & & & & & & 1 & & \\ & & & 1 & & 0 & & & \leftarrow \text{riga } j \\ & & & & & & 1 & & \\ & & & & & & & \ddots & \\ & & & & & & & & 1 \end{bmatrix}$$

ove ogni elemento non esplicitamente indicato è pari a zero, è detta matrice di permutazione. La matrice $P^{(ij)}$ è l'esempio più semplice di matrice di permutazione ed è ottenuta dalla matrice identica scambiando la riga i con la riga j. Tale matrice è tale che $P^{(ij)}A$ scambia fra loro la riga i con la riga j della matrice A, lasciando invariate le altre, e $AP^{(ij)}$ scambia fra loro la colonna i con la colonna j della matrice A, lasciando invariate le altre. Di più, $P^{(ij)}P^{(ij)} = I$, ossia, per l'unicità della matrice inversa, $(P^{(ij)})^{-1} = P^{(ij)}$

Ora, ritornando alla procedura di fattorizzazione precedente si ha che al primo passo la fattorizzazione viene applicata alla matrice $P_1 A$ anziché alla matrice

A, ove P_1 è la matrice di permutazione che scambia la prima riga con la riga relativa all'elemento di modulo massimo della prima colonna. Analogamente, al secondo passo la fattorizzazione viene applicata alla matrice $P_2A^{(1)}$ anziché alla matrice $A^{(1)}$, ove P_2 è la matrice di permutazione che scambia la seconda riga con la riga relativa all'elemento di modulo massimo della seconda sottocolonna a partire dall'elemento in posizione $(2,2)$ e al terzo passo la fattorizzazione viene applicata alla matrice $P_3A^{(2)}$ anziché alla matrice $A^{(2)}$, ove P_3 è la matrice di permutazione che scambia la terza riga con la riga relativa all'elemento di modulo massimo della terza sottocolonna a partire dall'elemento in posizione $(3,3)$. Chiaramente, se non è stato effettuato alcuno scambio la matrice di permutazione sarà la matrice identica.
Quindi, si ha

$$\begin{aligned} U &= M^{(3)}(P_3A^{(2)}) \\ &= M^{(3)}(P_3M^{(2)}(P_2A^{(1)})) \\ &= M^{(3)}(P_3M^{(2)}(P_2M^{(1)}(P_1A))) \\ &= M^{(3)}P_3M^{(2)}P_2M^{(1)}P_1A \end{aligned}$$

Tenuto conto che, come enunciato nella definizione, vale che

$$P_1P_1 = I, \quad P_2P_2 = I, \quad P_3P_3 = I$$

e quindi si può riscrivere

$$\begin{aligned} U &= M^{(3)}P_3M^{(2)}P_2M^{(1)}(P_2P_2)P_1A \\ &= M^{(3)}P_3M^{(2)}(P_2M^{(1)}P_2)P_2P_1A \\ &= M^{(3)}P_3M^{(2)}\tilde{M}^{(1)}P_2P_1A \quad \text{con } \tilde{M}^{(1)} = P_2M^{(1)}P_2 \\ &= M^{(3)}P_3M^{(2)}(P_3P_3)\tilde{M}^{(1)}(P_3P_3)P_2P_1A \\ &= M^{(3)}(P_3M^{(2)}P_3)(P_3\tilde{M}^{(1)}P_3)P_3P_2P_1A \\ &= M^{(3)}\tilde{\tilde{M}}^{(2)}\tilde{\tilde{M}}^{(1)}P_3P_2P_1A \quad \text{con } \tilde{\tilde{M}}^{(1)} = P_3\tilde{M}^{(1)}P_3,\ \tilde{\tilde{M}}^{(2)} = P_3M^{(2)}P_3 \end{aligned}$$

Ora, tenuto conto del tipo di matrici di permutazione che è possibile applicare, le matrici $\tilde{\tilde{M}}^{(1)}$, $\tilde{M}^{(2)}$ sono ancora matrici di tipo M, cosicché si può ripetere il procedimento precedentemente visto per determinare la matrice L e vale

$$L = \left(\tilde{M}^{(1)}\right)^{-1}\left(\tilde{\tilde{M}}^{(2)}\right)^{-1}\left(M^{(3)}\right)^{-1}$$

Inoltre, la matrice

$$P = P_3P_2P_1,$$

prodotto di matrici di permutazione è ancora una matrice di permutazione e tiene conto di tutti gli scambi effettuati nei vari passi di fattorizzazione. È possibile dimostrare che la tecnica dello scambio di righe, comunque scelto, garantisce la rimozione di eventuali elementi $a_{kk}^{(k)}$ nulli (anche se essi potrebbero essere numericamente piccoli). Infatti vale il seguente teorema.

Teorema 9.1 *Sia $A \in \mathbb{R}^{n\times n}$. Esiste una matrice di permutazione P tale che si può ottenere la fattorizzazione LU, ossia $PA = LU$.*

Di più, fra le molteplici strategie di scambio possibili, la strategia del pivot parziale per righe ha un'effetto stabilizzante sul procedimento di fattorizzazione in quanto

$$\begin{aligned} |l_{ik}| &= \frac{|a_{ik}^{(k)}|}{\max_{j=k,\ldots,n} |a_{jk}^{(k)}|} \leq 1 \quad i = k+1, \ldots, n \\ a_M^{(n-1)} &\leq 2^{n-1} a_M^{(1)} \end{aligned}$$

ove $a_M^{(k)}$ è l'elemento di modulo massimo al passo k, ovvero gli elementi della L sono tutti di modulo minore o uguale a 1 e si ha pure una previsione della massima crescita degli elementi di $U = A^{(n-1)}$.

Concludendo, l'algoritmo della fattorizzazione $PA = LU$ in forma compatta è dato da

per $k = 1, n-1$

calcolo $a_{ipiv\ k}^{(k-1)} = \max_{i=k,\ldots,n} |a_{ik}^{(k-1)}|$
se $ipiv \neq k$ scambia riga k con riga $ipiv$

per $i = k+1, n$
$m_{ik} = a_{ik}^{(k-1)} / a_{kk}^{(k-1)}$
per $j = k+1, n$
$a_{ij}^{(k)} = a_{ij}^{(k-1)} - m_{ik} a_{kj}^{(k-1)}$

Chiaramente, prima di andare ad effettuare la risoluzione dei due sistemi lineari, occorrerà permutare opportunamente il termine noto $\underline{b}$, in quanto la fattorizzazione $PA = LU$ si applica non al sistema $A\underline{x} = \underline{b}$, ma al sistema $PA\underline{x} = P\underline{b}$. Quindi, si risolveranno

$$\begin{aligned} L\underline{c} &= P\underline{b} \\ U\underline{x} &= \underline{c}. \end{aligned}$$

9.4 Confronto fra i metodi diretti e metodi iterativi

Come visto in sezione 7.2, affrontando il calcolo della soluzione di un problema su calcolatore, si hanno i seguenti tre tipi di errore.

Errore inerente. A causa dell'errore di rappresentazione sui dati anziché risolvere il sistema lineare

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

si risolve su calcolatore in realtà il sistema lineare perturbato

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

ove $A + \delta A$ e $b + \delta b$ hanno come elementi i numeri macchina con cui sono stati approssimati i corrispondenti elementi con un errore maggiorato dalla precisione di macchina. Si può dimostrare che vale che

$$e_x \leq \frac{K(A)}{1 - K(A)e_A}(e_A + e_b)$$

ove

$$\begin{aligned} e_x &= \|\underline{\delta x}\|/\|\underline{x}\| \quad \text{errore relativo su } \underline{x} \\ e_A &= \|\delta A\|/\|A\| \quad \text{errore relativo su } \underline{x} \\ e_b &= \|\underline{\delta b}\|/\|\underline{b}\| \quad \text{errore relativo su } \underline{b} \end{aligned}$$

e la quantità

$$K(A) = \|A\|\|A^{-1}\| \geq 1$$

è detta numero di condizionamento della matrice A.
Il numero di condizionamento misura la "sensibilità" del problema agli errori sui dati: se $K(A)$ è elevato l'errore relativo sulla soluzione può essere elevato per quanto e_a e e_b siano molto piccoli. La stima è comunque pessimistica in quanto vale per ogni possibile termine noto $\underline{b}$.
Si ricordi che l'errore inerente è indipendente dal metodo utilizzato per la risoluzione.

Errore analitico. I metodi diretti in aritmetica esatta danno soluzione esatta; non si ha quindi errore analitico.
Nel caso dei metodi iterativi occorre invece troncare la generazione della successione di vettori tramite un opportuno criterio di arresto cosı` da approssimare il limite e dando cosiòrigine ad un errore analitico.

Errore algoritmico. Entrambi i metodi sono ovviamente soggetti all'errore algoritmico. Tuttavia, come già ricordato, i metodi iterativi per loro natura sono meno sensibili all'errore algoritmico.
D'altro canto, la tecnica del pivot parziale per righe ha un effetto stabilizzante: la questione è comunque legata alla crescita degli elementi di U.
Si tenga presente che la fattorizzazione LU senza pivot è stabile se A è diagonalmente dominante in senso stretto o simmetrica definita positiva.

Costi computazionali Si distingue fra il costo di memorizzazione e il costo in termini di operazioni moltiplicative.
Per quanto concerne il costo di memorizzazione la differenza può essere significativa nel caso delle matrici sparse (matrici il cui numero di elementi diversi da zero è $O(n)$ anziché n^2). Infatti le operazioni di fattorizzazione possono far aumentare il numero di elementi non nulli, dando luogo alla necessità di predisporre altro spazio di memoria per la loro memorizzazione; mentre i metodi iterativi visti non alterano la struttura della matrice ed eventualmente può essere caricata in memoria una sola riga di A per volta.
Per quanto concerne il costo in termini di operazioni di tipo moltiplicativo si ha che è dell'ordine di $n^3/3$ per la fattorizzazione LU e pari a n^2·numero di iterazioni per i metodi iterativi. Nel caso di convergenza in meno di n passi il vantaggio può essere rilevante.

9.5 Esempi

Esempio 1. Si vuole calcolare la decomposizione $A = LU$ delle seguenti matrici 4×4

$$A = \begin{bmatrix} 3 & 1 & 1 & 1 \\ 2 & 1 & 0 & 0 \\ 2 & 0 & 1 & 0 \\ 2 & 0 & 0 & 1 \end{bmatrix} \text{ e } \quad B = \begin{bmatrix} 1 & 0 & 0 & 2 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 2 \\ 1 & 1 & 1 & 3 \end{bmatrix}.$$

- Passo 1: $m_{21} = 2/3$, $m_{31} = 2/3$, $m_{41} = 2/3$, quindi

$$M_1 A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -2/3 & 1 & 0 & 0 \\ -2/3 & 0 & 1 & 0 \\ -2/3 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 3 & 1 & 1 & 1 \\ 2 & 1 & 0 & 0 \\ 2 & 0 & 1 & 0 \\ 2 & 0 & 0 & 1 \end{bmatrix} = \left[\begin{array}{cccc} 3 & 1 & 1 & 1 \\ \hline 0| & 1/3 & -2/3 & -2/3 \\ 0| & -2/3 & 1/3 & -2/3 \\ 0| & -2/3 & -2/3 & 1/3 \end{array}\right]$$

  Passo 2: $m_{32} = -2$, $m_{33} = -2$, quindi

$$M_2(M_1 A) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 2 & 1 & 0 \\ 0 & 2 & 0 & 1 \end{bmatrix} \begin{bmatrix} 3 & 1 & 1 & 1 \\ 0 & 1/3 & -2/3 & -2/3 \\ 0 & -2/3 & 1/3 & -2/3 \\ 0 & -2/3 & -2/3 & 1/3 \end{bmatrix} = \left[\begin{array}{cccc} 3 & 1 & 1 & 1 \\ 0 & 1/3 & -2/3 & -2/3 \\ \hline 0 & 0| & -1 & -2 \\ 0 & 0| & -2 & -1 \end{array}\right]$$

  Passo 3: $m_{43} = 2$, quindi

$$M_3(M_2 M_1 A) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -2 & 1 \end{bmatrix} \begin{bmatrix} 3 & 1 & 1 & 1 \\ 0 & 1/3 & -2/3 & -2/3 \\ 0 & 0 & -1 & -2 \\ 0 & 0 & -2 & -1 \end{bmatrix} = \left[\begin{array}{cccc} 3 & 1 & 1 & 1 \\ 0 & 1/3 & -2/3 & -2/3 \\ 0 & 0 & -1 & -2 \\ \hline 0 & 0 & 0 & |3 \end{array}\right]$$

  Quindi si ha

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2/3 & 1 & 0 & 0 \\ 2/3 & -2 & 1 & 0 \\ 2/3 & -2 & 2 & 1 \end{bmatrix} \text{ e } U = \begin{bmatrix} 3 & 1 & 1 & 1 \\ 0 & 1/3 & -2/3 & -2/3 \\ 0 & 0 & -1 & -2 \\ 0 & 0 & 0 & 3 \end{bmatrix}$$

  La posizione degli elementi non nulli della matrice A (ovvero il tipo di *pattern*), fa sì che si abbia il fenomeno del *fill-in*, ossia l'introduzione di nuovi elementi non nulli a causa delle operazioni di fattorizzazione, fino a rendere completamente piene sia la matrice L, sia la matrice U.

- Passo 1: $m_{21} = 0$, $m_{31} = 0$, $m_{41} = 1$, quindi

$$M_1 B = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 2 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 2 \\ 1 & 1 & 1 & 3 \end{bmatrix} = \left[\begin{array}{cccc} 1 & 0 & 0 & 2 \\ \hline 0| & 1 & 0 & 2 \\ 0| & 0 & 1 & 2 \\ 0| & 1 & 1 & 1 \end{array}\right]$$

  Passo 2: $m_{32} = 0$, $m_{33} = -1$, quindi

$$M_2(M_1 B) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 2 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 2 \\ 1 & 1 & 1 & 1 \end{bmatrix} = \left[\begin{array}{cccc} 1 & 0 & 0 & 2 \\ 0 & 1 & 0 & 2 \\ \hline 0 & 0| & 1 & 2 \\ 0 & 0| & 1 & -1 \end{array}\right]$$

Passo 3: $m_{43} = -1$, quindi

$$M_3(M_2M_1B) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 2 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 2 \\ 1 & 1 & 1 & -1 \end{bmatrix} = \left[\begin{array}{cccc} 1 & 0 & 0 & 2 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 2 \\ \hline 0 & 0 & 0| & -3 \end{array}\right]$$

Quindi si ha

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix} \text{ e } \quad U = \begin{bmatrix} 1 & 0 & 0 & 2 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & -3 \end{bmatrix}$$

La posizione degli elementi non nulli della matrice B (ovvero il tipo di *pattern*), fa sì che non si abbia alcun fenomeno del *fill-in*. Si osservi che la matrice B si ottiene dalla matrice A con un'opportuna permutazione.

Esempio 2. Si vuole calcolare la decomposizione $PA = LU$ della matrice

$$A = \begin{bmatrix} 1 & 0 & 2 \\ -1 & t & 1 \\ 0 & -1 & 3 \end{bmatrix}, \quad t \in \mathcal{R}$$

con $\det(A) \neq 0$ per ogni $t \neq -1$.
Passo 1. Si ricerca preliminarmente il miglior elemento pivotale nella prima colonna.
Non è necessario fare scambi.
Si ha $m_{21} = -1$ e $m_{31} = 0$, quindi

$$M_1A = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 2 \\ -1 & t & 1 \\ 0 & -1 & 3 \end{bmatrix} = \left[\begin{array}{ccc} 1 & 0 & 2 \\ \hline 0| & t & 3 \\ 0| & -1 & 3 \end{array}\right]$$

Passo 2. Si ricerca preliminarmente il miglior elemento pivotale nella seconda sottocolonna.

Caso $|t| \geq |-1| = 1$: non è necessario fare scambi.
Si ha $m_{32} = -1/t$, quindi

$$M_2(M_1A) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1/t & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 2 \\ 0 & t & 3 \\ 0 & -1 & 3 \end{bmatrix} = \left[\begin{array}{ccc} 1 & 0 & 2 \\ 0 & t & 3 \\ \hline 0 & 0| & 3/t+3 \end{array}\right]$$

In definitiva

$$\begin{aligned} A &= LU = (M_2M_1)^{-1}U = M_1^{-1}M_2^{-1}U \\ &= \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1/t & 1 \end{bmatrix} U \\ &= \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -1/t & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 2 \\ 0 & t & 3 \\ 0 & 0 & 3/t+3 \end{bmatrix} \end{aligned}$$

Caso $|t| < |-1| = 1$**:** si scambiano fra loro la seconda e la terza riga considerando la matrice di permutazione P_2 e ottenendo la matrice $P_2(M_1A)$, con

$$P_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \text{ e } \quad P_2(M_1A) = \begin{bmatrix} 1 & 0 & 2 \\ 0 & -1 & 3 \\ 0 & t & 3 \end{bmatrix}.$$

Si ha $m_{32} = -t$, quindi

$$M_2(P_2(M_1A)) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & t & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 2 \\ 0 & -1 & 3 \\ 0 & t & 3 \end{bmatrix} = \left[\begin{array}{ccc} 1 & 0 & 2 \\ 0 & -1 & 3 \\ \hline 0 & 0 & 3t+3 \end{array}\right]$$

In definitiva, si ha

$$U = M_2P_2M_1A = M_2(P_2M_1P_2)P_2A =$$

essendo $P_2P_2 = I$; da cui

$$U = M_2\tilde{M}_1P_2A, \quad \text{con } \tilde{M}_1 = (P_2M_1P_2)$$

e infine

$$\begin{aligned} PA &= P_2A = \tilde{M}_1{}^{-1}M_2^{-1}U \\ &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & -t & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 2 \\ 0 & -1 & 3 \\ 0 & 0 & 3t+3 \end{bmatrix} \end{aligned}$$

10 Metodi diretti per la risoluzione di sistemi lineari: fattorizzazione QR

10.1 Metodo di ortonormalizzazione di Gram-Schmidt

Per descrivere il metodo di ortonormalizzazione di Gram-Schmidt occorre la seguente definizione preliminare.

Definizione 10.1 *Vettori ortogonali*
Siano $\underline{x}, \underline{y} \in \mathbb{R}^n$. *Si dice che i vettori* $\underline{x}$ *e* $\underline{y}$ *sono ortogonali fra di loro se*

$$\underline{x}^T \underline{y} = 0$$

Ora, siano $\underline{a}_1, \underline{a}_2, \ldots, \underline{a}_n \in \mathbb{R}^n$, n vettori linearmente indipendenti fra loro.
Si vogliono determinare $\underline{q}_1, \underline{q}_2, \ldots, \underline{q}_n \in \mathbb{R}^n$ vettori ortonormali, ossia tali che per ogni $i, j = 1, \ldots, n$

$$\underline{q}_i^T \underline{q}_j = 0 \quad \text{se } i \neq j$$

e per ogni $i = 1, \ldots, n$

$$\|\underline{q}_i\|_2 = 1$$

Il metodo di ortonormalizzazione di Gram-Schmidt procede nel modo seguente.

I passo. Si pone

$$\underline{q}_1 = \underline{a}_1 / \|\underline{a}_1\|_2$$

cosicché

$$\|\underline{q}_1\| = \|\underline{a}_1 / \|\underline{a}_1\|_2\|_2 = \|\underline{a}_1\|_2 / \|\underline{a}_1\|_2 = 1.$$

L'operazione è lecita in quanto il vettore $\underline{a}_1$ è diverso dal vettore nullo essendo gli $\underline{a}_i$ linearmente indipendenti fra loro, e quindi $\|\underline{a}_1\|_2 \neq 0$ (per le proprietà delle norme) e si può effettuare la normalizzazione.
Inoltre, i vettori $\underline{q}_1, \underline{a}_2, \ldots, \underline{a}_n$ sono a loro volta linearmente indipendenti fra loro.

II passo. Si calcola la quantità

$$\alpha_1 = \underline{q}_1^T \underline{a}_2,$$

detta proiezione di $\underline{a}_2$ nella direzione di $\underline{q}_1$, e si pone

$$\tilde{\underline{q}}_2 = \underline{a}_2 - \alpha_1 \underline{q}_1$$

cosicché i vettori $\tilde{\underline{q}}_2$ e $\underline{q}_1$ risultano essere ortogonali fra loro. Infatti, vale che

$$\begin{aligned}
\underline{q}_1^T \tilde{\underline{q}}_2 &= \underline{q}_1^T (\underline{a}_2 - \alpha_1 \underline{q}_1) \\
&= \underline{q}_1^T \underline{a}_2 - \alpha_1 \underline{q}_1^T \underline{q}_1 \\
&= \alpha_1 - \alpha_1 \|\underline{q}_1\|_2^2 \\
&= \alpha_1 - \alpha_1 \quad \text{essendo } \|\underline{q}_1\|_2 = 1 \\
&= 0.
\end{aligned}$$

Quindi si effettua la normalizzazione, ponendo

$$\underline{q}_2 = \tilde{\underline{q}}_2 / \|\tilde{\underline{q}}_2\|.$$

L'operazione è lecita in quanto il vettore $\tilde{\underline{q}}_2$ è diverso dal vettore nullo essendo i vettori $\underline{q}_1, \tilde{\underline{q}}_2, \underline{a}_3, \ldots, \underline{a}_n$ linearmente indipendenti fra loro.

III passo. Si calcolano la quantità

$$\beta_1 = \underline{q}_1^T \underline{a}_3,$$

detta proiezione di $\underline{a}_3$ nella direzione di $\underline{q}_1$ e la quantità

$$\beta_2 = \underline{q}_2^T \underline{a}_3,$$

detta proiezione di $\underline{a}_3$ nella direzione di $\underline{q}_2$ e si pone

$$\tilde{\underline{q}}_3 = \underline{a}_3 - \beta_1 \underline{q}_1 - \beta_2 \underline{q}_2$$

cosicché il vettore $\tilde{\underline{q}}_3$ risulta essere ortogonale ai vettori $\tilde{\underline{q}}_2$ e $\underline{q}_1$.
Infatti, vale

$$\begin{aligned} \underline{q}_1^T \tilde{\underline{q}}_3 &= \underline{q}_1^T (\underline{a}_3 - \beta_1 \underline{q}_1 - \beta_2 \underline{q}_2) \\ &= \underline{q}_1^T \underline{a}_3 - \beta_1 \underline{q}_1^T \underline{q}_1 - \beta_2 \underline{q}_1^T \underline{q}_2 \\ &= \beta_1 - \beta 1 = 0, \end{aligned}$$

essendo $\underline{q}_1^T \underline{q}_1 = \|\underline{q}_1\|_2^2 = 1$ e $\underline{q}_1^T \underline{q}_2 = 0$ e vale pure

$$\begin{aligned} \underline{q}_2^T \tilde{\underline{q}}_3 &= \underline{q}_2^T (\underline{a}_3 - \beta_1 \underline{q}_1 - \beta_2 \underline{q}_2) \\ &= \underline{q}_2^T \underline{a}_3 - \beta_1 \underline{q}_2^T \underline{q}_1 - \beta_2 \underline{q}_2^T \underline{q}_2 \\ &= \beta_2 - \beta 2 = 0, \end{aligned}$$

essendo $\underline{q}_2^T \underline{q}_2 = \|\underline{q}_2\|_2^2 = 1$ e $\underline{q}_2^T \underline{q}_1 = 0$.
Quindi si effettua la normalizzazione, ponendo

$$\underline{q}_3 = \tilde{\underline{q}}_3 / \|\tilde{\underline{q}}_3\|$$

L'operazione è lecita in quanto il vettore $\tilde{\underline{q}}_3$ è diverso dal vettore nullo essendo i vettori $\underline{q}_1, \underline{q}_2, \tilde{\underline{q}}_3, \underline{a}_4, \ldots, \underline{a}_n$ linearmente indipendenti fra loro.
Procedendo in questo modo, dopo n passi si ottengono i vettori $\underline{q}_1, \underline{q}_2, \ldots, \underline{q}_n$ richiesti.

10.2 Metodo di fattorizzazione QR

La procedura di ortonormalizzazione di Gram-Schmidt appena illustrata, può essere reinterpretata come una procedura di fattorizzazione QR, ove Q è una matrice ortogonale (ossia tale che $QQ^T = Q^T Q = I$) e R è una matrice triangolare superiore.
In quest'ottica, occorre innanzitutto scrivere le relazioni che legano gli elementi

di A a quelli della matrice prodotto QR, cui la prima viene uguagliata, ossia

$$\begin{aligned} a_{ij} &= \sum_{j=1}^{n} q_{ij} r_{jk} \\ &= \sum_{j=1}^{k} q_{ij} r_{jk} \end{aligned}$$

in quanto $r_{jk} = 0$ per $j = k+1, \ldots, n$ essendo R triangolare superiore.
Indicando con $\underline{a}_k$ la $k-$sima colonna della matrice A e riscrivendo tale formula in blocco per le colonne, si ha

$$\begin{aligned} \underline{a}_1 &= \underline{q}_1 r_{11} \\ \underline{a}_2 &= \underline{q}_1 r_{12} + \underline{q}_2 r_{22} \\ \underline{a}_3 &= \underline{q}_1 r_{13} + \underline{q}_2 r_{23} + \underline{q}_3 r_{33} \\ &\vdots \\ \underline{a}_k &= \underline{q}_1 r_{1k} + \underline{q}_2 r_{2k} + \ldots + \underline{q}_{k-1} r_{k-1k} + \underline{q}_k r_{kk} \\ &\vdots \end{aligned}$$

Quindi si può scrivere

$$\underline{q}_k r_{kk} = \underline{a}_k - (\underline{q}_1 r_{1k} + \underline{q}_2 r_{2k} + \ldots + \underline{q}_{k-1} r_{k-1k}) = \underline{a}_k - \sum_{j=1}^{k-1} \underline{q}_j r_{jk}.$$

Pertanto, scegliendo i coefficienti $r_{jk}, j = 1, \ldots, k-1$ come le proiezioni del vettore $\underline{a}_k$ nella direzione dei vettori $\underline{q}_j, j = 1, \ldots, k-1$, il vettore

$$\tilde{\underline{q}}_k = \underline{a}_k - \sum_{j=1}^{k-1} \underline{q}_j r_{jk}$$

risulta essere ortogonale ai vettori $\underline{q}_j$, $j = 1, \ldots, k-1$. Inoltre, scegliendo il coefficiente $r_{kk} = \|\tilde{\underline{q}}_k\|_2$ il vettore $\underline{q}_k = \tilde{\underline{q}}_k / r_{kk}$ risulta avere norma 2 unitaria. L'intera procedura corrisponde a quella precedentemente descritta come ortonormalizzazione di Gram-Schmidt.

Una volta ottenuta la fattorizzazione QR la si può utilizzare per la risoluzione del sistema lineare

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

procedendo in modo analogo a quanto visto nel caso della fattorizzazione LU. Più precisamente, si ottiene che si devono risolvere i due sistemi lineari

$$\begin{aligned} Q\underline{y} &= \underline{b} \\ R\underline{x} &= \underline{y} \end{aligned}$$

Il vantaggio dell'introduzione della matrice Q al posto della matrice A originaria, consiste nel fatto che, essendo Q ortogonale, ossia valendo $QQ^T = Q^TQ = I$ si ha che, per unicità dell'inversa,

$$Q^{-1} = Q^T,$$

cosicché il vettore $\underline{y} = Q^{-1}\underline{b}$ viene ottenuto semplicemente considerando il prodotto matrice-vettore

$$\underline{y} = Q^T\underline{b}$$

(e non la risoluzione del sistema lineare).
Per quanto riguarda il sistema $R\underline{x} = \underline{y}$, essendo R una matrice triangolare superiore, esso viene risolto con la procedura di risoluzione backward, già vista nel caso della fattorizzazione LU.
Prima di procedere alla scrittura dell'algoritmo per la determinazione della fattorizzazione QR con il procedimento indicato, è opportuno riflettere sulla seguente questione: si può modificare l'ordine delle operazioni di normalizzazione così da rendere più agevole l'introduzione di un'eventuale tecnica di pivot per colonne.
In effetti, il risultato è inalterato se si procedere per "aggiornamenti" successivi delle colonne di A, cosicché al $k-$simo passo le prime k colonne saranno ortonormali fra loro, mentre le rimanenti colonne lo saranno già rispetto alle prime k, ma non fra di loro.
Si può quindi scrivere il seguente algoritmo

per k=1,...,n
$r_{kk} = \|\underline{a}_k\|_2$
$\underline{a}_k = \underline{a}_k / r_{kk}$

per j=k+1,...,n
$r_{kj} = \underline{a}_k^T \underline{a}_j$
$\underline{a}_j = \underline{a}_j - r_{kj}\underline{a}_k$

che trasforma la matrice A nella matrice Q e contemporaneamente costruisce la matrice R.
L'algoritmo, scritto in questo modo, permette l'introduzione di un'eventuale tecnica di pivot, in cui al passo k si sceglie la miglior colonna fra le colonne rimanenti, ossia dalla $k-$sima all' $n-$sima.
Il costo della fattorizzazione in termini di operazioni moltiplicative è dell'ordine di n^3; mentre quello di risoluzione è dell'ordine di $n^2 + n^2/2$.
Occorre sottolineare che l'algoritmo della fattorizzazione QR in genere utilizzato non è quello sopra descritto, esso risulta in genere molto stabile e questo motiva il maggior costo computazionale rispetto alla fattorizzazione LU.

11 Metodi diretti per la risoluzione di sistemi lineari: fattorizzazione LL^H

11.1 La fattorizzazione LL^H

Sia $A \in C^{n\times n}$ una matrice hermitiana definita positiva, ossia tale che $A^H = A$ e per ogni $\underline{x} \neq \underline{0}$ e $\underline{x}^H A\underline{x} > 0$, vale il seguente teorema di fattorizzazione.

Teorema 11.1 *Sia $A \in C^{n\times n}$ una matrice hermitiana definita positiva, allora esiste ed è unica la fattorizzazione*

$$A = LL^H$$

con L matrice triangolare inferiore.

Per determinare tale fattorizzazione si può procedere nel modo seguente.
Vista l'uguaglianza tra A e LL^H, vale evidentemente

$$\begin{aligned} a_{ij} &= \sum_{k=1}^{n} l_{ik} l_{kj}^H \\ &= \sum_{k=1}^{n} l_{ik}\bar{l}_{jk} \qquad \text{poiché } l_{kj}^H = \bar{l}_{jk} \\ &= \sum_{k=1}^{\min(i,j)} l_{ik}\bar{l}_{jk} \quad \text{poiché } l_{ik} = 0 \text{ per } k > i \text{ e } l_{jk} = 0 \text{ per } k > j \end{aligned}$$

Si tratta quindi di determinare una procedura per calcolare gli elementi $l_{ij},\ i = 1,\dots,n\ j = 1,\dots,i$.
Si considera l'ordinamento seguente: si calcola l'elemento diagonale l_{11} e poi la relativa sottocolonna $l_{i1},\ i = 2,\dots,n$, poi il successivo elemento diagonale l_{22} e la sua sottocolonna $l_{i2},\ i = 3,\dots,n$ e così via fino all'ultimo passo in cui si calcola solo l'elemento l_{nn}, non essendoci relativa sottocolonna.
Consideriamo per primo il caso dell'elemento in posizione diagonale per cui vale

$$a_{jj} = \sum_{k=1}^{j} l_{jk}\bar{l}_{jk} = \sum_{k=1}^{j} |l_{jk}|^2 = \sum_{k=1}^{j-1} |l_{jk}|^2 + |l_{jj}|^2$$

ovvero

$$|l_{jj}|^2 = a_{jj} - \sum_{k=1}^{j-1} |l_{jk}|^2$$

Ora, vale che

$$\begin{aligned} \det(A) &= \det(LL^H) \underbrace{=}_{Binet} \det(L)\det(L^H) \\ &= \det(L)\overline{\det(L)} = |\det(L)|^2 \\ &= \left|\prod_{j=1}^{n} l_{jj}\right|^2 \quad \text{poiché } L \text{ è una matrice triangolare} \\ &= \prod_{j=1}^{n} |l_{jj}|^2 \end{aligned}$$

ove essendo A hermitiana definita positiva è $a_{jj} > 0$ per ogni $j = 1,\dots,n$ e $\det(A) > 0$.
Quindi, per ogni $j = 1,\dots,n$ è $|l_{jj}|^2 > 0$ cosicché

$$a_{jj} = \sum_{k=1}^{j-1} |l_{jk}|^2 + |l_{jj}|^2 > \sum_{k=1}^{j-1} |l_{jk}|^2.$$

È quindi lecito estrarre la radice quadrata e porre per ogni $j = 1, \ldots, n$

$$l_{jj} = \sqrt{a_{jj} - \sum_{k=1}^{j-1} |l_{jk}|^2} \in \mathbb{R}$$

Per quanto riguarda gli elementi della corrispondente sottodiagonale, ossia gli elementi $l_{ij}, i = j+1, \ldots, n$, vale che

$$a_{ij} = \sum_{k=1}^{j} l_{ik}\bar{l}_{jk} = \sum_{k=1}^{j-1} l_{ik}\bar{l}_{jk} + l_{ij}\bar{l}_{jj}$$

da cui, essendo $\bar{l}_{jj} = l_{jj}$,

$$l_{ij} = \left(a_{ij} - \sum_{k=1}^{j-1} l_{ik}\bar{l}_{jk} \right) / l_{jj}$$

In definitiva si può scrivere il seguente algoritmo:

$$\left[\begin{array}{l} \text{per j=1,...,n} \\ \qquad l_{jj} = \sqrt{a_{jj} - \sum_{k=1}^{j-1} |l_{jk}|^2} \\ \qquad \left[\begin{array}{l} \text{per i=j+1,...,n} \\ \qquad\qquad l_{ij} = \left(a_{ij} - \sum_{k=1}^{j-1} l_{ik}\bar{l}_{jk} \right) / l_{jj} \end{array} \right. \end{array} \right.$$

Il costo della fattorizzazione in termini di operazioni moltiplicative è dell'ordine di $n^3/6$, ossia il medesimo del metodo di Gauss, se questo viene opportunamente implementato per il caso di matrici simmetriche; mentre quello di risoluzione è dell'ordine di n^2.
È evidente che tale tipo di fattorizzazione non permette l'utilizzo di alcuna tecnica di pivot in quanto essa distruggerebbe la struttura simmetrica della matrice assegnata.
Tuttavia è possibile dimostrare che il procedimento è stabile e rispetto alla previsione della massima crescita degli elementi di L vale il seguente risultato

$$l_M \leq \sqrt{a_M}$$

ove a_M è l'elemento di modulo massimo della matrice A e l_M è l'elemento di modulo massimo della matrice L ottenuta.

12 Metodi per il calcolo di autovalori estremi

12.1 Premesse

Il problema del calcolo degli autovalori di una matrice $A \in \mathbb{C}^{n\times n}$ potrebbe essere in linea teorica affrontato calcolando le radici del polinomio caratteristico

$$p_n(\lambda) = \det(A - \lambda I) = 0.$$

Tuttavia, questo approccio non risulta in generale conveniente su calcolatore, in quanto i metodi che si usano per calcolare le radici di un polinomio di grado n sono in genere malcondizionati se gli autovalori della matrice non sono ben separati.
Una possibilità alternativa è invece quella di applicare alla matrice A delle opportune trasformazioni per similitudine (in genere mediante matrici ortogonali) così da trasformarla in una matrice B, di cui sia più facile il calcolo degli autovalori.
Infatti vale la proprietà che matrici simili hanno lo stesso polinomio caratteristico e quindi gli stessi autovalori, come enunciato nel seguente teorema.

Teorema 12.1 *Sia $A \in \mathbb{C}^{n\times n}$ e sia $B \in \mathbb{C}^{n\times n}$ simile ad A, ossia esista una matrice $S \in \mathbb{C}^{n\times n}$ invertibile tale che*

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

Allora gli autovalori di B coincidono con gli autovalori di A.

Dimostrazione Vale che

$$\begin{aligned}
\det(A - \lambda I) &= \det(SBS^{-1} - \lambda I) \\
&= \det(S(B - \lambda I)S^{-1} \\
&= \det(S)\det(B - \lambda I)\det(S^{-1}) \quad \text{teorema di Binet} \\
&= \det(S)\det(B - \lambda I)(\det(S))^{-1} \\
&= \det(B - \lambda I).
\end{aligned}$$

Pertanto le due matrici hanno i medesimi autovalori in quanto hanno il medesimo polinomio caratteristico.

12.2 Metodo delle potenze

Sia $A \in \mathbb{C}^{n\times n}$. Si vuole calcolare l'autovalore di modulo massimo e il corrispondente autovettore.
Si assuma che, detti $\lambda_1, \lambda_2, \ldots, \lambda_n$, gli autovalori della matrice A valga la relazione di ordinamento

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \ldots \geq |\lambda_n|$$

Questa assunzione è equivalente alla richiesta che

- λ_1 sia un autovalore semplice (ossia λ_1 è radice del polinomio caratteristico una sola volta)

- $-\lambda_1$ non sia autovalore di A
- $\overline{\lambda}_1$ non sia autovalore di A

Tale ipotesi può essere eventualmente indebolita nel caso di autovalore di modulo massimo di molteplicità r (ossia λ_1 è radice del polinomio caratteristico r volte) nel modo seguente

$$\begin{aligned} |\lambda_1| &= |\lambda_2| = \ldots = |\lambda_r| > |\lambda_{r+1}| \geq |\lambda_{r+1}| \geq \ldots \geq |\lambda_n| \\ \lambda_1 &= \lambda_2 = \ldots ... = \lambda_r \end{aligned}$$

Considerato un vettore di innesco $\underline{z}^{[0]}$, il metodo prevede di generare la seguente successione di vettori

$$\underline{z}^{[k]} = A\underline{z}^{[k-1]}, \quad k \geq 1$$

Nonostante non sia necessario per la convergenza del metodo, assumiamo per semplicità che, detti $\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_n$ gli autovettori corrispondenti agli autovalori $\lambda_1, \lambda_2, \ldots, \lambda_n$, essi siano fra loro linearmente indipendenti e quindi formino una base di $\mathbb{C}^n$.
Si può allora scrivere

$$\underline{z}^{[0]} = \sum_{j=1}^{n} \alpha_j \underline{v}_j$$

e assumiamo che sia $\alpha_1 \neq 0$, ossia nel vettore $\underline{z}^{[0]}$ è effettivamente presente un contributo dell'autovettore $\underline{v}_1$ che si vuole calcolare.
Ora, per come è stata costruita la successione, vale evidentemente che

$$\underline{z}^{[k]} = A\underline{z}^{[k-1]} = A^2\underline{z}^{[k-2]} = \ldots = A^k\underline{z}^{[0]}$$

e, ricordando l'espressione di $\underline{z}^{[0]}$, si ha

$$\begin{aligned} \underline{z}^{[k]} &= A^k \sum_{j=1}^{n} \alpha_j \underline{v}_j \\ &= \sum_{j=1}^{n} \alpha_j A^k \underline{v}_j \\ &= \sum_{j=1}^{n} \alpha_j \lambda_j^k \underline{v}_j \end{aligned}$$

in quanto, se $A\underline{v}_j = \lambda_j \underline{v}_j$ allora $A^2\underline{v}_j = A(A\underline{v}_j) = A(\lambda_j \underline{v}_j) = \lambda_j A\underline{v}_j = \lambda_j^2 \underline{v}_j$ e più in generale

$$A^k \underline{v}_j = \lambda_j^k \underline{v}_j,$$

cioè, noti gli autovalori e autovettori di una matrice A, la matrice A^k ha come autovalori le potenze $k-$sime degli autovalori di A e ha i medesimi autovettori.
Ora, mettendo in evidenza λ_1^k, si ha

$$\underline{z}^{[k]} = \lambda_1^k \left(\alpha_1 \underline{v}_1 + \sum_{j=2}^{n} \alpha_j \left(\frac{\lambda_j}{\lambda_1} \right)^k \underline{v}_j \right) \tag{10}$$

È quindi evidente che

$$\lim_{k\to+\infty} \underline{z}^{[k]} = \lim_{k\to+\infty} \lambda_1^k \left(\alpha_1 \underline{v}_1 + \sum_{j=2}^{n} \alpha_j \left(\frac{\lambda_j}{\lambda_1} \right)^k \underline{v}_j \right) = \lim_{k\to+\infty} \lambda_1^k \alpha_1 \underline{v}_1$$

in quanto, per ogni $j \geq 2$

$$\lim_{k\to+\infty} \left(\frac{\lambda_j}{\lambda_1} \right)^k = 0$$

essendo $|\lambda_1| > |\lambda_j|$ per ogni $j \geq 2$.
Quindi il vettore $\underline{z}^{[k]}$ tende a disporsi in direzione parallela a $\underline{v}_1$, ovvero tende ad essere un autovettore relativamente all'autovalore λ_1.
Lasciando un momento in sospeso la questione del calcolo effettivo dell'autovettore, in quanto nell'espressione appena vista compare il fattore λ_1^k, vediamo come è possibile calcolare l'autovalore di modulo massimo.
Per ogni $k \geq 0$, si dice quoziente di Rayleigh al passo k
la quantità

$$\sigma_k = \frac{(\underline{z}^{[k]})^H A \underline{z}^{[k]}}{(\underline{z}^{[k]})^H \underline{z}^{[k]}}$$

Ora, tenuto conto della (10)

$$\begin{aligned}
\sigma_k &= \frac{(\underline{z}^{[k]})^H A \underline{z}^{[k]}}{(\underline{z}^{[k]})^H \underline{z}^{[k]}} \\
&= \frac{\left(\lambda_1^k \left(\alpha_1 \underline{v}_1 + \sum_{j=2}^n \alpha_j \left(\frac{\lambda_j}{\lambda_1} \right)^k \underline{v}_j \right) \right)^H A \lambda_1^k \left(\alpha_1 \underline{v}_1 + \sum_{j=2}^n \alpha_j \left(\frac{\lambda_j}{\lambda_1} \right)^k \underline{v}_j \right)}{\left(\lambda_1^k \left(\alpha_1 \underline{v}_1 + \sum_{j=2}^n \alpha_j \left(\frac{\lambda_j}{\lambda_1} \right)^k \underline{v}_j \right) \right)^H \lambda_1^k \left(\alpha_1 \underline{v}_1 + \sum_{j=2}^n \alpha_j \left(\frac{\lambda_j}{\lambda_1} \right)^k \underline{v}_j \right)} \\
&= \frac{\left(\overline{\lambda}_1^k \left(\overline{\alpha}_1 \underline{v}_1^H + \sum_{j=2}^n \overline{\alpha}_j \left(\overline{\frac{\lambda_j}{\lambda_1}} \right)^k \underline{v}_j^H \right) \right) A \lambda_1^k \left(\alpha_1 \underline{v}_1 + \sum_{j=2}^n \alpha_j \left(\frac{\lambda_j}{\lambda_1} \right)^k \underline{v}_j \right)}{\left(\overline{\lambda}_1^k \left(\overline{\alpha}_1 \underline{v}_1^H + \sum_{j=2}^n \overline{\alpha}_j \left(\overline{\frac{\lambda_j}{\lambda_1}} \right)^k \underline{v}_j^H \right) \right) \lambda_1^k \left(\alpha_1 \underline{v}_1 + \sum_{j=2}^n \alpha_j \left(\frac{\lambda_j}{\lambda_1} \right)^k \underline{v}_j \right)}
\end{aligned}$$

e

$$\begin{aligned}
\lim_{k\to+\infty} \sigma_k &= \lim_{k\to+\infty} \frac{\left(\overline{\lambda}_1^k \left(\overline{\alpha}_1 \underline{v}_1^H + \sum_{j=2}^n \overline{\alpha}_j \left(\overline{\frac{\lambda_j}{\lambda_1}} \right)^k \underline{v}_j^H \right) \right) A \lambda_1^k \left(\alpha_1 \underline{v}_1 + \sum_{j=2}^n \alpha_j \left(\frac{\lambda_j}{\lambda_1} \right)^k \underline{v}_j \right)}{\left(\overline{\lambda}_1^k \left(\overline{\alpha}_1 \underline{v}_1^H + \sum_{j=2}^n \overline{\alpha}_j \left(\overline{\frac{\lambda_j}{\lambda_1}} \right)^k \underline{v}_j^H \right) \right) \lambda_1^k \left(\alpha_1 \underline{v}_1 + \sum_{j=2}^n \alpha_j \left(\frac{\lambda_j}{\lambda_1} \right)^k \underline{v}_j \right)} \\
&= \lim_{k\to+\infty} \frac{\overline{\lambda}_1^k \overline{\alpha}_1 \underline{v}_1^H A \lambda_1^k \alpha_1 \underline{v}_1}{\overline{\lambda}_1^k \overline{\alpha}_1 \underline{v}_1^H \lambda_1^k \alpha_1 \underline{v}_1} \\
&= \lim_{k\to+\infty} \frac{\underline{v}_1^H A \underline{v}_1}{\underline{v}_1^H \underline{v}_1} \\
&= \frac{\underline{v}_1^H A \underline{v}_1}{\underline{v}_1^H \underline{v}_1} \\
&= \frac{\underline{v}_1^H \lambda_1 \underline{v}_1}{\underline{v}_1^H \underline{v}_1}
\end{aligned}$$

$$= \quad \lambda_1 \frac{\|\underline{v}_1\|_2^2}{\|\underline{v}_1\|_2^2} = \lambda_1$$

ovvero l'autovalore di modulo massimo.
Nel calcolo su calcolatore si ha che

$$\lambda_1^k \to \begin{cases} \infty & \text{se} \quad |\lambda_1| > 1 \quad \text{overflow} \\ 0 & \text{se} \quad |\lambda_1| < 1 \quad \text{underflow} \end{cases}$$

cosicché per evitare questo problema si normalizza la successione dei vettori $\underline{z}^{[k]}$ in norma 2, ossia si considera la successione dei vettori

$$\underline{y}^{[k]} = \underline{z}^{[k]} / \|\underline{z}^{[k]}\|,$$

ove l'operazione di normalizzazione lascia invariate le direzioni.
Inoltre, per ogni $k \geq 0$, vale che il quoziente di Rayleigh è

$$\sigma_k = \frac{(\underline{y}^{[k]})^H A \underline{y}^{[k]}}{(\underline{y}^{[k]})^H \underline{y}^{[k]}} = (\underline{y}^{[k]})^H A \underline{y}^{[k]}$$

essendo $(\underline{y}^{[k]})^H \underline{y}^{[k]} = \|\underline{y}^{[k]}\|_2^2 = 1$.
In definitiva l'algoritmo del metodo delle potenze è il seguente

$$\begin{array}{l} y^{[0]} = z^{[0]} / \|z^{[0]}\|_2 \\ z^{[1]} = A y^{[0]} \\ \sigma_1 = (y^{[0]})^H z^{[1]} \\ \left[\begin{array}{l} \text{per } k \geq 1, \\ \quad y^{[k]} = z^{[k]} / \|z^{[k]}\|_2 \\ \quad z^{[k+1]} = A y^{[k]} \\ \quad \sigma_{k+1} = (y^{[k]})^H z^{[k+1]} \\ \quad \left[\begin{array}{l} \text{se } |\sigma_{k+1} - \sigma_k| \leq \varepsilon \\ \quad y^{[k+1]} = z^{[k+1]} / \|z^{[k+1]}\|_2; \\ \quad \mathit{break} \end{array} \right. \end{array} \right. \end{array}$$

Il costo computazione per iterazione è pari a $n^2 + 3n + \alpha$ con n dimensione della matrice A e α costo dell'operazione di estrazione di radice.
Per quanto riguarda invece le proprietà di convergenza, dal procedimento precedentemente riportato, è evidente che vale

$$|\sigma_k - \lambda_1| = O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right)$$

ossia la convergenza è tanto più veloce quanto $|\lambda_2| < |\lambda_1|$.
Nel caso in cui A sia una matrice hermitiana, si ha una velocità di convergenza maggiore, in quanto vale

$$|\sigma_k - \lambda_1| = O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^{2k}\right)$$

Si noti, infine, che l'ipotesi $\alpha_1 \neq 0$ relativamente al vettore di innesco $z^{[0]}$ non è così importante quando si effettui il calcolo su calcolatore, in quanto, visto gli errori di calcolo introdotti dall'uso dell'aritmetica finita, in poche iterazioni comparirà facilmente un contributo relativo all'autovettore $\underline{v}_1$.

12.3 Metodo delle potenze inverse

Sia $A \in \mathbb{C}^{n\times n}$. Si vuole calcolare l'autovalore di modulo minimo e il corrispondente autovettore.
Si assuma che, detti $\lambda_1, \lambda_2, \ldots, \lambda_n$, gli autovalori della matrice A valga la relazione di ordinamento

$$|\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_{n-1}| > |\lambda_n|$$

Questa assunzione è equivalente alla richiesta che

- λ_n sia un autovalore semplice (ossia λ_n è radice del polinomio caratteristico una sola volta)
- $-\lambda_n$ non sia autovalore di A
- $\overline{\lambda}_n$ non sia autovalore di A

Tale ipotesi può essere eventualmente indebolita nel caso di autovalore di modulo minimo di molteplicità r (ossia λ_1 è radice del polinomio caratteristico r volte) nel modo seguente

$$\begin{array}{rcl} |\lambda_1| & \geq & |\lambda_2| \geq \ldots \geq |\lambda_{n-r}| > |\lambda_{n-r+1}| = |\lambda_{n-r+2}| = \ldots = |\lambda_n| \\ \lambda_{n-r+1} & = & \lambda_{n-r+2} = \ldots = \lambda_n \end{array}$$

L'idea è quella di sfruttare la relazione esistente fra gli autovalori della matrice A e quelli della sua matrice inversa A^{-1}.
Infatti, se $Ax = \lambda x$ si ha pure

$$A^{-1}x = \frac{1}{\lambda}x,$$

ossia gli autovalori della matrice A^{-1} sono gli inversi degli autovalori della matrice A e gli autovettori sono i medesimi.
Quindi

$$\begin{array}{rcl} |\lambda_n| & = & \min\limits_{i=1,\ldots,n} |\lambda_i(A)| \\ & = & \min\limits_{i=1,\ldots,n} \dfrac{1}{|\lambda_i(A^{-1})|} \\ & = & \dfrac{1}{\max_{i=1,\ldots,n} |\lambda_i(A^{-1})|} \end{array}$$

ovvero è sufficiente applicare il metodo delle potenze alla matrice A^{-1} e invertire il quoziente di Rayleigh ottenuto come approssimazione dell'autovalore.
In definitiva l'algoritmo del metodo delle potenze inverse è il seguente

$y^{[0]} = z^{[0]}/\|z^{[0]}\|_2$
$z^{[1]} = A^{-1}y^{[0]}$
$\sigma_1 = (y^{[0]})^H z^{[1]}$
per $k \geq 1$,
  $y^{[k]} = z^{[k]}/\|z^{[k]}\|_2$
  $z^{[k+1]} = A^{-1}y^{[k]}$
  $\sigma_{k+1} = (y^{[k]})^H z^{[k+1]}$
  se $|\sigma_{k+1} - \sigma_k| \leq \varepsilon$
    $y^{[k+1]} = z^{[k+1]}/\|z^{[k+1]}\|_2$;
    $\sigma_{k+1} = 1/\sigma_{k+1}$
    break

È importante sottolineare che non occorre (e non si deve) calcolare la matrice inversa A^{-1}, ma è sufficiente ricondurre il problema del calcolo del vettore

$$z^{[k+1]} = A^{-1}y^{[k]}$$

al calcolo della soluzione del sistema lineare

$$Az^{[k+1]} = y^{[k]}$$

In assenza di particolari proprietà di struttura della matrice A, un metodo conveniente può essere quello della fattorizzazione di Gauss con pivot parziale, in quanto i sistemi lineari hanno sempre la stessa matrice A, che quindi può essere fattorizzata "una tantum" al costo di $n^3/3$ operazioni di tipo moltiplicativo; mentre ad ogni passo del metodo si risolveranno solo i due sistemi con matrice triangolare inferiore e superiore.

12.4 Calcolo dell'autovalore più vicino ad un valore α assegnato

Il calcolo dell'autovalore più vicino ad un valore α assegnato può essere ad esempio di aiuto quando si possegga una stima di un certo autovalore della matrice A e lo si voglia calcolare con maggior precisione.
Tale calcolo può essere effettuato applicando il metodo delle potenze alla matrice

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

Infatti, poiché

$$\lambda_i(B) = \frac{1}{\lambda_i(A - \alpha I)} = \frac{1}{\lambda_i(A) - \alpha}$$

(se $A\underline{x} = \lambda\underline{x}$ allora $(A - \alpha I)\underline{x} = (\lambda - \alpha)\underline{x}$ e la proprietà è vera solo perché la matrice I è una matrice particolare per cui ogni vettore $\underline{x}$ è autovettore relativamente all'autovalore 1), si ha che

$$\max_{i=1,\ldots,n} |\lambda_i(B)| \quad = \quad \frac{1}{\min_{i=1,\ldots,n} |\lambda_i(A) - \alpha|}$$

ossia l'autovalore più vicino a α e quindi, detto j l'indice per cui si realizza tale massimo, vale che

$$\lambda_j(A) = \frac{1}{\lambda_j(B)} + \alpha$$

In definitiva l'algoritmo del metodo è il seguente

$$
\begin{array}{l}
B = A - \alpha I \\
y^{[0]} = z^{[0]}/\|z^{[0]}\|_2 \\
z^{[1]} = B^{-1}y^{[0]} \\
\sigma_1 = (y^{[0]})^H z^{[1]} \\
\left[\begin{array}{l}
\text{per } k \geq 1, \\
\quad y^{[k]} = z^{[k]}/\|z^{[k]}\|_2 \\
\quad z^{[k+1]} = B^{-1}y^{[k]} \\
\quad \sigma_{k+1} = (y^{[k]})^H z^{[k+1]} \\
\quad \left[\begin{array}{l}
\text{se } |\sigma_{k+1} - \sigma_k| \leq \varepsilon \\
\qquad y^{[k+1]} = z^{[k+1]}/\|z^{[k+1]}\|_2; \\
\qquad \sigma_{k+1} = 1/\sigma_{k+1} + \alpha \\
\qquad break
\end{array}\right.
\end{array}\right.
\end{array}
$$

Come nel caso del metodo della potenza inversa, il calcolo del vettore

$$z^{[k+1]} = B^{-1}y^{[k]}$$

viene ricondotto a quello del calcolo della soluzione del sistema lineare

$$Bz^{[k+1]} = y^{[k]}$$